Robust Permutation Test for Equality of Distributions under Covariate-Adaptive Randomization

Mauricio Olivares

University of Illinois at Urbana–Champaign

lvrsgnz2@illinois.edu

November 27, 2020

JOB MARKET PAPER
(LINK TO THE LATEST VERSION)

Abstract

Though stratified randomization achieves more balance on baseline covariates than pure randomization, it does affect the way we conduct inference. This paper considers the classical two-sample goodness-of-fit testing problem in randomized controlled trials (RCTs) when the researcher employs a particular type of stratified randomization—covariate-adaptive randomization (CAR). When testing the null hypothesis of equality of distributions between experimental groups in this setup, we first show that stratification leaves a mark on the test statistic's limit distribution, making it difficult, if not impossible, to obtain critical values. We instead propose an alternative approach to conducting inference based on a permutation test that *i)* is asymptotically exact in the sense that the limiting rejection probability under the null hypothesis equals the nominal α level, *ii)* is applicable under relatively weak assumptions commonly satisfied in practice, and *iii)* works for randomization schemes that are popular among empirically oriented researchers, such as stratified permuted block randomization.

The proposed test's main idea is that by transforming the original statistic by one minus its bootstrap p-value, it becomes asymptotically uniformly distributed on $[0, 1]$. Thus, the transformed test statistic—also called *prepivoted*—has a fixed limit distribution that is free of unknown parameters, effectively removing the effect of stratification. Consequently, a permutation test based on the prepivoted statistic produces a test whose limiting rejection probability equals the nominal level. We present further numerical evidence of the proposed test's advantages in a Monte Carlo exercise, showing our permutation test outperforms the existing alternatives. We illustrate our method's empirical relevance by revisiting a field experiment by Butler and Broockman (2011) on the effect of race on state legislators' responsiveness to help their constituents register to vote during elections in the United States. Lastly, we provide the companion `RATest` `R` package to facilitate and encourage applying our test in empirical research.

Keywords: Covariate-adaptive randomization, stratified block randomization, permutation test, prepivoting, Goodness-of-fit.

JEL Classification: C12, C14, C22.

1 Introduction

Consider first the most straightforward way a researcher carries out randomization in a controlled experiment—simple randomization. In this experimental design, every individual is as likely to be assigned to the treatment or control group. While simple randomization takes care of selection bias, it does not guard the researcher against imbalances over baseline covariates, which may result in loss of statistical efficiency or low estimation precision, even if these imbalances occur purely by chance (Imbens and Rubin, 2015, Chapter 9). This problem further worsens when sample sizes are small or the number of covariates to balance over increases.

In such circumstances, covariate-adaptive randomization (CAR) is a popular randomization technique that exploits observable characteristics—such as geographic, demographic, or other factors before random assignment—to inform the treatment and achieve balance. This form of stratified randomization is reasonably easy to implement and improves upon simple randomization, primarily if the baseline covariates are correlated with the outcome of interest. In essence, CAR consists of two steps—first, define strata as different combinations of covariate levels, and then assign treatment to achieve balance within each stratum. Thus, CAR techniques are relevant in the experimental design, and this fact explains their popularity among empirically oriented researchers.[1]

This paper presents theoretical, empirical, and simulation evidence showing that when testing the null hypothesis of equality of distributions, balancing over covariates using CAR techniques has a detrimental effect on inference. In particular, our first result shows that stratification leaves a mark on the distribution of the classical two-sample Kolmogorov–Smirnov (2SKS) test statistic, making it difficult, if not impossible, to obtain critical values. The complexity attributed to stratification may lead to severe size distortions if we use the asymptotic null distribution's critical values obtained under simple randomization. Indeed, we present simulation results showing that the limiting rejection probability of the 2SKS test can be substantially below its nominal level.

One might wonder whether we can use permutation-based critical values instead of asymptotic ones to bypass the 2SKS test's size distortions, especially because randomization comes from random treatment assignment, making intuitive sense to consider randomization inference for testing. First, we argue that permutation inference based on the 2SKS statistic is not immune to stratification's adverse effects, *i.e.*, the permutation test that does not account for CAR fails to control the type 1 error rate, even in large samples. Next, to demonstrate the quantitative importance of this phenomenon, we present simulation evidence showing that

[1] Bruhn and McKenzie (2009) present a comprehensive review of how these methods are used in development economics. See also Duflo and Banerjee (2017), Bai (2019), and the references therein for a more recent count of these techniques in economics. More broadly, Hu et al. (2014) examine a large class of CAR schemes in clinical trials.

the permutation test based on the 2SKS statistic is not a reliable procedure for the testing problem of interest—the empirical rejection probabilities under the null hypothesis are shockingly different from the nominal level. Consequently, applying conventional testing procedures based on simple randomization may lead to invalid results when the randomization scheme is covariate-adaptive.

To overcome this problem, we introduce a novel permutation test for the aforementioned null hypothesis under CAR. As we will explain below, our proposed test's main idea is that by transforming the 2SKS statistic by its bootstrap cumulative distribution function (CDF), it becomes asymptotically uniformly distributed on $[0, 1]$. Thus, the new test statistic—also called *prepivoted* (Beran, 1987)—has a fixed limit distribution that is free of unknown parameters, effectively removing the effect of stratification. Our main result shows that the permutation test based on the prepivoted statistic will have rejection probability that tends to α for testing equality of distributions under CAR.

Since prepivoting offers an alternative way of rendering test statistics that do not depend on the fundamentals, recent studies exploit this idea to restore the permutation tests' asymptotic validity. The theory developed in Chung and Romano (2016) appears particularly attractive when comparing equality of means of multidimensional observations based on a modified maximum statistic. More recently, Cohen and Fogarty (2020) propose a unified framework to conduct permutation-based inference for Neyman's weak null hypothesis for a large class of test statistics—like the difference-in-means statistic or the absolute difference-in-means—based on Gaussian prepivoting. Our paper describes further steps in this direction by extending this idea to testing problems concerning the entire distributions, not only some aspects of them like their means.[2]

It is essential to realize that, in order to apply the prepivoting method, one must establish the consistency of the bootstrap under CAR. This paper proposes an exchangeable bootstrap approach to estimating the 2SKS statistic's asymptotic null distribution under CAR. Since the conditions we provide allow for several choices of weights for the bootstrap approximation, we carry out the Bayesian bootstrap (Rubin, 1981). Besides the desired consistency property, our bootstrap procedure has the practical advantage that the researcher does not need to know the control variables that give rise to strata. This advantage becomes meaningful in field experiments where ethical considerations play a central role. For example, field experiments frequently hide pre-treatment characteristics to fulfill their IRB commitments to keep the subjects in the experiment anonymous, thus ensuring minimal risk (Duflo and Banerjee, 2017, Chapter 5).

[2]We can find permutation tests based on modified test statistics that do not depend on the fundamentals, like prepivoting here, in other contexts. Notable examples include the pioneering works of Neuhaus (1993) and Janssen (1997, 1999). More recently, Chung and Romano (2013, 2016) generalize this principle to handle general finite-dimensional testing problems, whereas Chung and Olivares (2020) consider a modified test statistic for the classical goodness-of-fit testing problem with an estimated nuisance parameter.

To better understand our method and show its empirical relevance, we present a reappraisal of the field experiment by Butler and Broockman (2011) about the effect of race on state legislators' responsiveness to help their constituents register to vote during the 2008 U.S. elections. Based on our permutation test, we find empirical evidence suggesting that legislators show more responsiveness to those constituents who, based on their race, are inferred to be of the same party. However, the response rates to each racial alias are indistinguishable once we signal the constituent's partisanship, thus complementing the results in Butler and Broockman (2011). We accompany our analysis with the `RATest` R package—available on CRAN—to ease and encourage the application of our test in empirical research.

Previous research has realized and discussed the consequences of CAR techniques on the way we conduct inference.[3] The seminal works of Birkett (1985) and Forsythe (1987) document that the simple two-sample t-test (2StT) is conservative under CAR via Monte Carlo simulation, raising concerns about its validity if adaptive randomization is present. Shao, Yu, and Zhong (2010) formalize the statistical properties of the 2StT under CAR, sparking an increasing body of research seeking to understand this phenomenon for a large class of CAR techniques and experimental designs. Along this line, Ma, Hu, and Zhang (2015); Ye (2018); Ma et al. (2020) study t and Wald tests' theoretical properties under CAR schemes and propose corrections based on the asymptotic critical values. Bugni, Canay, and Shaikh (2018, 2019) extend this approach to linear regression with strata fixed effects models and multiple treatments. Bai (2019) shows that matched-paired designs—a type of CAR scheme with only two units per stratum—is optimal among all stratified randomization designs in the sense of minimizing the difference-in-means estimator's (second moment of the) ex-post bias. Other extensions include survival models (Ye, Yi, and Shao, 2020), adaptive randomization in network data (Zhou, Li, and Hu, 2020), quantile regression Zhang and Zheng (2020), and randomization inference (Simon and Simon, 2011; Bugni, Canay, and Shaikh, 2018).

However, except for Zhang and Zheng (2020), all the papers above only consider making inference about low-dimensional parameters, whether it is the average treatment effect or the slope coefficients in a regression model. Unlike these methods, we revisit the classical goodness-of-fit testing problem, *i.e.*, in our testing problem, the parameter of interest is the entire outcome distributions rather than one aspect of them, such as their mean. Thus, our testing problem posits two significant challenges. First, even though we can characterize the interconnection between the randomization scheme and inference, our test statistic's limit distribution depends

[3]We focus on CAR only, but alternative randomization schemes and their potential effects on statistical inference are also present in literature. Notable examples include the pioneering works of Begg and Iglewicz (1980); Atkinson (1982); Smith (1984a,b) on model-based randomization for estimation efficiency, and more recently Baldi Antognini and Zagoraiou (2011). Zhang et al. (2007); Hu and Rosenberger (2006); Rosenberger and Sverdlov (2008); Hu, Zhang, and He (2009) introduce adaptive randomization techniques based on outcomes in addition to covariates. Re-randomization is considered in Morgan and Rubin (2012); Kuznetsova and Tymofyeyev (2013); Basse and Airoldi (2018); Cohen and Fogarty (2020). Alternatively, Bertsimas, Johnson, and Kallus (2015) provides an approach based on optimization as opposed to randomization.

on the fundamentals and stratification, making it difficult, if not impossible, to obtain critical values. Second, and more importantly, one cannot restore our test procedures' validity by simple studentization, as in the 2StT case, e.g. Bugni, Canay, and Shaikh (2018). In this paper, we propose a new approach to sidestep these difficulties based on the prepivoting idea of Beran (1987, 1988).

The layout of the article is organized as follows. In the next section, we introduce the statistical environment, notation, and the statistical problem at hand. We study the adverse effects of stratification on how we conduct inference via the classical 2SKS and permutation tests in Section 3. The same section shows how these testing procedures fail to control the type 1 error, even asymptotically. Section 4 introduces our permutation test and establishes this paper's main results under general conditions to address this difficulty. Under weak assumptions, we show that the permutation test based on the prepivoted statistic has limiting rejection probability under the null hypothesis equal to the nominal level. Section 5 contains some simulation results, and we dedicate Section 6 to the empirical illustration. Finally, a summary of this paper's contributions and conclusions are collected in Section 7. Appendices A–D contain the proofs, auxiliary lemmas, and additional material.

2 Statistical Environment

2.1 Setup and Notation

We consider the standard randomized experiment setup, where Y denotes the (continuous) outcome of interest, and Z is a vector of pre-treatment covariates. Let A be a treatment indicator such that $A = 1$ if the experimental unit receives treatment, and $A = 0$ otherwise. Define $Y(1)$ as the potential outcome if the experimental unit belongs to the treatment group, and $Y(0)$ if it belongs to the control group. The following rule determines the observed outcomes

$$Y = Y(0) + (Y(1) - Y(0))A_i \ .$$

Throughout the paper, we maintain the following assumption about the data available to the researcher. This assumption is standard in the type of econometric applications we have in mind:

A. 1. *The data are an independent and identically distributed sample* $\{(Y_i, A_i, Z_i) : 1 \leq i \leq N\}$ *from the distribution of* (Y, A, Z)*, denoted* $\mathcal{Q}$*.*

To show how exactly CAR works, it is useful to introduce the stratification rule as a function of baseline covariates. Let $S : \text{supp}(Z) \to \mathscr{S}$ be a discrete function that generates the strata, with $p(s) = \mathbb{P}\{S = s\} > 0$ for all $s \in \mathscr{S}$.

Remark 1. Some authors separate the elements in Z into two subsets, one that the researcher uses to inform the treatment and another subset as part of the working model (Shao, Yu, and Zhong, 2010; Ma, Hu, and Zhang, 2015; Ma et al., 2020). If the researcher uses the elements in Z that are part of the randomization in the test's construction, then a correctly specified model between Y and those covariates is required to construct a valid test. We make no such distinction. ■

If Z consists of p baseline covariates, and each covariate has s_j levels, $j = 1, \ldots, p$, the total number of strata is $|\mathscr{S}| = \prod_{j=1}^{p} s_j < \infty$. For every experimental unit $1 \leq i \leq N$, we generate A_i after we observe Z_i. Collect the treatment indicators and strata, respectively, into $\mathbf{A}_k = (A_1, \ldots, A_k) \in \{0,1\}^k$ and $\mathbf{S}_k = (S_1, \ldots, S_k) \in \mathscr{S}^k$ for $1 \leq k \leq N$, where $S_i = S(Z_i)$. Denote $Y_{1,i} = Y_i$ among the treated, and $Y_{0,i} = Y_i$ among the non-treated, and collect all these outcomes in one vector as $\boldsymbol{X} = (Y_{1,1}, \ldots, Y_{1,m}, Y_{0,1}, \ldots, Y_{0,n}) = (X_1, \ldots, X_N)$.

Remark 2. Stratification through a discrete function S is a common feature of covariate-adaptive designs, where the researcher typically categorizes continuous covariates to form strata, e.g., Shao, Yu, and Zhong (2010); Ma, Hu, and Zhang (2015); Bugni, Canay, and Shaikh (2018, 2019). However, discretizing continuous covariates comes at the expense of losing information and the additional effort to define the categories judiciously. This problematic has led to new CAR schemes that achieve balance without breaking continuous covariates down into categories. See Hu et al. (2014)—and references therein—for a thorough review of the literature in this regard. ■

Consider the following device:

$$\mathcal{D}_N(s) = \sum_{i=1}^{N} (A_i - \lambda)\, \mathbb{1}_{\{S_i = s\}}\ ,\ s \in \mathscr{S},\ \lambda \in (0,1)\ . \tag{1}$$

The previous function measures the within-stratum degree of imbalance for a pre-specified fraction λ.[4] Typically $\lambda = 1/2$, meaning that the research design allocates half of the subjects to the treatment group in every stratum. Thus, $\mathcal{D}_N(s) > 0$ means that there are more subjects in the treatment group relative to the control group, and analogously if $\mathcal{D}_N(s) < 0$. Throughout this paper, we assume λ is the same regardless of the stratum though we can relax this requirement and allow for different target proportions for different strata, e.g. Bugni, Canay, and Shaikh (2019) or Ye, Yi, and Shao (2020).

For $s \in \mathscr{S}$, let $m(s) = |\{1 \leq i \leq N : A_i = 1, S_i = s\}|$ and similarly $n(s)$ with $A_i = 0$

[4] There are different measures of imbalance besides the one we consider here. For example the overall imbalance measure, $\mathcal{D}_N = \sum_{i=1}^{N} (A_i - \lambda)$, or the marginal imbalance of Pocock and Simon (1975). Different measures of imbalance give rise to different CAR designs. For example, Hu and Hu (2012) procedure minimizes the weighted average of overall, within-stratum, and marginal imbalance, whereas the model-based approach in Smith (1984a) defines imbalance to achieve optimality results. See Rosenberger and Lachin (2015) for a review.

replacing $A_i = 1$. We now discuss an additional assumption about the treatment mechanism

A. 2. *The treatment assignment mechanism is such that:*

i) $\Big\{(Y_i(1), Y_i(0), Z_i) \,:\, 1 \leq i \leq N\Big\} \perp\!\!\!\perp \mathbf{A}_N \,|\, \mathbf{S}_N$.

ii)

$$\left\{ \left\{ \frac{\mathcal{D}_N(s)}{\sqrt{N}} \right\}_{s \in \mathscr{S}} \middle| \, \mathbf{S}_N \right\} \xrightarrow{\text{d}} \mathcal{N}\left(0, \Sigma_D\right)$$

where $\Sigma_D = diag\{\, p(s)\,\tau(s) \,:\, s \in \mathscr{S}\}$ *with* $0 \leq \tau(s) \leq \lambda(1-\lambda)$ *for all* $s \in \mathscr{S}$.

Assumption A.2 is Bugni, Canay, and Shaikh (2019, Assumption 2.2). The first part of this assumption asserts that, while the treatment assignments and the observed outcomes are dependent, treatment assignments do not affect the potential outcomes, conditionally on strata. Moreover, given strata, Z may contain covariates not used for CAR, so we do not have to specify a model between observed outcomes and these additional covariates (see Remark 1).

The idea behind Assumption A.2 *ii)* is that, conditionally on strata, the fraction of units in the treatment group concentrates around the target proportion λ across strata as the sample increases. This condition holds for the most commonly used CAR schemes, such as stratified permuted block randomization (Fisher, 1934; Zelen, 1974), covariate-adaptive biased coin design (Efron, 1971; Baldi Antognini and Zagoraiou, 2011), and covariate-adaptive urn design (Wei, 1978; Baldi Antognini and Giovagnoli, 2004). See Baldi Antognini and Zagoraiou (2015) and Lemmas B.11–B.13 in Bugni, Canay, and Shaikh (2018) for more details.

Assumption A.2 can be either strengthen or weaken, depending on the nature of the treatment assignment mechanism and the experimental design. In the former case, we can replace A.2 *ii)* with the more restrictive $\mathcal{D}_N(s) = o_p(N(s)^{1/2})$. In the latter, one can weaken it by instead considering $\mathcal{D}_N(s) = o_p(N(s))$ for every $s \in \mathscr{S}$, where $N(s) = m(s) + n(s)$, e.g. Bugni, Canay, and Shaikh (2019); Zhang and Zheng (2020). We stick to our formulation because it *i)* covers the most common CAR schemes—certainly the ones in this paper—and *ii)* simplifies the asymptotic theory.

Remark 3. Alternatively, one may consider probability bounds for the overall and marginal imbalances (see footnote 4) and derive the asymptotic properties for Wald tests under sequential randomization algorithms (Hu and Hu, 2012) and the marginal procedures in Pocock and Simon (1975). See Corollary 3.1 and Theorem 3.3 in Ma, Hu, and Zhang (2015). Lastly, we can sidestep discretization and work with continuous covariates directly by assuming

$$N^{-1/2} \sum_{i=1}^{N} (2A_i - 1)\, Z_i \xrightarrow{\text{d}} \boldsymbol{\xi}\ ,$$

where $\boldsymbol{\xi}$ is a p-dimensional random vector with $\mathbb{E}(\boldsymbol{\xi}) = \mathbf{0}$. This last condition applies to CAR

designs that deal with continuous covariates, such as simple randomization, re-randomization (Morgan and Rubin, 2012), pair-wise sequential randomization (Qin et al., 2018), and Atkinson's D_A-biased coin design (Atkinson, 1982), thus encompassing a large class of model-based randomization methods that attain certain optimality criteria (e.g. Smith, 1984b,a; Baldi Antognini and Zagoraiou, 2011). ■

2.2 Testing Problem

Let $F_1(\cdot)$ and $F_0(\cdot)$ denote the distribution functions of random variables $Y(1)$ and $Y(0)$, respectively. We wish to test the hypothesis

$$H_0 : F_1 = F_0 \quad \text{vs} \quad H_1 : F_1 \neq F_0 \ . \tag{2}$$

One possible candidate for a test statistic for hypothesis (2) is the 2SKS test statistic. To fix notation, consider the empirical counterparts of F_1 and F_0 and denote

$$\hat{F}_1(y) = \frac{1}{m}\sum_{i=1}^{N} \mathbb{1}_{\{Y_i \leq y\}} A_i = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{X_i \leq y\}} \text{ and } \hat{F}_0(y) = \frac{1}{n}\sum_{i=1}^{N} \mathbb{1}_{\{Y_i \leq y\}}(1-A_i) = \frac{1}{n}\sum_{j=m+1}^{N} \mathbb{1}_{\{X_j \leq y\}} \ ,$$

as the empirical CDF of treatment and control groups, respectively. Thus, the 2SKS statistic is given by

$$K_{m,n}(\boldsymbol{X}) = \sup_y |V_{m,n}(y; \boldsymbol{X})| \ , \tag{3}$$

where

$$V_{m,n}(y; \boldsymbol{X}) = \sqrt{\frac{mn}{N}} \left(\hat{F}_1(y) - \hat{F}_0(y) \right) \tag{4}$$

is the classical two-sample empirical process

2.3 Construction of a Permutation Test

Before turning to the theoretical results, we first illustrate the construction of a permutation tests to asses H_0 in (2). To define the test, we introduce further notation. Define $\mathbf{G}_N$ as the set of all permutations π of $\{1, \ldots, N\}$, with $|\mathbf{G}_N| = N!$. Given $\boldsymbol{X} = \boldsymbol{x}$, recompute $K_{m,n}(\boldsymbol{x})$ for all permutations $\pi \in \mathbf{G}_N$ and denote by $K_{m,n}^{(1)}(\boldsymbol{x}) \leq K_{m,n}^{(2)}(\boldsymbol{x}) \leq \cdots \leq K_{m,n}^{(N!)}(\boldsymbol{x})$ the ordered values of $\{K_{m,n}(\boldsymbol{x}_\pi) \ : \ \pi \in \mathbf{G}_N\}$, where $\boldsymbol{x}_\pi$ denotes the action of $\pi \in \mathbf{G}_N$ on $\boldsymbol{x}$.

Let $k = N! - \lfloor N!\,\alpha \rfloor$ and define

$$M^+(\boldsymbol{x}) = \left|\{1 \le j \le N! : K^{(j)}_{m,n}(\boldsymbol{x}) > K^{(k)}_{m,n}(\boldsymbol{x})\}\right|$$
$$M^0(\boldsymbol{x}) = \left|\{1 \le j \le N! : K^{(j)}_{m,n}(\boldsymbol{x}) = K^{(k)}_{m,n}(\boldsymbol{x})\}\right| .$$

Using this notation, the permutation test is given by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 & K_{m,n}(\boldsymbol{x}) > K^{(k)}_{m,n}(\boldsymbol{x}) \\ a(\boldsymbol{x}) & K_{m,n}(\boldsymbol{x}) = K^{(k)}_{m,n}(\boldsymbol{x}) \\ 0 & K_{m,n}(\boldsymbol{x}) < K^{(k)}_{m,n}(\boldsymbol{x}) \end{cases}, \quad \text{where} \quad a(\boldsymbol{x}) = \frac{N!\,\alpha - M^+(\boldsymbol{x})}{M^0(\boldsymbol{x})} . \tag{5}$$

Alternatively, the permutation test rejects H_0 in (2) if $K_{m,n}(\boldsymbol{x})$ exceeds the upper α quantile of the permutation distribution:

$$\hat{R}^K_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\{K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}) \le t\}} . \tag{6}$$

We can interpret the permutation distribution as the conditional distribution of $K_{m,n}(\boldsymbol{X}_\pi)$ given $\boldsymbol{X}$, where π is a random permutation uniformly distributed over $\mathbf{G}_N$, and $\boldsymbol{X}_\pi = (X_{\pi(1)}, \ldots, X_{\pi(N)})$. To see why, we observe that $K_{m,n}(\boldsymbol{X}_\pi)$ and $K_{m,n}(\boldsymbol{X}_{\pi'})$ are equally likely for any $\pi, \pi' \in \mathbf{G}_N$, conditionally on $\boldsymbol{X}$ (Lehmann and Romano, 2005, Theorem 15.2.2).

Remark 4. The above construction of the permutation test can be computationally burdensome for moderately large N, which is typically the case in practice. In these scenarios, we may alternatively rely on a stochastic approximation without affecting the permutation test's theoretical properties by sampling permutations π from $\mathbf{G}_N$ with or without replacement. More formally, let $\hat{\mathbf{G}}_N = \{\pi_1, \ldots, \pi_M\}$, where π_1 is the identity permutation and $\pi_2, \ldots, \pi_M$ are i.i.d. uniform on $\mathbf{G}_N$. The same construction follows if we replace $\mathbf{G}_N$ with $\hat{\mathbf{G}}_N$, and the approximation is arbitrarily close for M sufficiently large (Romano, 1989, Section 4). From now on we focus on $\mathbf{G}_N$ while in practice we fall back on $\hat{\mathbf{G}}_N$. See also Algorithm 1 in Section 4.3. ■

3 The Adverse Effects of Stratification on Inference

We now demonstrate that when testing the null hypothesis of equality of distributions, balancing over covariates using CAR techniques has a detrimental effect on inference. This section's main result is that the asymptotic distribution of the 2SKS statistic depends on stratification, making it difficult, if not impossible, to obtain valid critical values. Moreover, we show that permutation-based inference is not exempt from this effect. Thus, naively relying on a permutation test that is incompatible with adaptive randomization to conduct inference can lead to

severe size distortions, and is therefore invalid, even in large samples.

3.1 Asymptotic Results under CAR

We begin by investigating the effects of CAR on the asymptotic behavior of the 2SKS test statistic. We introduce the following notations. Let

$$\mathbb{C}_1(y_1, y_2) = \lambda(1-\lambda)\{F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2)\} \ , \tag{7}$$

where $a \wedge b = \min\{a, b\}$. Let $\mathbb{G}_2$ and $\mathbb{G}_3$ be two Gaussian processes with mean zero and respective covariance structures given by

$$\begin{aligned}\mathbb{C}_2(y_1, y_2) = \sum_{s\in\mathscr{S}} p(s)\,\tau(s)\Big(&(1-\lambda)^2\,\mathbb{E}\left(m_1(y_1, Z_i)|S_1 = s\right)\mathbb{E}\left(m_1(y_2, Z_i)|S_1 = s\right) \\ &+ \lambda(1-\lambda)\,\mathbb{E}\left(m_1(y_1, Z_i)|S_1 = s\right)\mathbb{E}\left(m_0(y_2, Z_i)|S_1 = s\right) \\ &+ \lambda(1-\lambda)\,\mathbb{E}\left(m_1(y_2, Z_i)|S_1 = s\right)\mathbb{E}\left(m_0(y_1, Z_i)|S_1 = s\right) \\ &+ \lambda^2\,\mathbb{E}\left(m_0(y_1, Z_i)|S_1 = s\right)\mathbb{E}\left(m_0(y_2, Z_i)|S_1 = s\right)\Big) \ ,\end{aligned} \tag{8}$$

and

$$\begin{aligned}\mathbb{C}_3(y_1, y_2) = \lambda^2(1-\lambda)^2 \sum_{s\in\mathscr{S}} p(s)\Big(&\mathbb{E}[m_1(y_1, Z)|S = s]\,\mathbb{E}[m_1(y_2, Z)|S = s] \\ &+ \mathbb{E}[m_0(y_1, Z)|S = s]\,\mathbb{E}[m_0(y_2, Z)|S = s] - 2\,\mathbb{E}[m_1(y_1, Z)|S = s]\,\mathbb{E}[m_0(y_2, Z)|S = s]\Big) \ ,\end{aligned} \tag{9}$$

where for each $s \in \mathscr{S}$ and $a \in \{0, 1\}$, $m_a(y, Z) = F_a(y|Z) - F_a(y)$.

The following theorem describes the behavior of the 2SKS test statistic for treatment assignment mechanisms satisfying assumption A.2.

Theorem 1. *Suppose the distribution of the data satisfies assumption A.1 and that the treatment assignment is such that assumption A.2 holds. Then the two-sample empirical process* $\{V_{m,n}(y, \boldsymbol{X}) : y \in \mathbb{R}\}$ *converges weakly under the null hypothesis to* $\mathbb{H}(\cdot)$*. Here* $\mathbb{H}$ *is a tight Gaussian process with mean zero and covariance structure*

$$\mathbb{C}(\mathbb{H}(y_1), \mathbb{H}(y_2)) = \frac{1}{\lambda(1-\lambda)}\Big(\mathbb{C}_1(y_1, y_2) + \mathbb{C}_2(y_1, y_2) + \mathbb{C}_3(y_1, y_2)\Big) \ . \tag{10}$$

Furthermore, under the null hypothesis, $K_{m,n}$ *converges in distribution to* $K = \sup_y|\mathbb{H}(y)|$ *with CDF* $J(\cdot, F_1, F_0)$ *given by*

$$J(t, F_1, F_0) = \mathbb{P}_{F_1,F_0}\Big\{K \leq t\Big\} \ .$$

It is instructive to compare the key features of Theorem 1 with the particular case when covariates play no role in randomization. In simple randomized designs, the process $V_{m,n}(\cdot)$ converges weakly to $\mathbb{G}_1$, an F_0–Brownian bridge process under the null hypothesis (Van der Vaart and Wellner, 1996, Theorem 3.7.1). However, under the general assumptions on the treatment assignment mechanisms in Theorem 1, the asymptotic distribution of the process $V_{m,n}(\cdot)$ is no longer the Brownian bridge but rather a different Gaussian process. Indeed, we show in Appendix A that we can write the process $\mathbb{H}$ as the sum of three independent components

$$\mathbb{H}(y) = \underbrace{\mathbb{G}_1(y)}_{\text{standard}} + \underbrace{\mathbb{G}_2(y) + \mathbb{G}_3(y)}_{\text{shift due to stratification}} \quad .$$

We note that the first summand is the standard F_0–Brownian bridge process. However, the remaining terms in the preceding expression are generally different than 0, yielding a more complicated covariance structure (10). Thus, stratification leaves a mark on the asymptotic null distribution of the two-sample empirical process—the new asymptotic null distribution depends on the nature of the data generating process and the treatment assignment mechanism. We synthesize the ongoing discussion in the following corollary.

Corollary 1. *Under simple randomization, but otherwise under the conditions of Theorem 1, the covariance structure in (10) reduces to* $F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2)$.

Since no asymptotic critical values are available, one may wonder whether we can use data-dependent "critical values"—such as permutation-based critical values—instead of asymptotic ones, mainly because randomization comes from random treatment assignment, making intuitive sense to consider randomization inference for testing.

We argue that permutation-based inference is generally not valid in the presence of CAR schemes. To see why, we note that, in light of Theorem 1, the 2SKS statistic is not asymptotically pivotal. Therefore one can deduce that the corresponding permutation test fails to control the Type 1 error even asymptotically. This conclusion is an immediate consequence of the fact that the permutation distribution based on the 2SKS statistic behaves like the limit distribution *as if* the randomization was simple, not like the true unconditional limiting distribution under CAR. The following theorem due to Chung and Olivares (2020) formally presents this fact. Note that the null hypothesis is not assumed.

Theorem 2. *Consider testing the hypothesis (2). If assumptions A.1–A.2, then the permutation distribution (6) based on the 2SKS statistic is such that*

$$\sup_t \left| \hat{R}^K_{m,n}(t) - J_1(t) \right| \xrightarrow{\text{P}} 0 \ ,$$

where $J_1(\cdot)$ *denotes the CDF of* $K_1 \equiv \sup_y |\mathbb{G}_{\bar{P}}(y)|$*, where* $\mathbb{G}_{\bar{P}}$ *is a* $\bar{P}$*-Brownian bridge corresponding to the mixture distribution* $\bar{P}$ *given by*

$$P(y) = \sum_{s \in \mathscr{S}} p(s) \Big\{ \lambda F_1(y|S=s) + (1-\lambda) F_0(y|S=s) \Big\} \ . \tag{11}$$

The permutation test based on the 2SKS statistic under CAR fails to control the type 1 error rate, even in large samples, since the 2SKS statistic is not asymptotically pivotal—the limiting distribution depends on stratification. We confirm this phenomenon in the simulation studies in Section 5, where the empirical rejection probabilities under the null hypothesis are substantially different from the nominal level.

With this setup, our goal is to find an asymptotically valid permutation test for H_0. First, we introduce an exchangeable bootstrap approach to consistently estimating the 2SKS statistic's asymptotic null distribution under CAR. Second, we find a permutation test whose limiting rejection probability under the null hypothesis equals the nominal level in large samples. The next section formalizes these ideas.

4 Main Results: Restoring Asymptotic Validity

The main results in this section—Theorems 3 and 4—show that we can indeed develop asymptotically valid permutation test for (2) under CAR. These results depend on an insightful idea by Beran (1987, 1988), based on the inverse CDF property. More specifically, one can transform the original 2SKS statistic by its bootstrap CDF. Then, once we ensure the consistency of the bootstrap, the newly transformed statistic—which is referred to as *prepivoted*—is asymptotically uniformly distributed on $[0, 1]$, and thereby restoring the feasibility of an asymptotically distribution-free test statistic.

This section starts by presenting an exchangeable bootstrap method to approximate the limiting null distribution of the 2SKS statistic under CAR. We specialize in the Bayesian bootstrap (Rubin, 1981), but our conditions allow for different bootstrap weights. Then, we introduce the the new permutation test based on the prepivoted statistic (14). We show that the permutation test based on the prepivoted statistic has rejection probability that tends to α for testing equality of distributions under CAR.

4.1 Exchangeable Bootstrap under CAR

Before establishing the consistency of the exchangeable bootstrap, we introduce the following condition for the bootstrap weights.

A. 3. *For each N, let $(\omega_1, \ldots, \omega_N)$ be an exchangeable, nonnegative random vector independent of data $\{(Y_i, A_i, Z_i) : 1 \leq i \leq N\}$, such that the following conditions are satisfied under F_1 and F_0*

$$\sup_N \left\{ \int_0^\infty \sqrt{P\left(|\omega_1 - \bar{\omega}_N| > x\right)} dx \right\} < \infty,$$
$$N^{-1/2}\, \mathbb{E} \max_{1 \leq i \leq N} |\omega_i - \bar{\omega}_N| \xrightarrow{\mathrm{P}} 0,$$
$$N^{-1} \sum_{i=1}^N (\omega_i - \bar{\omega}_N)^2 \xrightarrow{\mathrm{P}} c^2 > 0 \ .$$

Throughout this paper, the bootstrap weights $\omega_1, \ldots, \omega_N$ are i.i.d. from the uniform Dirichlet distribution. This choice of weights leads to the so-called Bayesian bootstrap Rubin (1981), thus satisfying the conditions stated in assumption A.3 (Van der Vaart and Wellner, 1996, Section 3.6.2).[5] Instead of sampling each Y_i independently with replacement and equal probability $1/N$, the Bayesian bootstrap uses a posterior probability distribution centered at $1/N$ for each Y_i, but the probability of selection changes from sample to sample. Rubin (1981) shows that the Bayesian bootstrap procedure leads to a Dirichlet posterior distribution and is based on a conjugate prior for the Dirichlet.

Consider the *weighted bootstrap* analogues of the empirical CDF,

$$\hat{F}_1^\omega(y) = \frac{1}{m} \sum_{i=1}^N \omega_i \mathbb{1}_{\{Y_i \leq y\}} A_i \quad \text{and} \quad \hat{F}_0^\omega(y) = \frac{1}{n} \sum_{i=1}^N \omega_i \mathbb{1}_{\{Y_i \leq y\}} (1 - A_i) \ .$$

The two-sample weighted bootstrap empirical process is given by

$$V_{m,n}^\omega(y; \boldsymbol{X}) = \sqrt{\frac{mn}{N}} \left\{ \hat{F}_1^\omega(y) - \hat{F}_0^\omega(y) - \left(\hat{F}_1(y) - \hat{F}_0(y) \right) \right\} = \sqrt{\frac{mn}{N}} \left\{ \tilde{F}_1(y) - \tilde{F}_0(y) \right\} \tag{12}$$

where $\tilde{F}_a \equiv \hat{F}_a - \hat{F}_a^\omega$, $a \in \{0, 1\}$. The 2SKS based on it as

$$K_{m,n}^\omega(\boldsymbol{X}) = \sup_y \left| V_{m,n}^\omega(y, \boldsymbol{X}) \right| \ . \tag{13}$$

The following theorem states the consistency of the weighted bootstrap.

Theorem 3. *Suppose the distribution of the data satisfies assumption A.1 and that the treatment assignment is such that assumption A.2 holds. For each N, let $(\omega_1, \ldots, \omega_N)$ be weights satisfying assumption A.3. Then, conditionally on data, the process $\{V_{m,n}^\omega(y, \boldsymbol{X}) : y \in \mathbb{R}\}$ converges weakly under the null hypothesis to $\mathbb{H}(\cdot)$ in probability. Here $\mathbb{H}(\cdot)$ is a tight Gaussian process as in Theorem 1.*

[5] Alternative examples of weights satisfying assumption A.3 are the multinomial weights, multinomial replicates, and the wild bootstrap, to name a few.

Furthermore, conditionally on data, $K^{\omega}_{m,n}$ converges in distribution to $K = \sup_y |\mathbb{H}(y)|$ with CDF $J(\cdot, F_1, F_0)$ defined in Theorem 1.

To gain further intuition about the previous result, let $J_{m,n}(F_1, F_0)$ be the distribution of $K_{m,n}(\boldsymbol{X})$, and $J_{m,n}(\cdot, F_1, F_0)$ be the corresponding CDF defined by

$$J_{m,n}(t, F_1, F_0) = \mathbb{P}_{F_1,F_0}\Big\{K_{m,n} \leq t\Big\} \ .$$

Following Beran (1988), we define the prepivoted statistic as

$$T_{m,n}(\boldsymbol{X}) = J_{m,n}\left(K_{m,n}(\boldsymbol{X}), \tilde{F}_1, \tilde{F}_0\right) \ . \tag{14}$$

The previous theorem shows that the bootstrap CDF $J_{m,n}(\cdot, \tilde{F}_1, \tilde{F}_0)$ converges in probability to $J(\cdot, F_1, F_0)$ in supremum norm. Since $J_{m,n}(\cdot, F_1, F_0)$ itself converges to a continuous $J(\cdot, F_1, F_0)$ by Theorem 1, it follows that $T_{m,n}(\boldsymbol{X}) = J_{m,n}\left(K_{m,n}(\boldsymbol{X}), \tilde{F}_1, \tilde{F}_0\right)$ converges weakly to the uniform $[0, 1]$.

Remark 5. An alternative approach to the exchangeable bootstrap is the covariate-adaptive bootstrap (CAB), originally due to Shao, Yu, and Zhong (2010). In a nutshell, the CAB proceeds by first resampling $\mathbf{S}_N$ with replacement to generate a new vector of assignments $\mathbf{A}_N$, and then by resampling $\boldsymbol{X}$ with replacement for each cell defined by combinations of strata and treatment indicators. One benefits of the CAB is that CAB samples are cross-sectionally independent given data Zhang and Zheng (2020). However, researchers need to know the treatment assignment rule and the baseline covariates used in stratification to implement the CAB. While this knowledge is commonly available in most RCTs, this is not always the case when ethical considerations play a central role. For example, field experiments—like the one we consider in our empirical application—frequently hide pre-treatment characteristics to fulfill their IRB commitments to keep the subjects in the experiment anonymous, thus ensuring minimal risk (Duflo and Banerjee, 2017, Chapter 5). Our exchangeable bootstrap approach bypasses this difficulty. See Section 6 for more discussion. ■

4.2 Asymptotically Valid Permutation Test under CAR

We now turn to our key theoretical result. Let $\pi \in \mathbf{G}_N$ be a random permutation of $\{1, \ldots, N\}$ as in Section 2.3. The permutation distribution based on $T_{m,n}$ is given by

$$\hat{R}^{T}_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in \mathbf{G}_N} \mathbb{1}_{\left\{T_{m,n}(x_{\pi(1)}, \ldots, x_{\pi(N)}) \leq t\right\}} \ . \tag{15}$$

We seek the limiting behavior of $\hat{R}^T_{m,n}$ and its upper α-quantile, which we now denote $\hat{r}_{m,n}$, where

$$\hat{r}_{m,n}(1-\alpha) = \inf\{t : \hat{R}^T_{m,n}(t) \geq 1-\alpha\} \ .$$

The following theorem shows that the proposed test is asymptotically valid, *i.e.*, the permutation distribution based on $T_{m,n}(\boldsymbol{X})$ is asymptotically uniformly distributed on $[0,1]$. Consequently, the α-upper quantiles $\hat{r}_{m,n}$ can be used as "critical values" for the prepivoted statistic. Note that the null hypothesis is not assumed.

Theorem 4. *Consider testing the hypothesis (2). If assumptions A.1–A.3 hold with $\tau(s) = 0$ for all $s \in \mathscr{S}$, then the permutation distribution $\hat{R}^T_{m,n}(\cdot)$ of $T_{m,n}(\boldsymbol{X})$ defined in (15) satisfies*

$$\sup_{0\leq t\leq 1} \left|\hat{R}^T_{m,n}(t) - U(t)\right| \xrightarrow{\mathrm{P}} 0 \ ,$$

where $U(\cdot)$ is the CDF of the uniform distribution on $[0,1]$. Furthermore, $\hat{r}_{m,n}(1-\alpha) \xrightarrow{\mathrm{P}} 1-\alpha$.

Remark 6. From the construction of the permutation test in (5) based on $T_{m,n}(\boldsymbol{X})$, we have

$$\Pr\left\{T_{m,n}(\boldsymbol{X}) > \hat{r}_{m,n}\right\} \leq \mathbb{E}\left[\phi(\boldsymbol{X})\right] \leq \Pr\left\{T_{m,n}(\boldsymbol{X}) \geq \hat{r}_{m,n}\right\} \ .$$

Then, Theorem 4 implies $\mathbb{E}\left[\phi(\boldsymbol{X})\right] \to \alpha$ (Lehmann and Romano, 2005, Section 15.2.2). ■

Remark 7. There is no loss in power in using permutation critical values. To see why, let $r_{m,n}$ be the $1-\alpha$ quantile of the distribution of $T_{m,n}$. Typically the test based on $T_{m,n}$ rejects when $T_{m,n} > r_{m,n}$, where $r_{m,n}$ is nonrandom. We have that $r_{m,n} \to 1-\alpha$. Assume that $T_{m,n}$ weakly converges to some limit law $U'(\cdot)$ under some sequence of alternatives that are contiguous to some distribution satisfying the null hypothesis. Then the power of the test would tend to $1 - U'(U^{-1}(1-\alpha))$. Thus, under the premises of the preceding Theorems 3 and 4, we have that $\hat{r}_{m,n}$, obtained from the permutation distribution, satisfies $\hat{r}_{m,n} \xrightarrow{\mathrm{P}} 1-\alpha$. The same result follows under a sequence of contiguous alternatives, thus implying that the permutation test has the same limiting local power as the test which uses nonrandom critical values. ■

Remark 8. Bugni, Canay, and Shaikh (2018) consider a covariate-adaptive permutation test for testing equality of means under CAR. Unlike the standard construction in Section 2.3, the covariate-adaptive permutation test only permutes indices within strata, thus respecting stratification. However, we do not consider this approach mainly because we need to know the baseline covariates used in stratification—we permute data within strata. We cannot meet this requirement always, particularly when anonymity of experimental subjects may be at stake, like in the field experiment we consider in Section 6. ■

4.3 Implementation of the new Permutation Test

Given the previous sections' theoretical results, we now elaborate on some missing implementation details of the proposed test. In particular, the following algorithm below illustrates how the `RATest` package calculates the proposed permutation test. Note that we rely on a stochastic approximation to the permutation distribution as in Remark 4.

Algorithm 1

1. *Take a permutation of data* $\boldsymbol{X}_{\pi_j}$, $\pi_j \in \mathbf{G}_N$ *and calculate the 2SKS statistic,* $K_j \equiv K_{m,n}(\boldsymbol{X}_{\pi_j})$.
2. *For* $b = 1, \ldots, B$,
   (a) *Draw weights* $\omega_1^b, \ldots, \omega_N^b$ *from a uniform Dirichlet distribution.*
   (b) *Sample data according to the probabilities defined by the Dirichlet draws.*
   (c) *Use these resampled data to calculate the 2SKS statistic. Call this new statistic* $K_{j,b}^*$.
3. *The prepivoted statistic for* π_j *is the fraction of the values* $\{K_{j,b}^* : 1 \leq b \leq B\}$ *that are less than or equal to* K_j*, i.e., one minus the bootstrap p-value, given by*

$$T_{m,n}(\boldsymbol{X}_{\pi_j}) = J_{m,n}(K_j, \hat{F}_1, \hat{F}_0) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\left\{K_{j,b}^* \leq K_j\right\}} \ .$$

4. *Repeat Steps 1–3 for* $1 \leq j \leq M$*, and collect these values into* $\{T_{m,n}(\boldsymbol{X}_{\pi_j}) : 1 \leq j \leq M\}$.
5. *The permutation test rejects the null hypothesis if the observed prepivoted statistic exceeds the upper-*α *quantile of the permutation distribution:*

$$\hat{R}_{m,n}^T(t) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}_{\left\{T_{m,n}(\boldsymbol{X}_{\pi_j}) \leq t\right\}} \ .$$

Remark 9. We may characterize drawing weights from a uniform Dirichlet distribution by drawing from the N-dimensional unit simplex. Alternatively, we can achieve this by drawing Gamma $(1, 1)$ distributed numbers and normalizing these to sum to 1. However, since a Gamma $(1, 1)$ is equivalent to an Exp(1) distribution, we can define the weights as $\omega_i = \eta_i/\bar{\eta}$, $1 \leq i \leq N$, where $\eta_i \sim \text{Exp}(1)$ and $\bar{\eta} = N^{-1}\sum_{i=1}^{N} \eta_i$. ■

Remark 10. In practice, Algorithm 1 can be expensive to compute as the sample size increases. It involves resampling twice—once with replacement for the bootstrap, and once without it, for the permutation test. Whether there is some computationally more efficient algorithm to compute our test is something we leave as an interesting topic for future research. ■

5 Monte Carlo Experiments

In this section, we study the proposed test's finite sample performance through a Monte Carlo exercise compared to two other methods, namely the 2SKS test and the permutation test based on the same 2SKS statistic. The main focus is on the finite-sample implications of the asymptotic validity of our permutation test. To accomplish our goal, we adhere to the design in Bugni, Canay, and Shaikh (2018). The following rule governs the potential outcomes:

$$Y_i(a) = \mu_a + m_a(Z_i) + \sigma_a(Z_i)\varepsilon_i(a), \;\; a \in \{0,1\} \;, \; 1 \leq i \leq N \;,$$

where $\{Z_i, \varepsilon_i(1), \varepsilon_i(0) \;:\; 1 \leq i \leq N\}$ are i.i.d. This gives rise to the observed outcomes

$$\begin{aligned} Y_i &= \mu_0 + m_0(Z_i) + \{(\mu_1 - \mu_0) + (m_1(Z_i) - m_0(Z_i))\} A_i + u_i \\ u_i &= \sigma_1(Z_i)\varepsilon_i(1)A_i + \sigma_0(Z_i)\varepsilon_i(0)(1 - A_i) \;. \end{aligned}$$

We compare our permutation test with the following procedures:

2SKS: This test is the classical 2SKS test described in Section 3, *i.e.*, this test does not take into account CAR. We use the asymptotic approximation to its distribution under the null hypothesis (Simard and L'Ecuyer, 2011, Section 3). We rely on `ks.test`, the base `R` implementation to compute the p-value of the 2SKS test. See Drew, Glen, and Leemis (2000) and Marsaglia, Tsang, and Wang (2003) for a review of the computational aspects involved in calculating the exact 2SKS distribution for some of the most popular existing methods.

Naive: This test is the permutation test of Section 3 based on the 2SKS statistic. We call it "naive" in the same spirit as in Bugni, Canay, and Shaikh (2018), *i.e.* because this test ignores the effects of stratification on inference. We sample 1000 permutations for the stochastic approximation of the permutation distribution (see Remark 4). See the `R` package `RATest` for additional documentation.

5.1 Size

Arguing as in Bugni, Canay, and Shaikh (2018), we consider the following two models to investigate the empirical size:

Model 1 (Linear Model): Let $Z_i \sim$Beta$(2,2)$, $\sigma_0(Z_i) = 1$, $\sigma_1(Z_i) = \sigma_1$, $\varepsilon_i(1) \sim \mathcal{N}(0,1)$, $\varepsilon_i(0) \sim \mathcal{N}(0,1)$, and $m_1(Z_i) = m_0(Z_i) = \gamma Z_i$.

Model 2 (Non-linear, t distribution, homogeneous): Let $Z_i \sim$Unif$(-2,2)$, $\sigma_0(Z_i) = Z_i^2$,

$\sigma_1(Z_i) = Z_i^2\sigma_1$, $\varepsilon_i(1) \sim \frac{1}{3}t_3$, $\varepsilon_i(0) \sim \frac{1}{3}t_3$, and

$$m_1(Z_i) = m_0(Z_i) = \begin{cases} \gamma Z_i & \text{if} \quad Z_i \in [-1,1] \\ \gamma(2 - Z_i^2) & \text{otherwise.} \end{cases} .$$

Table 1 shows the rejection probabilities under the null hypothesis at $\alpha = 0.05$, *i.e.*, we impose the restrictions $\sigma_1 = 1$ and $\mu_1 = \mu_0 = 0$. In our simulations we use $5,000$ replications and sample size $N = 200$. For each model, different combinations of target proportions $\lambda \in \{0.5, 0.7\}$ and strata $|\mathscr{S}| \in \{4, 10\}$ give rise to four parameter configurations. When $\lambda = 0.5$, we consider four different CAR schemes—simple randomization (SRS), covariate-adaptive Wei's biased-coined design (WEI) with $\varphi(x) = 0.5(1 - x)$, covariate-adaptive Efron's biased-coined design (BCD) with $\gamma = 0.75$, and stratified block randomization (SBR). See Appendix C for more details.

Table 1: Size of $\alpha = 0.05$ tests $H_0 : F_1 = F_0$.

		$\lvert\mathcal{S}\rvert = 4$, $\lambda = 0.5$, $\gamma = 1$, $\sigma_1 = 1$			$\lvert\mathcal{S}\rvert = 10$, $\lambda = 0.5$, $\gamma = 1$, $\sigma_1 = 1$		
Model	CAR	2SKS	Naive PermTest	PermTest	2SKS	Naive PermTest	PermTest
	SRS	0.0532	0.0486	0.0443	0.0486	0.0508	0.0560
1	WEI	0.0250	0.0226	0.0328	0.0216	0.0220	0.0307
	BCD	0.0144	0.0132	0.0260	0.0142	0.0146	0.0328
	SBR	0.0118	0.0130	0.0273	0.0122	0.0102	0.0300
	SRS	0.0496	0.0476	0.0580	0.0478	0.0526	0.0370
2	WEI	0.0444	0.0433	0.0406	0.0144	0.0144	0.0267
	BCD	0.0414	0.0408	0.0520	0.0074	0.0074	0.0172
	SBR	0.0334	0.0407	0.0465	0.0068	0.0060	0.0124
		$\lvert\mathcal{S}\rvert = 4$, $\lambda = 0.7$, $\gamma = 1$, $\sigma_1 = 1$			$\lvert\mathcal{S}\rvert = 10$, $\lambda = 0.7$, $\gamma = 1$, $\sigma_1 = 1$		
1	SRS	0.0492	0.0510	0.0460	0.0526	0.0468	0.0480
	SBR	0.0128	0.0146	0.0306	0.0102	0.0116	0.0280
2	SRS	0.0528	0.0490	0.0500	0.0452	0.0506	0.0500
	SBR	0.0392	0.0400	0.0446	0.0060	0.0066	0.0108

Rejection probabilities based on 5000 replications for the three tests defined in the text, four different CAR schemes, and two data generating processes. The symbols 2SKS, Naive PermTest, and PermTest stand for the classical 2SKS test, the permutation test based on the classical 2SKS, and the proposed permutation test robust to CAR, respectively. $N = 200$ across experiments. We use 1000 permutations for the stochastic approximation of the permutation distribution, and 1000 weighted bootstrap samples.

All three tests perform as expected under simple randomization. These tests control the type 1 error rate in this setup, so the numerical discrepancies from the nominal size are due to simulation noise. We note that the 2SKS test under rejects quite severely in Model 1, while it suffers from modest size distortions for Model 2 under WEI and BCD schemes. However, the size distortions increase when the number of strata increases, regardless of the randomization scheme or the model generating the outcomes. Meanwhile, the naive permutation test exhibits

considerable size distortions in Model 1 but performs reasonably well under rejection under Model 2. Similar to the 2SKS test's behavior, the naive permutation test performs very poorly when the number of strata increases, delivering rejection probabilities considerably below the nominal level.[6]

In contrast, our permutation test outperforms the existing alternatives across all specifications considered in our numerical exercise. For both models, the size is close to the nominal level when the treatment assignment mechanism follows WEI and BCDschemes, and the number of strata is small. However, the size is more distorted under SBR, particularly when the number of strata increases.[7]

6 Empirical Illustration

To illustrate the proposed method in this paper, we present a reappraisal of the field experiment by Butler and Broockman (2011) about the effect of race on state legislators' responsiveness to help their constituents register to vote during the 2008 elections in the United States. One may observe legislators engaging in discrimination based on race for at least two reasons. First, legislators may better represent those who share their characteristics. Second, this behavior may arise due to strategic partisanship, *i.e.*, legislators appeal primarily to constituents who are likely to vote for them.[8] To assess the interconnection between political discrimination and representation, the authors conduct an experiment involving $4,859$ U.S. state legislators who received fictitious emails from a constituent with either a commonly regarded Black or White name. These names were randomly assigned using stratified block randomization to balance over baseline covariates, namely the state, legislative chamber, political party, and whether the legislator was up for reelection. The authors also randomly signal voters' partisanship by asking about Democratic primary elections, Republican primary elections, or primary elections without explicitly mentioning any party. The final sample contains the $4,859$ emails, including whether the state legislator responded at all, the treatment indicator, and partisanship signal. See Butler and Broockman (2011) for a more detailed description of the data, summary statistics, and theoretical background on this topic.

A defining characteristic of this field experiment is that it may cause reputational harm

[6] We observe similar size distortions when we consider the so-called covariate-adaptive permutation test, *i.e.*, when we permute the data within stratum, and therefore we omit them here. See Rosenberger and Lachin (2015, Chapter 9) and Bugni, Canay, and Shaikh (2018, Remark 4.14) for more details and discussion.

[7] The average bootstrap sample contains roughly 63.2% of the original observations and omits 26.8%. To see why, observe that the probability that a particular observation is *not* chosen from a set of N observations is $1 - 1/N$, so the probability that the observation is not chosen N times is $(1 - 1/N)^N$, which converges to $1/\mathrm{e} \approx 0.368$ as $N \to \infty$. This may affect the performance of the 2SKS statistic, especially for the sample size considered in our numerical exercise.

[8] For example, Black constituents are far more likely to align with the Democratic party—84% of the Black voters registered as of 2017 (Pew Research Center, 2018).

given the study's subject. As a result, the public data file does not include the control variables because researchers pledge to the ethics committee to keep the legislators in the experiment anonymous. One of our permutation test's main advantages is that we do not require knowledge about strata to implement it, thus providing a flexible framework to perform asymptotically valid permutation inference for the hypothesis of interest in the presence of CAR—like the one used in this experiment—while simultaneously safeguarding anonymity. This characteristic makes our approach an attractive one to study randomized experiments involving public officials.

Table 2 shows the empirical results.[9] The first three columns report response rates when we do not signal the constituent's partisanship. Meanwhile, the remaining three columns randomize the constituent's partisanship signal in the letter. Column 1 reports a statistically significant mean difference of 5.1% in the response rates, where constituents with putatively Black names receive fewer responses than their White counterparts. However, this difference in responses disappears once we signal partisanship (column 4).

Table 2: Response Rates: Overall and Party-specific Effects.

	No Partisanship Signal			Partisanship Signal		
	Overall	Republican Legislator	Democratic Legislator	Overall	Republican Legislator	Democratic Legislator
Black Alias	55.3%	58.9%	52.4%	55.7%	56%	51.8%
	$m = 806$	$m = 360$	$m = 446$	$m = 1622$	$m = 723$	$m = 723$
White Alias	60.5%	67.0%	55.1%	55.8%	60.8%	55.6%
	$n = 812$	$n = 364$	$n = 448$	$n = 1619$	$n = 899$	$n = 896$
Race Differential	−5.1%	−8.1%	−2.7%	−0.1%	−4.8%	3.7%
	$p = 0.04$	$p = 0.04$	$p = 0.42$	$p = 0.95$	$p = 0.12$	$p = 0.11$
Equality of Distributions	$p = 0.02$	$p = 0.10$	$p = 0.41$	$p = 0.93$	$p = 0.16$	$p = 0.11$

This table reports response rates in percentage points as a result of randomized putatively Black or White aliases. The first three columns provide response rates when the constituent's partisanship is not signaled, and the remaining three when the constituent's partisanship is signaled. The label "Republican Legislator" indicates the subsample of republican representatives, and similarly for "Democratic Legislator." The last two rows report p-values for two-tailed t–tests for equality of means and our permutation test for equality of distributions between experimental groups, respectively. We adjust these p–values to account for multiple hypothesis testing following Holm (1979) procedure. We use 1000 permutations for the stochastic approximation of the permutation distribution. When partishanship is not signaled, we use 1000 weighted bootstrap samples, otherwise 600 due to memory storage.

To test whether state legislators respond more favorably to voters who, based on their race, are more likely to be of the same political party, the columns 2–3 condition on legislator's party affiliation. We observe a statistically significant higher response rate to the White alias than the Black alias when the legislator is Republican. In contrast, there is no statistically

[9]We adjust the p-values to account for multiple hypothesis testing following Holm (1979) method. For unadjusted p-values, see Butler and Broockman (2011, Tables 1–3). See Chung and Olivares (2020, Section 4) for a discussion about multiple testing adjustments for permutation based inference for hypotheses like the one we consider here.

significant mean difference in response rates between Black and White aliases when the legislator is Democratic. Lastly, since one may argue that legislators respond more favorably to co-partisans, columns 4–6 show response rates for Republican and Democratic legislators when we signal partisanship. We can see that there is no statistically significant mean difference in response rates between experimental groups when we signal partisanship regardless of the legislator's party affiliation.

Our permutation test complements these findings in several important ways. First—when we do not signal partisanship—we reject the hypothesis that the response distributions between experimental groups are the same in the overall case and when the legislator is from the Democratic party (columns 1 and 4). On the other hand, when the legislator is Republican, our permutation test fails to reject the hypothesis of equality of distributions. Second, when we signal partisanship, our permutation test fails to reject the null hypothesis that the response distributions between aliases are equal across specifications.

Butler and Broockman (2011) investigate potential heterogeneity in the treatment effect to shed some light on legislators' responsiveness when they receive the partisanship signal. In particular, they show that legislators from both parties discriminate at similar rates once we take race into account—White Democrats and White Republicans respond more often to White aliases, and their response rates are statistically indistinguishable. These findings conflict with our conclusions when we signal partisanship. However, once we adjust for multiple hypothesis testing, Butler and Broockman (2011, Table 3) initial results are no longer statistically significant.

Thus, the conclusions based on our proposed test suggest a clear pattern: legislators show more responsiveness to those constituents who, based on their race, are *believed* to be of the same party (no signal), but the response rates to each racial alias are indistinguishable from one another once the uncertainty disappears (signal).

7 Conclusions

This paper introduces an asymptotically robust permutation test for testing equality of distributions under CAR, that is, our permutation test has rejection probability that tends to α. From a theoretical point of view, stratifying impacts inference negatively and may lead to severe size distortions. Our first result shows that the limiting rejection probability of the standard 2SKS test can be substantially below its nominal level. We then show that this problem carries over to permutation-based inference indeed. Our second result establishes that in this setup, the permutation test that does not account for CAR fails to control the type 1 error rate, even in large samples. To demonstrate the quantitative importance of this phenomenon,

we present simulation evidence showing that the 2SKS and permutation tests are not reliable procedures for the testing problem of interest—the empirical rejection probabilities under the null hypothesis are shockingly different from the nominal level.

This paper's main contribution—Theorems 3 and 4—shows that we can indeed develop asymptotically valid permutation test for testing equality of distributions under CAR. Our main results exploit Beran (1987, 1988)'s idea and transform the initial 2SKS statistic by its bootstrap CDF. We establish the consistency of the exchangeable bootstrap under CAR in Theorem 3. Then, the transformed statistic—also called prepivoted—becomes asymptotically uniformly distributed on $[0, 1]$, effectively removing the effect of stratification. We show in Theorem 4 that the permutation test based on the prepivoted statistic has rejection probability that tends to α for testing equality of distributions under CAR.

Our theoretical and simulation results imply that the size control could be improved, often notably, outperforming the existing alternatives. Therefore, we recommend that researchers use the permutation test we develop in this paper for testing equality of distributions when randomization is covariate-adaptive. We also provide open-source software implementation, the `RATest` R package, to apply our proposed method straightforwardly. We illustrate our method's empirical relevance by revisiting a field experiment by Butler and Broockman (2011) about the effect of race on state legislators' responsiveness to help their constituents register to vote during the 2008 elections in the United States.

Acknowledgments

I am deeply indebted to my main advisers Eunyi Chung, Roger Koenker, Dan Bernhardt, and Xiaofeng Shao for their continuous guidance, support, patience, and encouragement. I would also like to thank seminar participants at Boston University, ITAM, and UIUC for useful comments and feedback that led to considerable improvement of the paper. All errors are my own.

Appendix

NOTATION: The classes $\mathscr{F}$ in all of the applications in this Appendix are collections of indicator functions of lower rectangles in $\mathbb{R}$. Thus, the empirical processes in this paper can be viewed as random maps into $\ell^\infty(\mathscr{F})$—the space of all bounded functions on $\mathbb{R}$ equipped with the uniform norm—and weak convergence is understood as convergence in distribution in $\ell^\infty(\mathscr{F})$. We are going to assume that the class $\mathscr{F}$ is pointwise measurable (Van der Vaart and Wellner, 1996, Example 2.3.4), ruling out measurability problems with regards suprema.

Throughout this appendix, if ξ is a random variable defined on a probability space $(\Omega, \mathscr{B}, P)$, it is assumed that $\xi_1, \ldots, \xi_N$ are coordinate projections on the product space $(\Omega^N, \mathscr{B}^N, P^N)$, and the expectations are computed for P^N. If auxiliary variables—independent of the ξs—are involved, we use a similar convention. In that case, the underlying probability space is assumed to be of the form $(\Omega^N, \mathscr{B}^N, P^N) \times (\mathcal{Z}, \mathscr{C}, Q)$, with $\xi_1, \ldots, \xi_N$ equal to the coordinate projections on the first N coordinates and the additional variables depending only on the N+1st coordinate.

Symbols $\mathcal{O}_p(1)$ and $o_p(1)$ stand for being bounded in probability and convergence to zero in probability, respectively. All vector are column vectors. We use $\lfloor \cdot \rfloor$ to denote the largest smaller integer. We use $\xrightarrow{\mathrm{P}}$ to denote convergence in probability, and $\xrightarrow{\mathrm{d}}$ to denote convergence in distribution, respectively. For two random variables ξ and η, write $\xi \stackrel{\mathrm{d}}{=} \eta$ if they have the same distribution.

A Proof of the Main Results

A.1 Proof of Theorem 1

We begin the proof by noting some preliminary facts which will be useful in the analysis of the asymptotic behavior of $K_{m,n}$. As a first step, develop

$$V_{m,n}(y;\boldsymbol{X}) = \sqrt{\frac{mn}{N}}\left\{\hat{F}_1(y) - \hat{F}_0(y)\right\}$$

under the null hypothesis as

$$\begin{aligned}
V_{m,n}(y;\boldsymbol{X}) &= \sqrt{\frac{mn}{N}}\left\{\left(\hat{F}_1(y) - F_1(y)\right) - \left(\hat{F}_0(y) - F_0(y)\right)\right\} \\
&= \sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)A_i - \frac{1}{n}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(0)\leq y\}} - F_0(y)\right)(1-A_i)\right\} \\
&= \left(\frac{\sqrt{mn}}{N}\right)\frac{1}{\sqrt{N}}\left\{\frac{N}{m}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)A_i - \frac{N}{n}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(0)\leq y\}} - F_0(y)\right)(1-A_i)\right\} \\
&= \left(\frac{\sqrt{mn}}{N}\right)\frac{1}{\sqrt{N}}\left\{\left(\frac{\mathcal{D}_N}{N}+\lambda\right)^{-1}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)A_i\right. \\
&\left. - \left(1-\frac{\mathcal{D}_N}{N}-\lambda\right)^{-1}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(0)\leq y\}} - F_0(y)\right)(1-A_i)\right\} \\
&= \left(\frac{\sqrt{mn}}{N}\right)\left(\left(\frac{\mathcal{D}_N}{N}+\lambda\right)^{-1}\left(1-\frac{\mathcal{D}_N}{N}-\lambda\right)^{-1}\right)\times \\
&\frac{1}{\sqrt{N}}\left\{\left(1-\frac{\mathcal{D}_N}{N}-\lambda\right)\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)A_i - \left(\frac{\mathcal{D}_N}{N}+\lambda\right)\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(0)\leq y\}} - F_0(y)\right)(1-A_i)\right\} \\
&= M_{N,1}\Big(M_{N,2}(y) + M_{N,3}(y)\Big)
\end{aligned}$$

where

$$M_{N,1} = \left(\frac{\sqrt{mn}}{N}\right)\left(\frac{\mathcal{D}_N}{N}+\lambda\right)^{-1}\left(1-\frac{\mathcal{D}_N}{N}-\lambda\right)^{-1}$$
$$M_{N,2}(y) = \frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{N}\left((1-\lambda)\left(\mathbb{1}_{\{Y_i(1)\leq y\}}-F_1(y)\right)A_i - \lambda\left(\mathbb{1}_{\{Y_i(0)\leq y\}}-F_0(y)\right)(1-A_i)\right)\right\}$$
$$M_{N,3}(y) = -\frac{\mathcal{D}_N}{\sqrt{N}}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\left(\mathbb{1}_{\{Y_i(1)\leq y\}}-F_1(y)\right)A_i + \left(\mathbb{1}_{\{Y_i(0)\leq y\}}-F_0(y)\right)(1-A_i)\right)\right\}$$

Assumption 2 (ii) implies that $M_{N,1} \xrightarrow{\text{P}} (\lambda(1-\lambda))^{-1/2}$ as $N \to \infty$. Similarly, assumptions A.1, A.2 (ii) and Lemma B.5, allow us to conclude that

$$\sup_y |M_{N,3}(y)| \xrightarrow{\text{P}} 0, \quad \text{as} \quad N \to \infty$$

under the null hypothesis. Moreover, Lemma B.1 shows that $M_{N,2}(\cdot)$ weakly converges to $\mathbb{G}_1(\cdot)+\mathbb{G}_2(\cdot)+\mathbb{G}_3(\cdot)$, where $(\mathbb{G}_1,\mathbb{G}_2,\mathbb{G}_3)(\cdot)$ are three independent Gaussian processes with covariance functions $\mathbb{C}_1(y_1,y_2)$, $\mathbb{C}_2(y_1,y_2)$, and $\mathbb{C}_3(y_1,y_2)$ respectively. Therefore $V_{m,n}(\cdot)$ converges weakly in $\ell^\infty(\mathscr{F})$ under the null hypothesis to a tight Gaussian process $\mathbb{H}(\cdot)$; it has mean zero with covariance structure:

$$\mathbb{C}_H(y_1,y_2) = \frac{1}{\lambda(1-\lambda)}\big(\mathbb{C}_1(y_1,y_2)+\mathbb{C}_2(y_1,y_2)+\mathbb{C}_3(y_1,y_2)\big)\,.$$

This concludes the proof of the first part of the theorem. Note that the maps $v \to \|v\|$ from $\ell^\infty(\mathscr{F})$ into $\mathbb{R}$ are continuous with respect to the supremum norm. Then, a direct application of the continuous mapping theorem (Van der Vaart, 2000, Theorem 18.11) yields the final result. This finishes the proof.

A.2 Proof of Corollary 1

Consider the setup and notation of Theorem 1. Under simple randomization, $m_1(y|Z) = m_0(y|Z) = 0$ for every y, where $m_a(y|Z)$, $a \in \{0,1\}$ are defined in (B.3)–(B.2). This implies

that the covariances $\mathbb{C}_2$ and $\mathbb{C}_3$ in (8)–(9) are zero too. Therefore,

$$\begin{aligned}\mathbb{C}_H(y_1, y_2) &= \frac{1}{\lambda(1-\lambda)}\Big(\mathbb{C}_1(y_1, y_2) + \mathbb{C}_2(y_1, y_2) + \mathbb{C}_3(y_1, y_2)\Big) \\ &= \frac{1}{\lambda(1-\lambda)}\Big(\lambda(1-\lambda)\left(F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2)\right)\Big) \\ &= F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2) \ ,\end{aligned}$$

as desired.

A.3 Proof of Theorem 2

The proof follows closely the arguments in the proof of Chung and Olivares (2020, Theorem A.2). Independent of the $\boldsymbol{X}$, let $(\pi(1), \ldots, \pi(N))$ and $(\pi'(1), \ldots, \pi'(N))$ be two independent random permutations of $\{1, \ldots, N\}$. We will denote $\boldsymbol{X}_\pi = (X_{\pi(1)}, \ldots, X_{\pi(N)})$; $\boldsymbol{X}_{\pi'}$ is defined the same way with π replaced by π'.

We seek to show that

$$\Big(K_{m,n}(\boldsymbol{X}_\pi), K_{m,n}(\boldsymbol{X}_{\pi'})\Big) \xrightarrow{\text{d}} (K_1, K_1') \ , \tag{A.1}$$

where K_1 and K_1' are independent with common CDF $J_1(\cdot)$. Then Hoeffding's Condition (Lehmann and Romano, 2005, Theorem 15.2.3) implies that

$$\sup_t \left|\hat{R}^K_{m,n}(t) - J_1(t)\right| \xrightarrow{\text{P}} 0 \ ,$$

completing the proof of the theorem. In the following, we prove (A.1) in two steps.

Step 1. Apply the coupling construction of Chung and Romano (2013) as described in Appendix D. More specifically, couple data $\tilde{\boldsymbol{X}}$ with an auxiliary sample of N i.i.d. observations $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_N)$ from the mixture distribution $\bar{P}$ with

$$\bar{P}(y) = \sum_{s \in \mathscr{S}} p(s)\left\{\lambda F_1(y|S=s) + (1-\lambda)F_0(y|S=s)\right\} \ .$$

See Appendix D for a detailed exposition of the coupling construction and notation.

Step 2. We now argue that the permutation distribution based on $\boldsymbol{X}$ should behave approximately like the behavior of the permutation distribution based on $\bar{\boldsymbol{X}}$. In view of the arguments in the proof of Lemma 5.1 in Chung and Romano (2013), it suffices to verify the following two conditions

$$\left(K_{m,n}(\bar{\boldsymbol{X}}_{\pi}), K_{m,n}(\bar{\boldsymbol{X}}_{\pi'})\right) \xrightarrow{\mathrm{d}} (K_1, K_1') \tag{A.2}$$

$$K_{m,n}(\bar{\boldsymbol{X}}_{\pi,\pi_0}) - K_{m,n}(\boldsymbol{X}_{\pi}) \xrightarrow{\mathrm{P}} 0 \ , \tag{A.3}$$

where the permutation π_0 is properly defined in Appendix D. Condition (A.2) follows by the same reasoning as in the proof of Chung and Olivares (2020, Lemma B.1).

To show (A.3), we first construct an auxiliary process $\tilde{V}_{m,n}$ that is stochastically equivalent to $V_{m,n}(\cdot;\boldsymbol{X})$ in the wide sense *i.e.* they have the same finite-dimensional distributions. Independently for each $s \in \mathscr{S}$ and independently of $(\mathbf{A}_N, \mathbf{S}_N)$, let $\{Y_i^s(1), Y_i^s(0) : 1 \leq i \leq N\}$ be i.i.d. with marginal distribution equal to the distribution of $(Y_i(1), Y_i(0))|\, S_i = s$.

The auxiliary process $\tilde{V}_{m,n}$ is thus given by

$$\tilde{V}_{m,n}(y) \equiv \sqrt{\frac{mn}{N}} \sum_{s\in\mathscr{S}} \left\{ \frac{1}{m} \sum_{i=\mathscr{N}_N(s)+1}^{\mathscr{N}_N(s)+m(s)} \mathbb{1}_{\{Y_i^s(1)\leq y\}} - \frac{1}{n} \sum_{i=\mathscr{N}_N(s)+m(s)+1}^{\mathscr{N}_N(s)+m(s)+n(s)} \mathbb{1}_{\{Y_i^s(0)\leq y\}} \right\} \ ,$$

where $\mathscr{N}_N(s) \equiv \sum_{i=1}^N \mathbb{1}_{\{S_i<s\}}$ for each s. Intuitively, the auxiliary process $\tilde{V}_{m,n}$ orders units by strata, and then by $A_i = 1$ first and $A_i = 0$ second within strata. This construction—combined with the i.i.d. assumption on data and assumption A.2 *i)*—ensures the distribution of $V_{m,n}(\cdot;\boldsymbol{X})$ is the same as $\tilde{V}_{m,n}$, since

$$\left\{V_{m,n}(y;\boldsymbol{X})|\{A_i,S_i\}_{i=1}^N\right\} \overset{\mathrm{d}}{=} \left\{\tilde{V}_{m,n}(y)|\{A_i,S_i\}_{i=1}^N\right\} \ .$$

Furthermore, $V_{m,n}(\cdot;\tilde{\boldsymbol{X}})$ from the coupling construction (appendix D), and $\tilde{V}_{m,n}$ have the same distribution by the same reasoning. With this in mind, a sufficient condition for (A.3) is given by showing $\mathcal{V}_{m,n}(y) = V_{m,n}(y;\bar{\boldsymbol{X}}_{\pi\pi_0}) - V_{m,n}(y;\tilde{\boldsymbol{X}}_{\pi}) \xrightarrow{\mathrm{P}} 0$ uniformly over $y \in \mathbb{R}$.

To prove the desired result, it is useful to rewrite $\mathcal{V}_{m,n}$ as follows

$$\mathcal{V}_{m,n}(y) = \sqrt{\frac{n}{mN}} \left\{ \sum_{i=1}^N \left(\mathbb{1}_{\{\bar{X}_{\pi_0(i)}\leq y\}} - \mathbb{1}_{\{\tilde{X}_i\leq y\}} \right) W_{\pi(i)} \right\} \ , \tag{A.4}$$

where W_i is defined as

$$W_i = \begin{cases} 1 & \text{if } \pi(i) \leq m \\ -\frac{m}{n} & \text{if } \pi(i) > m \end{cases}, \; 1 \leq i \leq N \; .$$

The argument provided here follows closely the arguments in the proof of Chung and Olivares (2020, Lemma B.2) and the coupling construction in appendix D. First, we note that $\mathbb{E}\left[\mathcal{V}_{m,n}(y)\right] = 0$ by independence of data and $W_{\pi(i)}$. To investigate the variance, observe that the elements in $\bar{\boldsymbol{X}}_{\pi_0}$ and $\tilde{\boldsymbol{X}}$ are the same except for $\mathcal{C}$ of them (see appendix D for more details). This makes all the terms in the difference $\mathcal{V}_{m,n}(y)$ zero, except for at most $\mathcal{C}$ of them. Conditioning on the random drawing of indices in the coupling construction—hence conditioning on $\mathcal{C}$ and π_0—and on the permutation π, the variance of $\mathcal{V}_{m,n}(y)$ is determined by

$$\mathbb{V}\left[\mathcal{V}_{m,n}(y)\right] = \mathbb{E}\left[\mathbb{V}\left(\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right)\right] + \mathbb{V}\left[\mathbb{E}\left(\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right)\right] \tag{A.5}$$

by the law of total variance. We claim that both terms in previous display are zero, asymptotically. Note that the conditional variance in the first term in (A.5) is bounded above

$$\mathbb{V}\left[\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right] = \frac{n}{Nm}\mathcal{C}\,\mathbb{V}\left[W_{\pi(i)}\left(\mathbb{1}_{\{\bar{X}_{\pi_0(i)} \leq y\}} - \mathbb{1}_{\{\tilde{X}_i \leq y\}}\right) \middle| \mathcal{C}, \pi, \pi_0\right] \leq \frac{n}{m}\frac{\mathcal{C}}{N}\mathcal{O}(1) \; .$$

We show in (D.3) that $\mathbb{E}(\mathcal{C}/N) \leq N^{-1/2}$ and so the first term on the right hand side of (A.5) converges to 0. Another application of the law of total variances applied to the second term in (A.5) yields

$$\mathbb{V}\left[\mathbb{E}\left(\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right)\right] = \mathbb{E}\left\{\mathbb{V}\left[\mathbb{E}\left(\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right) \middle| \mathcal{C}, \pi_0\right]\right\} + \mathbb{V}\left\{\mathbb{E}\left[\mathbb{E}\left(\mathcal{V}_{m,n}(y) \middle| \mathcal{C}, \pi, \pi_0\right) \middle| \mathcal{C}, \pi_0\right]\right\} .$$

Let $\mathcal{A}$ be the number of observations among those $\mathcal{C}$ observations that have $W_{\pi(i)} = 1$. Conditioning on the random drawing of indices in the coupling construction—hence conditioning on $\mathcal{C}$ and π_0—, the distribution of $\mathcal{A}$ is hypergeometric with $\mathcal{C}$ draws out of N elements, among which m have $W_{\pi(i)} = 1$. This gives

$$\mathbb{E}[\mathcal{A}|\mathcal{C}, \pi_0] = \mathcal{C}\left(\frac{m}{N}\right), \quad \text{and} \quad \mathbb{V}[\mathcal{A}|\mathcal{C}, \pi_0] = \mathcal{C}\left(\frac{m}{N}\right)\left(\frac{n}{N}\right)\left(\frac{N-\mathcal{C}}{N-1}\right) \; .$$

With this in mind, it can be shown that

$$\mathbb{E}\Big\{\mathbb{V}\Big[\mathbb{E}\left(\mathcal{V}_{m,n}(y)|\,\mathcal{C},\pi,\pi_0\right)|\,\mathcal{C},\pi_0\Big]\Big\} = \frac{1}{N-1}\left[\mathbb{E}(\mathcal{C}) - \mathbb{E}(\mathcal{C}^2)\left(\frac{1}{N}\right)\right]\mathcal{O}(1) = o(1)$$
$$\mathbb{V}\Big\{\mathbb{E}\Big[\mathbb{E}\left(\mathcal{V}_{m,n}(y)|\,\mathcal{C},\pi,\pi_0\right)|\,\mathcal{C},\pi_0\Big]\Big\} = 0\ .$$

Then (A.4) converges to 0 in quadratic mean. Since both processes defining $\mathcal{V}_{m,n}(y)$ are asymptotically equicontinuous, the convergence in probability holds uniformly. This finishes the proof of the Theorem.

A.4 Proof of Theorem 3

The process $V^{\omega}_{m,n}(y;\boldsymbol{X})$ can equivalently be written as

$$\begin{aligned}
V^{\omega}_{m,n}(y;\boldsymbol{X}) &= \sqrt{\frac{mn}{N}}\left\{\hat{F}^{\omega}_1(y) - \hat{F}^{\omega}_0(y) - \left(\hat{F}_1(y) - \hat{F}_0(y)\right)\right\} \\
&= \sqrt{\frac{mn}{N}}\left\{\frac{1}{m}\sum_{i=1}^{N}(\omega_i - 1)\left(\mathbb{1}_{\{Y_i(1)\le y\}} - F_1(y)\right)A_i\right. \\
&\left. -\frac{1}{n}\sum_{i=1}^{N}(\omega_i - 1)\left(\mathbb{1}_{\{Y_i(0)\le y\}} - F_0(y)\right)(1 - A_i)\right\}
\end{aligned}$$

Develop the above expression in the same way as in the proof of Theorem 1 to conclude $V^{\omega}_{m,n}(y) = M_{n,1}\Big(M^{\omega}_{n,2}(y) + M^{\omega}_{n,3}(y)\Big)$, where

$$M_{N,1} = \left(\frac{\sqrt{mn}}{N}\right)\left(\frac{\mathcal{D}_N}{N} + \lambda\right)^{-1}\left(1 - \frac{\mathcal{D}_N}{N} - \lambda\right)^{-1}$$
$$M^{\omega}_{N,2}(y) = \frac{1}{\sqrt{N}}\left\{\sum_{i=1}^{N}(\omega_i - 1)\left((1-\lambda)\left(\mathbb{1}_{\{Y_i(1)\le y\}} - F_1(y)\right)A_i - \lambda\left(\mathbb{1}_{\{Y_i(0)\le y\}} - F_0(y)\right)(1 - A_i)\right)\right\}$$
$$M^{\omega}_{N,3}(y) = -\frac{\mathcal{D}_N}{\sqrt{N}}\left\{\frac{1}{N}\sum_{i=1}^{N}(\omega_i - 1)\left(\left(\mathbb{1}_{\{Y_i(1)\le y\}} - F_1(y)\right)A_i + \left(\mathbb{1}_{\{Y_i(0)\le y\}} - F_0(y)\right)(1 - A_i)\right)\right\}$$

Note that $M_{N,1}$ is the same as in the proof of Theorem 1, therefore $M_{N,1} \xrightarrow{\text{P}} (\lambda(1-\lambda))^{-1/2}$ as $N \to \infty$. Lemma B.9 allows us to conclude that

$$\sup_y \left| M^{\omega}_{N,3}(y) \right| \xrightarrow{\text{P}} 0, \quad \text{as} \quad N \to \infty \ .$$

Lastly, conditional weak convergence of $M^{\omega}_{N,2}(\cdot)$ to $\mathbb{G}_1(\cdot) + \mathbb{G}_2(\cdot) + \mathbb{G}_3(\cdot)$ is established in Lemma B.6, where $(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3)(\cdot)$ are three independent Gaussian processes with covariance functions $\mathbb{C}_1(s,t)$, $\mathbb{C}_2(s,t)$, and $\mathbb{C}_3(s,t)$, defined in Lemma B.2. Therefore $V^{\omega}_{m,n}(\cdot)$ converges weakly in $\ell^{\infty}(\mathscr{F})$ to a tight Gaussian process $\mathbb{H}(\cdot)$ given data; this process is defined in Theorem 1, concluding the proof of the first part of the theorem. Note that the maps $v \to \|v\|$ from $\ell^{\infty}(\mathscr{F})$ into $\mathbb{R}$ are continuous with respect to the supremum norm. Then, a direct application of the continuous mapping theorem (Van der Vaart, 2000, Theorem 18.11) yields the final result. This finishes the proof.

A.5 Proof of Theorem 4

The proof follows closely the arguments in the proof of Chung and Romano (2016, Theorem 2.6). Fix $\delta > 0$ and denote

$$\mathscr{P} \equiv \left\{ \pi \in \mathbf{G}_N \ : \ \sup_y \left| \tilde{F}_{1,\pi}(y) - \bar{P}(y) \right| \leq \delta, \ \sup_y \left| \tilde{F}_{0,\pi}(y) - \bar{P}(y) \right| \leq \delta \right\} \ , \tag{A.6}$$

where $\bar{P}$ is the mixture distribution given by (11), and

$$\tilde{F}_{a,\pi}(y) = \frac{1}{m} \sum_{i=1}^{m} (\omega_i - 1) \mathbb{1}_{\{Y_{a,\pi(i)}\}}, \quad a \in \{0,1\} \ .$$

Then, rewrite the permutation distribution (15) as follows

$$\hat{R}^T_{m,n}(t) = \frac{1}{N!} \sum_{\pi \in \mathscr{P}} \mathbb{1}_{\left\{ J_{m,n}\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}), \tilde{F}_{1,\pi}, \tilde{F}_{0,\pi} \right) \leq t \right\}} + \frac{1}{N!} \sum_{\pi \in \mathscr{P}^c} \mathbb{1}_{\left\{ J_{m,n}\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}), \tilde{F}_{1,\pi}, \tilde{F}_{0,\pi} \right) \leq t \right\}} \ .$$

We derive the limiting behavior of $\hat{R}^T_{m,n}$ in three steps.

Step 1 We begin by showing we can rewrite

$$\hat{R}^T_{m,n}(t) = \frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_{m,n}\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}),\tilde{F}_{1,\pi},\tilde{F}_{0,\pi}\right)\leq t\right\}} + o_p(1)\ . \tag{A.7}$$

To this end, it suffices to show that $(N!)^{-1}|\mathscr{P}| \xrightarrow{\mathrm{P}} 1$, where $|v|$ denotes the cardinality of v. In view of (A.6), the previous sufficient condition follows if we can verify

$$\frac{1}{N!}\sum_{\pi} \mathbb{1}_{\left\{\sup_y\left|\tilde{F}_{1,\pi}(y)-\bar{P}(y)\right|\leq\delta\right\}} \xrightarrow{\mathrm{P}} 1\ , \tag{A.8}$$

and similarly if we replace $\tilde{F}_{1,\pi}$ with $\tilde{F}_{0,\pi}$. By Markov's inequality, a sufficient condition for (A.8) is given by

$$\frac{1}{N!}\sum_{\pi} \mathrm{P}\left\{\sup_y\left|\tilde{F}_{1,\pi}(y) - \bar{P}(y)\right| \leq \delta\right\} \to 1\ . \tag{A.9}$$

By the contiguity results in Chung and Romano (2013, Section 5), we can deduce (A.9) from the basic assumption of how it behaves under an i.i.d. sequence $\xi_1,\ldots,\xi_m$ distributed according to $\bar{P}$ given in (11), combined with the fact that

$$\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\omega_i-1)\left(\mathbb{1}_{\{Y_{1,i}\leq y\}} - \bar{P}\right) \overset{\mathrm{d}}{=} \frac{1}{\sqrt{m}}\sum_{s\in\mathscr{S}}\sum_{i=\mathscr{N}_N(s)+1}^{\mathscr{N}_N(s)+m(s)}\left(\omega_i-1)(\mathbb{1}_{\{Y_i^s(1)\leq y\}} - \bar{P}(y)\right)\right)\ ,$$

where the equality in distribution follows by the same reasoning as in the proof of Theorem 2 and independence between the weights $\omega_1,\ldots,\omega_m$ and data.

We begin by establishing the consistency of the exchangeable bootstrap based on $\xi_1,\ldots,\xi_m$ i.i.d. from $\bar{P}$. Observe first that $\mathscr{F}$ being $\bar{P}$-Donsker implies

$$\left\{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left(\mathbb{1}_{\{\xi_i\leq y\}} - \bar{P}\right)\ : y\in\mathbb{R}\right\}$$

converges weakly to a $\bar{P}$-Brownian bridge process. Since the weights $\omega_1,\ldots,\omega_m$ satisfy assumption A.3, we also have that

$$\left\{\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(\omega_i-1)\left(\mathbb{1}_{\{\xi_i\leq y\}} - \bar{P}\right)\ : y\in\mathbb{R}\right\}$$

converges weakly to the same $\bar{P}$-Brownian bridge process, and

$$\mathrm{P}\left\{\sup_y \left|\frac{1}{m}\sum_{i=1}^{m}(\omega_i - 1)\left(\mathbb{1}_{\{\xi_i \le y\}} - \bar{P}(y)\right)\right| \le \delta\right\} \to 1$$

by the multiplier central limit theorem (Van der Vaart and Wellner, 1996, Theorem 2.9.6). Then,

$$\mathrm{P}\left\{\sup_y \left|\tilde{F}_{1,\pi}(y) - \bar{P}(y)\right| \le \delta\right\} \to 1$$

by Chung and Romano (2013, Lemma 5.3), thus implying (A.9). An analogous argument follows if we replace $\tilde{F}_{1,\pi}$ with $\tilde{F}_{0,\pi}$.

Step 2 We know from previous step that $\tilde{F}_{1,\pi}(y) \xrightarrow{\mathrm{P}} \bar{P}(y)$ and $\tilde{F}_{0,\pi}(y) \xrightarrow{\mathrm{P}} \bar{P}(y)$ in uniform norm. Recall that $J_1(\cdot)$ is the CDF of the supremum of a $\bar{P}$-Brownian bridge. Then, with probability tending to one, we can bound the first term on the right hand side of (A.7) by

$$\begin{aligned}\frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_1\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)})\right)\le t-\varepsilon\right\}} &\le \frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_{m,n}\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}),\tilde{F}_{1,\pi},\tilde{F}_{0,\pi}\right)\le t\right\}} \\ &\le \frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_1\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)})\right)\le t+\varepsilon\right\}} \qquad \text{(A.10)}\end{aligned}$$

for arbitrary $\varepsilon > 0$.

Step 3 We know from Theorem 2 that

$$\sup_t \left|\hat{R}^K_{m,n}(t) - J_1(t)\right| \xrightarrow{\mathrm{P}} 0 \ ,$$

with $J_1(\cdot)$ continuous and strictly increasing at $J_1^{-1}(\cdot)$ by Beran and Millar (1986, Proposition 2). Then, by the continuous mapping theorem for randomization distributions, Chung and Romano (2016, Lemma A.6), we have that

$$\frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_1\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)})\right)\le t-\varepsilon\right\}} \xrightarrow{\mathrm{P}} t - \varepsilon \ ,$$

and

$$\frac{1}{N!}\sum_{\pi\in\mathscr{P}} \mathbb{1}_{\left\{J_1\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)})\right)\le t+\varepsilon\right\}} \xrightarrow{\mathrm{P}} t + \varepsilon \ .$$

Then, for any $\varepsilon > 0$, condition (A.10) reduces to

$$t - \varepsilon \leq \frac{1}{N!} \sum_{\pi \in \mathscr{P}} \mathbb{1}_{\left\{J_{m,n}\left(K_{m,n}(x_{\pi(1)},\ldots,x_{\pi(N)}),\tilde{F}_{1,\pi},\tilde{F}_{0,\pi}\right) \leq t\right\}} \leq t + \varepsilon \ .$$

This finishes the proof of the Theorem.

B Auxiliary Lemmas

Lemma B.1. *Suppose assumptions A.1 and A.2 hold. Then, $M_{N,2}(\cdot)$ converges weakly in $\ell^\infty(\mathscr{F})$ under the null hypothesis to a tight Gaussian process with mean 0 and covariance structure given by*

$$\mathbb{C}(y_1, y_2) = \mathbb{C}_1(y_1, y_2) + \mathbb{C}_2(y_1, y_2) + \mathbb{C}_3(y_1, y_2) \ ,$$

where $\mathbb{C}_1$, $\mathbb{C}_2$, and $\mathbb{C}_3$ are given in (7)–(9).

Proof. Fix y and note that the properties of projection mappings (Brockwell and Davis, 1991, Proposition 2.3.2 and Chapter 2.7) allow us to decompose $\mathbb{1}_{\{Y_i(1)\leq y\}}$ into

$$\mathbb{1}_{\{Y_i(1)\leq y\}} = \mathbb{E}\left(\mathbb{1}_{\{Y_i(1)\leq y\}}|S_i\right) + \varepsilon_{i,1}(y), \quad \text{with} \quad \mathbb{E}(\varepsilon_{i,1}(y)|S_i) = 0 \ , \tag{B.1}$$

$1 \leq i \leq N$. Moreover, observe that

$$\mathbb{E}\left(\mathbb{1}_{\{Y_i(1)\leq y\}}|S_i\right) = \mathbb{E}\left(\mathbb{E}(\mathbb{1}_{\{Y_i(1)\leq y\}}|Z_i)|S_i\right) = \mathbb{E}\Big(F_1(y|Z_i)|S_i\Big) \ ,$$

where the first inequality follows by the tower property of conditional expectations and the fact S_i is a function of Z_i. Denote

$$m_1(y, Z_i) = F_1(y|Z_i) - F_1(y) \tag{B.2}$$

$$m_0(y, Z_i) = F_0(y|Z_i) - F_0(y) \ . \tag{B.3}$$

Plug $m_1(y, Z_i)$ into equation (B.1) to obtain $\mathbb{1}_{\{Y_i(1)\leq y\}} = \mathbb{E}\left(m_1(y, Z_i)|S_i\right) + F_1(y) + \varepsilon_{i,1}(y)$.

Repeat the same argument with $Y_i(1)$ replaced by $Y_i(0)$. Then $M_{N,2}(y)$ can be written as

$$\begin{aligned}
M_{N,2}(y) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda) \left(\mathbb{1}_{\{Y_i(1) \leq y\}} - F_1(y) \right) A_i - \lambda \left(\mathbb{1}_{\{Y_i(0) \leq y\}} - F_0(y) \right) (1 - A_i) \Big\} \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda) \Big(\mathbb{E}\left(m_1(y, Z_i)|S_i\right) + \varepsilon_{i,1}(y) \Big) A_i - \lambda \Big(\mathbb{E}\left(m_0(y, Z_i)|S_i\right) + \varepsilon_{i,0}(y) \Big) (1 - A_i) \Big\} \\
&= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda) \varepsilon_{i,1}(y) A_i - \lambda\, \varepsilon_{i,0}(y) (1 - A_i) \Big\} \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda)\, \mathbb{E}\left(m_1(y, Z_i)|S_i\right) A_i - \lambda\, \mathbb{E}\left(m_0(y, Z_i)|S_i\right) (1 - A_i) \Big\} \,.
\end{aligned}$$

Expand the second summand and rearrange

$$\begin{aligned}
M_{N,2}(y) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda) \varepsilon_{i,1}(y) A_i - \lambda\, \varepsilon_{i,0}(y) (1 - A_i) \Big\} \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left\{ A_i \left((1-\lambda)\, \mathbb{E}\left(m_1(y, Z_i)|S_i\right) + \lambda\, \mathbb{E}\left(m_0(y, Z_i)|S_i\right) \right) - \lambda\, \mathbb{E}\left(m_0(Z_i)|S_i\right) \right\} \,.
\end{aligned}$$

We add the following zero

$$\lambda \left(\sum_{i=1}^{N} (1-\lambda) \Big(\mathbb{E}\left(m_1(y, Z_i)|S_i\right) - \mathbb{E}\left(m_1(y, Z_i)|S_i\right) \Big) + \lambda \Big(\mathbb{E}\left(m_0(y, Z_i)|S_i\right) - \mathbb{E}\left(m_0(y, Z_i)|S_i\right) \Big) \right)$$

to the second summand of $M_{N,2}$ to obtain

$$\begin{aligned}
M_{N,2}(y) &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda) \varepsilon_{i,1}(y) A_i - \lambda\, \varepsilon_{i,0}(y) (1 - A_i) \Big\} \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (A_i - \lambda) \Big((1-\lambda)\, \mathbb{E}\left(m_1(y, Z_i)|S_i\right) + \lambda\, \mathbb{E}\left(m_0(y, Z_i)|S_i\right) \Big) \\
&+ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \lambda (1-\lambda) \Big(\mathbb{E}\left(m_1(y, Z_i)|S_i\right) - \mathbb{E}\left(m_0(y, Z_i)|S_i\right) \Big) \,.
\end{aligned}$$

If we see $\mathbb{E}\left(m_a(y, Z_i)|S_i\right)$, for $a \in \{0, 1\}$, as a function $\mathbb{E}\left(m_a(y, Z_i)|\cdot\right)$ defined on $\mathscr{S}$ and extended

to $\mathbb{R}$, we can compose it with S_i and get

$$\mathbb{E}\left(m_a(y,Z_i)|S_i\right) = \sum_{s\in\mathscr{S}} \mathbb{E}\left(m_a(y,Z_i)|S_i=s\right)\mathbb{1}_{\{S_i=s\}} \ .$$

Therefore,

$$\begin{aligned}
M_{N,2}(y) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big\{(1-\lambda)\varepsilon_{i,1}(y)A_i - \lambda\varepsilon_{i,0}(y)(1-A_i)\Big\} \\
&+ \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(A_i-\lambda)\Bigg(\sum_{s\in\mathscr{S}}(1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_i=s\right)\mathbb{1}_{\{S_i=s\}} \\
&+ \sum_{s\in\mathscr{S}}\lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_i=s\right)\mathbb{1}_{\{S_i=s\}}\Bigg) \\
&+ \frac{\lambda(1-\lambda)}{\sqrt{N}}\sum_{i=1}^{N}\Big(\mathbb{E}\left(m_1(y,Z_i)|S_i\right) - \mathbb{E}\left(m_0(y,Z_i)|S_i\right)\Big) \\
&= G_{N,1}(y) + G_{N,2}(y) + G_{N,3}(y) \ ,
\end{aligned}$$

where

$$G_{N,1}(y) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big\{(1-\lambda)\varepsilon_{i,1}(y)A_i - \lambda\varepsilon_{i,0}(y)(1-A_i)\Big\} \tag{B.4}$$

$$\begin{aligned}
G_{N,2}(y) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}(A_i-\lambda)\Bigg(&\sum_{s\in\mathscr{S}}(1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_i=s\right)\mathbb{1}_{\{S_i=s\}} \\
&+ \sum_{s\in\mathscr{S}}\lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_i=s\right)\mathbb{1}_{\{S_i=s\}}\Bigg)
\end{aligned} \tag{B.5}$$

$$G_{N,3}(y) = \frac{\lambda(1-\lambda)}{\sqrt{N}}\sum_{i=1}^{N}\Big(\mathbb{E}\left(m_1(y,Z_i)|S_i\right) - \mathbb{E}\left(m_0(y,Z_i)|S_i\right)\Big) \ . \tag{B.6}$$

The result follows immediately from Lemma B.2 and the continuous mapping theorem, finishing the proof of Lemma.

□

Lemma B.2. *Suppose assumptions A.1 and A.2 hold. Let $G_{N,1}(\cdot)$, $G_{N,2}(\cdot)$, and $G_{N,3}(\cdot)$ defined as in (B.4)-(B.6), respectively. Then, $\left(G_{N,1}, G_{N,2}, G_{N,3}\right)(\cdot)$ converges weakly in $\ell^\infty(\mathscr{F})$ under the null hypothesis to a tight Gaussian process $\left(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3\right)(\cdot)$ where its marginals $\mathbb{G}_1(\cdot)$, $\mathbb{G}_2(\cdot)$,*

and $\mathbb{G}_3(\cdot)$ *are mutually independent, zero-mean Gaussian processes with covariance structure given by Eqs.* (7)–(9)*, respectively.*

Proof. The proof of the Lemma essentially follows the same construction from Bugni, Canay, and Shaikh (2018, Lemma B.2), extending their results to the uniform case (See also Zhang and Zheng, 2020, Lemma E.2). We separate the proof into three steps.

Step 1. We begin the proof of the Lemma by showing the following asymptotic expansion

$$(G_{N,1}, G_{N,2}, G_{N,3})(\cdot) \overset{\mathrm{d}}{=} \left(G^*_{N,1}, G_{N,2}, G_{N,3}\right)(\cdot) + o_p(1) \tag{B.7}$$

holds uniformly over $y \in \mathbb{R}$, and the process $G^*_{N,1}$—to be defined shortly—*i)* is independent of both $G_{N,2}$ and $G_{N,3}$, and *ii)* weakly converges to $\mathbb{G}_1(\cdot)$ with covariance structure given in (7). We break down the proof of the asymptotic representation (B.7) into two steps.

Step 1.a. We first construct an auxiliary stochastic process $\tilde{G}_{N,1}$ that is stochastically equivalent to $G_{N,1}$ in the wide sense *i.e.* they have the same finite-dimensional distributions. Let

$$\begin{aligned}
\varepsilon^s_{i,1}(y) &= \mathbb{1}_{\{Y^s_i(1)\le y\}} - \mathbb{E}\left(\mathbb{1}_{\{Y_i(1)\le y\}} | S_i = s\right) \\
\varepsilon^s_{i,0}(y) &= \mathbb{1}_{\{Y^s_i(0)\le y\}} - \mathbb{E}\left(\mathbb{1}_{\{Y_i(0)\le y\}} | S_i = s\right)
\end{aligned}$$

where, independently for each $s \in \mathscr{S}$ and independently of $(\mathbf{A}_N, \mathbf{S}_N)$, $\{Y^s_i(1), Y^s_i(0) : 1 \le i \le N\}$ are i.i.d. with marginal distribution equal to the distribution of $(Y_i(1), Y_i(0)) | S_i = s$.

For each s let $\mathscr{N}_N(s) \equiv \sum_{i=1}^N \mathbb{1}_{\{S_i<s\}}$. The auxiliary process $\tilde{G}_{N,1}$ is given by

$$\tilde{G}_{N,1}(y) \equiv \sum_{s\in\mathscr{S}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=\mathscr{N}_N(s)+1}^{\mathscr{N}_N(s)+m(s)} (1-\lambda)\varepsilon^s_{i,1}(y) - \frac{1}{\sqrt{N}} \sum_{i=\mathscr{N}_N(s)+m(s)+1}^{\mathscr{N}_N(s)+m(s)+n(s)} \lambda\varepsilon^s_{i,0}(y) \right\} . \tag{B.8}$$

Intuitively, the auxiliary process $\tilde{G}_{N,1}$ orders units by strata, and then by $A_i = 1$ first and $A_i = 0$ second within strata. This construction—combined with the i.i.d. assumption on data and assumption A.2 (a)—ensures the distribution of $G_{N,1}$ is the same as $\tilde{G}_{N,1}$, since

$$\left\{G_{N,1}(y) | \{A_i, S_i\}_{i=1}^N\right\} \overset{\mathrm{d}}{=} \left\{\tilde{G}_{N,1}(y) | \{A_i, S_i\}_{i=1}^N\right\} .$$

Moreover, since $G_{N,2}(\cdot)$ and $G_{N,3}(\cdot)$ are both functions of $\{A_i, S_i\}_{i=1}^N$ then

$$\Big(G_{N,1}(y), G_{N,2}(y), G_{N,3(y)}\Big) \overset{\mathrm{d}}{=} \Big(\tilde{G}_{N,1}(y), G_{N,2}(y), G_{N,3}(y)\Big) \ .$$

Step 1.b. We further define a process $G^*_{N,1}$ that, *1)* converges weakly to a tight Gaussian process with mean 0 and covariance structure as in (7) and, *2)* satisfies

$$\sup_y \Big|\tilde{G}_{N,1}(y) - G^*_{N,1}(y)\Big| \xrightarrow{\mathrm{P}} 0 \tag{B.9}$$

$$G^*_{N,1}(\cdot) \perp\!\!\!\perp (G_{N,2}, G_{N,3})(\cdot) \ , \tag{B.10}$$

where $G^*_{N,1}(\cdot)$ is given by

$$G^*_{N,1}(y) \equiv \sum_{s\in\mathscr{S}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=\lfloor N\mathcal{N}(s)\rfloor+1}^{\lfloor N(\mathcal{N}(s)+\lambda p(s))\rfloor} (1-\lambda)\varepsilon^s_{i,1}(y) - \frac{1}{\sqrt{N}} \sum_{i=\lfloor N(\mathcal{N}(s)+\lambda p(s))\rfloor+1}^{\lfloor N(\mathcal{N}(s)+p(s))\rfloor} \lambda\varepsilon^s_{i,0}(y) \right\} \ , \tag{B.11}$$

for $\mathscr{N}(s) \equiv P\{S_i < s\}$ for all $s \in \mathscr{S}$. Weak convergence of $G^*_{N,1}$ follows from Lemma B.3, whereas condition (B.9) holds by Lemma B.4. Lastly, the independence condition in (B.10) holds because $G^*_{N,1}(\cdot)$ depends on $\{Y^s_i(1), Y^s_i(0)\}_{i=1}^N$ only, which is independent of $\{A_i, S_i\}_{i=1}^N$ by construction.

Combining Steps 1.a–1.b imply the asymptotic representation (B.7) holds.

Step 2. We now show the weak limits of $G_{N,2}$ and $G_{N,3}$. More specifically, we will show that $G_{N,2}$ and $G_{N,3}$ weakly converge to zero-mean Gaussian processes $\mathbb{G}_2$ and $\mathbb{G}_3$ with covariance structure as in (8)–(9). Consider $G_{N,2}$ first. Observe we can rewrite it as

$$\begin{aligned} G_{N,2}(y) &= \sum_{s\in\mathscr{S}} \sum_{i=1}^N \frac{(A_i-\lambda)}{\sqrt{N}} \mathbb{1}_{\{S_i=s\}} \Big((1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_i=s\right) + \lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_i=s\right) \Big) \\ &= \sum_{s\in\mathscr{S}} \frac{D_N(s)}{\sqrt{N}} \Big((1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_i=s\right) + \lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_i=s\right) \Big) \ , \end{aligned}$$

where $m_j(y, Z_i)$, $j \in \{0, 1\}$ is given by equations (B.2)–(B.3). Fix y and observe that Assumption A.2 implies that $G_{N,2}(y)|\mathbf{S}_N$ converges in distribution to a multivariate normal distribution

with mean zero and covariance

$$\sum_{s\in\mathscr{S}} p(s)\,\tau(s)\Big((1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_1=s\right)+\lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_1=s\right)\Big)^2 . \tag{B.12}$$

Let $\mathscr{E}=\{F_0(y|S):y\in\mathbb{R}\}$ with constant envelope function C and bounded L_1-bracketing numbers of size $2\varepsilon\|C\|$. Then $G_{N,2}(\cdot)$, seen as a random function on $\ell^\infty(\mathscr{E})$, converges weakly—conditionally on $\{S_i\}_{i=1}^N$—to a process $\mathbb{G}_2$ under the null hypothesis. Here $\mathbb{G}_2$ is a tight Gaussian process with mean zero and covariance structure given by

$$\begin{aligned}\mathbb{C}_2(y_1,y_2)=\sum_{s\in\mathscr{S}} p(s)\,\tau(s)\Big(&(1-\lambda)^2\,\mathbb{E}\left(m_1(y_1,Z_i)|S_1=s\right)\mathbb{E}\left(m_1(y_2,Z_i)|S_1=s\right)\\&+\lambda(1-\lambda)\,\mathbb{E}\left(m_1(y_1,Z_i)|S_1=s\right)\mathbb{E}\left(m_0(y_2,Z_i)|S_1=s\right)\\&+\lambda(1-\lambda)\,\mathbb{E}\left(m_1(y_2,Z_i)|S_1=s\right)\mathbb{E}\left(m_0(y_1,Z_i)|S_1=s\right)\\&+\lambda^2\,\mathbb{E}\left(m_0(y_1,Z_i)|S_1=s\right)\mathbb{E}\left(m_0(y_2,Z_i)|S_1=s\right)\Big) .\end{aligned}$$

Similarly, $G_{N,3}(\cdot)$ converges weakly in $\ell^\infty(\mathscr{E})$ to a process $\mathbb{G}_3$ under the null hypothesis. Here $\mathbb{G}_3$ is a tight Gaussian process with mean zero and covariance structure given by

$$\begin{aligned}\mathbb{C}_3(y_1,y_2)=\lambda^2(1-\lambda)^2\sum_{s\in\mathscr{S}} p(s)\Big(&\mathbb{E}[m_1(y_1,Z)|S=s]\,\mathbb{E}[m_1(y_2,Z)|S=s]\\&+\mathbb{E}[m_0(y_1,Z)|S=s]\,\mathbb{E}[m_0(y_2,Z)|S=s]-2\,\mathbb{E}[m_1(y_1,Z)|S=s]\,\mathbb{E}[m_0(y_2,Z)|S=s]\Big) .\end{aligned}$$

Step 3. Lastly, we show that $\left(G^*_{N,1},G_{N,2},G_{N,3}\right)(\cdot)$ weakly converges to a process $(\mathbb{G}_1,\mathbb{G}_2,\mathbb{G}_3)(\cdot)$ where its marginals $\mathbb{G}_1(\cdot)$, $\mathbb{G}_2(\cdot)$, and $\mathbb{G}_3(\cdot)$ are mutually independent. In what follows we consider fixed y but the results carry over by the Cramér–Wold device if we instead fix $y_1,\ldots,y_k\in\mathbb{R}$, $k\in\mathbb{N}$.

By the Cramér–Wold device (Van der Vaart, 2000, Section 2.3), and the marginal convergence of *Steps 1–2*, we have that

$$\left(G^*_{N,1}(y),G_{N,2}(y),G_{N,3}(y)\right)\xrightarrow{\text{d}}(\mathbb{G}_1(y),\mathbb{G}_2(y),\mathbb{G}_3(y))$$

jointly in finite dimension, where $\mathbb{G}_1(y)$, $\mathbb{G}_2(y)$, and $\mathbb{G}_3(y)$ are given as before. *Steps 1–2*

imply that $G^*_{N,1}(\cdot)$, $G_{N,2}(\cdot)$, and $G_{N,3}(\cdot)$ are—individually—asymptotically equicontinuous by Prohorov's theorem (Van der Vaart, 2000, Theorem 18.2). Consequently, $\left(G^*_{N,1}, G_{N,2}, G_{N,3}\right)(\cdot)$ is—jointly—asymptotically equicontinuous and then $\left(G^*_{N,1}, G_{N,2}, G_{N,3}\right)(\cdot)$ weakly converges to $(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3)(\cdot)$ by Van der Vaart and Wellner (1996, Theorem 1.5.4).

In a similar fashion as above, note that for any fixed y,

$$\mathbb{P}\left(G^*_{N,1}(y) \leq t_1, G_{N,2}(y) \leq t_2, G_{N,3}(y) \leq t_3\right) = \mathbb{P}\left(G^*_{N,1}(y) \leq t_1\right)\mathbb{P}\Big(G_{N,2}(y) \leq t_2, G_{N,3}(y) \leq t_3\Big)$$

by the asymptotic representation in (B.7). Observe that

$$\begin{aligned}
\mathbb{P}\left(G_{N,2}(y) \leq t_2, G_{N,3}(y) \leq t_3\right) &= \mathbb{E}\Big\{\mathbb{E}\left(\mathbb{1}_{\{G_{N,2}(y)\leq t_2\}}\mathbb{1}_{\{G_{N,3}(y)\leq t_3\}}|\mathbf{S}_N\right)\Big\} \\
&= \mathbb{E}\Big\{\mathbb{P}\left(G_{N,2}(y) \leq t_2|\mathbf{S}_N\right)\mathbb{1}_{\{G_{N,3}(y)\leq t_3\}}\Big\} \\
&= \mathbb{E}\Big\{\left[\mathbb{P}\left(G_{N,2}(y) \leq t_2|\mathbf{S}_N\right) - \mathbb{P}\left(\mathbb{G}_2(y) \leq t_2\right)\right]\mathbb{1}_{\{G_{N,3}(y)\leq t_3\}}\Big\} \\
&+ \mathbb{E}\Big\{\mathbb{P}\left(\mathbb{G}_2(y) \leq t_2\right)\mathbb{1}_{\{G_{N,3}(y)\leq t_3\}}\Big\} \qquad \text{(B.13)}
\end{aligned}$$

Consider two cases. First, if $\mathbb{P}\left(\mathbb{G}_1(y) \leq \cdot\right)$, $\mathbb{P}\left(\mathbb{G}_2(y) \leq \cdot\right)$, and $\mathbb{P}\left(\mathbb{G}_3(y) \leq \cdot\right)$ are continuous at t_1,t_2, and t_3 respectively for fixed y, then the weak convergence results of Steps 1–2, and dominated convergence theorem (Williams, 1991, Theorem 5.9) applied to (B.13), allow us to conclude

$$\mathbb{P}\left(G^*_{N,1}(y) \leq t_1\right)\mathbb{P}\left(G_{N,2}(y) \leq t_2, G_{N,3}(y) \leq t_3\right) \to \mathbb{P}\left(\mathbb{G}_1(y) \leq t_1\right)\mathbb{P}\left(\mathbb{G}_2(y) \leq t_2\right)\mathbb{P}\left(\mathbb{G}_3(y) \leq t_3\right) . \qquad \text{(B.14)}$$

The same conclusion follows if we now consider the case when $\mathbb{P}\left(\mathbb{G}_1(y) \leq \cdot\right)$, $\mathbb{P}\left(\mathbb{G}_2(y) \leq \cdot\right)$, and $\mathbb{P}\left(\mathbb{G}_3(y) \leq \cdot\right)$ are discontinuous for some t_1,t_2, and t_3—repeat the same argument as in the proof of Lemma B.2 in Bugni, Canay, and Shaikh (2018) combined with the fact that the processes $\mathbb{G}_1$, $\mathbb{G}_2$, and $\mathbb{G}_3$ are Gaussian.

Complete the argument as for the weak convergence above, *i.e.*, invoke asymptotic equicontinuity of $\left(G^*_{N,1}, G_{N,2}, G_{N,3}\right)(\cdot)$ and (B.14) to conclude that the marginals $\mathbb{G}_1(\cdot)$, $\mathbb{G}_2(\cdot)$, and $\mathbb{G}_3(\cdot)$ are mutually independent. This finishes the proof of Lemma.

□

Lemma B.3. *Suppose assumptions A.1 and A.2 hold. Then $G^*_{N,1}(\cdot)$ defined in (B.11) weakly*

converges in $\ell^\infty([0,1]\times\mathscr{F})$ *under the null hypothesis to* $\mathbb{G}_1$*, a tight Gaussian process with mean zero and covariance structure given by*

$$\mathbb{C}(\mathbb{G}_1(y_1),\mathbb{G}_1(y_2)) = \lambda(1-\lambda)\left(F_0(y_1\wedge y_2) - F_0(y_1)F_0(y_2)\right) \ .$$

Proof. The proof is based on the partial-sum representation of the process $G^*_{N,1}(\cdot)$ (Van der Vaart and Wellner, 1996, Chapter 2.12). Let $\cup_{s\in\mathscr{S}}\chi_s$ be a partition of the sample space into lower rectangles in $\mathbb{R}$ corresponding to strata $s\in\mathscr{S}$, and let $\mathscr{F}_s$ be the class of functions $f\mathbb{1}_{\{\chi_s\}}$ when f ranges over $\mathscr{F}$. Since the class $\mathscr{F}$ is Donsker—its bracketing numbers are of the polynomial order $(1/\varepsilon)^2$)—then each class $\mathscr{F}_s$ is Donsker (Van der Vaart and Wellner, 1996, Theorem 2.10.6).

Consider the following partial-sum process—also known as *sequential empirical process*—given by

$$\begin{aligned}\mathbb{Z}_{N,1}(t,\varepsilon^s_1(y),s) &= \frac{1}{\sqrt{N}}\sum_{i=1}^{\lfloor Nt\rfloor}(1-\lambda)\varepsilon^s_{i,1}(y)\\ &= (1-\lambda)\sqrt{\frac{\lfloor Nt\rfloor}{N}}\mathbb{G}_{\lfloor Nt\rfloor}\,\varepsilon^s_1(y)\ , \quad \tfrac{\lfloor N\lambda p(s)\rfloor}{N}\le t\le\tfrac{\lfloor N\lambda p(s)\rfloor+1}{N}\ . \end{aligned} \tag{B.15}$$

Since the class $\mathscr{F}_s$ is Donsker, then Van der Vaart and Wellner (1996, Theorem 2.12.1) implies that $\mathbb{Z}_{N,1}$ weakly converges in $\ell^\infty([0,1]\times\mathscr{F}_s)$ to a tight Gaussian process $\mathbb{Z}_1$—the *Kiefer–Müller process*. This process has mean zero and covariance structure

$$\mathbb{C}(\mathbb{Z}_1(t_1,y_1,s),\mathbb{Z}_1(t_2,y_2,s) = (1-\lambda)^2(t_1\wedge t_2)\left(F_1(y_1\wedge y_2) - F_1(y_1)F_1(y_2)\right) \ . \tag{B.16}$$

In particular, (B.16) reduces to $\lambda\,p(s)(1-\lambda)^2\left(F_0(y_1\wedge y_2) - F_0(y_1)F_0(y_2)\right)$ for $t_1=t_2=\lambda\,p(s)$. Repeating an analogous argument for $\varepsilon^s_{i,0}(y)$ we can conclude that $\mathbb{Z}_{N,0}$ weakly converges in $\ell^\infty([0,1]\times\mathscr{F}_s)$ to another Kiefer–Müller process $\mathbb{Z}_0$ with mean zero and covariance structure

$$\mathbb{C}(\mathbb{Z}_0(t_1,y_1,s),\mathbb{Z}_0(t_2,y_2,s) = \lambda^2(1-\lambda)p(s)\left(F_0(y_1\wedge y_2) - F_0(y_1)F_0(y_2)\right) \ . \tag{B.17}$$

Exploiting the fact that the finite union of Donsker classes—across $s\in\mathscr{S}$, and experimental groups—is Donsker too, then we conclude that $G^*_{N,1}(\cdot)$ defined in (B.11) weakly converges to a tight Gaussian process $\mathbb{G}_1$—the two-sample version of the Kiefer–Müller process—with mean

zero and covariance structure under the null hypothesis given by

$$\begin{aligned}\mathbb{C}(\mathbb{G}_1(y_1), \mathbb{G}_1(y_2)) &= \lambda(1-\lambda)\Big(\lambda\left\{F_1(y_1 \wedge y_2) - F_1(y_1)F_1(y_2)\right\} \\ &\quad + (1-\lambda)\left\{F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2)\right\}\Big) \\ &= \lambda(1-\lambda)F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2) \ ,\end{aligned} \tag{B.18}$$

where in the last equality we have used the fact that $F_0 = F_1$ holds under the null hypothesis $F_0 = F_1$. This finishes the proof. □

Lemma B.4. *Suppose assumptions A.1 and A.2 hold. Then*

$$\sup_y \left|\tilde{G}_{N,1}(y) - G^*_{N,1}(y)\right| \xrightarrow{\mathrm{P}} 0 \ .$$

where the processes $\tilde{G}_{N,1}$ and $G^*_{N,1}$ are given by equations (B.8) and (B.11), respectively.

Proof. In view of Markov's inequality, it suffices to show

$$\mathbb{E}\left(\sup_y \left|\tilde{G}_{N,1}(y) - G^*_{N,1}(y)\right|\right) \to 0 \ . \tag{B.19}$$

To this end, fix y and an arbitrary $s \in \mathscr{S}$. Consider the following expression

$$\left\{\sum_{i=1}^{m(s)} \varepsilon^s_{i,1}(y) - \sum_{i=1}^{\lfloor N\lambda p(s)\rfloor} \varepsilon^s_{i,1}(y)\right\} - \left\{\sum_{i=m(s)+1}^{m(s)+n(s)} \varepsilon^s_{i,0}(y) - \sum_{i=\lfloor N\lambda p(s)\rfloor+1}^{\lfloor Np(s)\rfloor} \varepsilon^s_{i,0}(y)\right\} \ , \tag{B.20}$$

and focus on the first summand between braces. By construction $m(s)$ is distributed as a binomial $B(N, \lambda p(s))$. Therefore

$$\begin{aligned}\mathbb{P}\left(|m(s) - \lfloor N\lambda p(s)\rfloor| \geq N\right) &= \mathbb{P}\left(m(s) \leq \lfloor N\lambda p(s)\rfloor - N\right) - \mathbb{P}\left(m(s) \leq N + \lfloor N\lambda p(s)\rfloor\right) \\ &\leq \exp\left\{-2\left(N + 2 + \frac{1}{N}\right)\right\} - \exp\left\{-2\left(N - 2 - \frac{1}{N}\right)\right\}\end{aligned}$$

by Hoeffding's Inequality (Pollard, 1984, Appendix B). We reach a similar conclusion for the

second summand between braces using the same argument. Then, (B.20) can be formulated as

$$\left\{\sum_{i=1}^{m(s)}\varepsilon_{i,1}^{s}(y)-\sum_{i=1}^{\lfloor N\lambda p(s)\rfloor}\varepsilon_{i,1}^{s}(y)\right\}-\left\{\sum_{i=m(s)+1}^{m(s)+n(s)}\varepsilon_{i,0}^{s}(y)-\sum_{i=\lfloor N\lambda p(s)\rfloor+1}^{\lfloor Np(s)\rfloor}\varepsilon_{i,0}^{s}(y)\right\}=\sum_{i=1}^{r_{1,m}(s)}\varepsilon_{i,1}^{s}(y)-\sum_{i=1}^{r_{0,n}(s)}\varepsilon_{i,0}^{s}(y)$$

with $r_{1,m}(s)$ and $r_{0,m}(s)$ two integer-valued random variables such that $r_{1,m}(s)\xrightarrow{\mathrm{P}}0$, $r_{0,n}(s)\xrightarrow{\mathrm{P}}0$, as $N\to\infty$. Building on this result, we obtain

$$\tilde{G}_{N,1}(y)-G_{N,1}^{*}(y)=\sum_{s\in\mathscr{S}}\left\{\frac{1}{\sqrt{N}}\sum_{i=1}^{r_{1,m}(s)}(1-\lambda)\varepsilon_{i,1}^{s}(y)-\frac{1}{\sqrt{N}}\sum_{i=1}^{r_{0,n}(s)}\lambda\varepsilon_{i,0}^{s}(y)\right\}.$$

Combine the weak convergence result of Lemma B.3 with the fact that $r_{1,m}(s)\xrightarrow{\mathrm{P}}0$ and $r_{0,n}(s)\xrightarrow{\mathrm{P}}0$ to conclude that the process $\left(\tilde{G}_{N,1}-G_{N,1}^{*}\right)(\cdot)$ weakly converges to zero by Durrett and Resnick (1977, Theorem 3). Then (B.19) follows by Van der Vaart and Wellner (1996, Lemma 2.3.11), thus finishing the proof. □

Lemma B.5. *Suppose assumptions A.1 and A.2 hold. Then*

$$\sup_{y}\left|\frac{1}{N}\sum_{i=1}^{N}\left(\left(\mathbb{1}_{\{Y_i(1)\le y\}}-F_1(y)\right)A_i+\left(\mathbb{1}_{\{Y_i(0)\le y\}}-F_0(y)\right)(1-A_i)\right)\right|\xrightarrow{\mathrm{P}}0\,. \tag{B.21}$$

Proof. For notational convenience, write

$$\Psi_N(y)=\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}_{\{Y_i(1)\le y\}}-F_1(y)\right)A_i\,.$$

Arguing as in the proof of Bugni, Canay, and Shaikh (2018, Lemma B.3), independently for each $s\in\mathscr{S}$ and $(\mathbf{A}_N,\mathbf{S}_N)$, let $\{(Y_i^s(1),Y_i^s(0)):1\le i\le N\}$ be i.i.d. with marginal distribution equal to the distribution of $(Y_i(1),Y_i(0))|S_i=s$. Note that

$$\Psi_N(y)\overset{\mathrm{d}}{=}\sum_{s\in\mathscr{S}}\frac{m(s)}{N}\left(\frac{1}{m(s)}\sum_{i=1}^{m(s)}\left(\mathbb{1}_{\{Y_i^s(1)\le y\}}-F_1(y)\right)\right).$$

To show that $\Psi_N(y) \xrightarrow{\mathrm{P}} 0$ uniformly over $y \in \mathbb{R}$, it suffices to establish for any $\varepsilon > 0$

$$\mathrm{P}\left\{\sup_y \left| \frac{1}{m(s)} \sum_{i=1}^{m(s)} \left(\mathbb{1}_{\{Y_i^s(1) \le y\}} - F_1(y) \right) \right| > \varepsilon \right\} \to 0 \ .$$

The almost sure representation theorem (Van der Vaart, 2000, Theorem 2.19) allows us to construct a sequence $\tilde{m}(s)/N \stackrel{\mathrm{d}}{=} m(s)/N$ with $\tilde{m}(s)/N \xrightarrow{\text{a.s.}} \lambda p(s) > 0$ as $N \to \infty$. Then, using the independence of $(\mathbf{A}_N, \mathbf{S}_N)$ and $\{\mathbb{1}_{\{Y_i^s(1) \le y\}} : 1 \le i \le N\}$, we see that for any $\varepsilon > 0$,

$$\begin{aligned} P\left\{\sup_y \left| \frac{1}{m(s)} \sum_{i=1}^{m(s)} \mathbb{1}_{\{Y_i^s(1) \le y\}} \right| > \varepsilon \right\} &= P\left\{\sup_y \left| \frac{1}{\tilde{m}(s)} \sum_{i=1}^{\tilde{m}(s)} \mathbb{1}_{\{Y_i^s(1) \le y\}} \right| > \varepsilon \right\} \\ &= \mathbb{E}\left[P\left\{\sup_y \left| \frac{1}{\tilde{m}(s)} \sum_{i=1}^{\tilde{m}(s)} \mathbb{1}_{\{Y_i^s(1) \le y\}} \right| > \varepsilon \,\middle|\, \tilde{m}(s) \right\} \right] \ . \end{aligned} \tag{B.22}$$

Let $\cup_{s \in \mathscr{S}} \chi_s$ be a partition of the sample space into lower rectangles in $\mathbb{R}$ corresponding to strata $s \in \mathscr{S}$, and let $\mathscr{F}_s$ be the class of functions $f\mathbb{1}_{\{\chi_s\}}$ when f ranges over $\mathscr{F}$. Since the class $\mathscr{F}$ is Donsker then each class $\mathscr{F}_s$ is Donsker (Van der Vaart and Wellner, 1996, Theorem 2.10.6). Then for fixed m,

$$\sup_y \left| \frac{1}{m} \sum_{i=1}^{m} \left(\mathbb{1}_{\{Y_i^s(1) \le y\}} - F_1(y) \right) \right| \xrightarrow{\mathrm{P}} 0 \tag{B.23}$$

by Glivenko–Cantelli theorem. The independence of $\tilde{m}(s)$ and $\{\mathbb{1}_{\{Y_i^s(1) \le y\}} : 1 \le i \le N\}$, and (B.23) imply that

$$P\left\{\sup_y \left| \frac{1}{\tilde{m}(s)} \sum_{i=1}^{\tilde{m}(s)} \mathbb{1}_{\{Y_i^s(1) \le y\}} \right| > \varepsilon \,\middle|\, \tilde{m}(s) \right\} = P\left\{\sup_y \left| \frac{1}{\tilde{m}(s)} \sum_{i=1}^{\tilde{m}(s)} \mathbb{1}_{\{Y_i^s(1) \le y\}} \right| > \varepsilon \right\} \to 0 \ , \tag{B.24}$$

by (B.23) and Van der Vaart and Wellner (1996, Theorem 3.5.1). The desired conclusion follows by (B.24) and a direct application of dominated convergence theorem (Williams, 1991, Theorem 5.9) to (B.22). Apply the same reasoning to the second summand in (B.21) to conclude

$$\sup_y \left| \frac{1}{N} \sum_{i=1}^{N} \left(\left(\mathbb{1}_{\{Y_i(0) \le y\}} - F_0(y) \right) (1 - A_i) \right) \right| \xrightarrow{\mathrm{P}} 0 \ .$$

Triangle inequality yields the desired result. □

Lemma B.6. *Suppose assumptions A.1 and A.2 hold. If* $(\omega_1, \ldots, \omega_N)$ *are weights satisfying Assumption 3. then, conditionally on data,* $M^{\omega}_{N,2}(\cdot)$ *converges weakly in* $\ell^{\infty}(\mathscr{F})$ *to a tight Gaussian under the null hypothesis. The limit process has mean* 0 *and covariance structure given by*

$$\mathbb{C}(y_1, y_2) = \mathbb{C}_1(y_1, y_2) + \mathbb{C}_2(y_1, y_2) + \mathbb{C}_3(y_1, y_2) \ ,$$

where $\mathbb{C}_1$, $\mathbb{C}_2$, *and* $\mathbb{C}_3$ *are given in* (7)–(9).

Proof. The lemma is proved by mirroring the arguments used in the proofs of Lemmas B.1–B.2, so we omit some details. In fact, arguing as in Lemma B.1, we can show that

$$M^{\omega}_{N,2}(y) = G^{\omega}_{N,1}(y) + G^{\omega}_{N,2}(y) + G^{\omega}_{N,3}(y) \ ,$$

where

$$G^{\omega}_{N,1}(y) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \Big\{ (1-\lambda)(\omega_i - 1)\varepsilon_{i,1}(y) A_i - \lambda(\omega_i - 1)\varepsilon_{i,0}(y)(1 - A_i) \Big\} \tag{B.25}$$

$$\begin{aligned} G^{\omega}_{N,2}(y) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (A_i - \lambda) \Bigg(& \sum_{s \in \mathscr{S}} (1-\lambda)(\omega_i - 1)\, \mathbb{E}\left(m_1(y, Z_i) | S_i = s\right) \mathbb{1}_{\{S_i = s\}} \\ & + \sum_{s \in \mathscr{S}} \lambda(\omega_i - 1)\, \mathbb{E}\left(m_0(y, Z_i) | S_i = s\right) \mathbb{1}_{\{S_i = s\}} \Bigg) \end{aligned} \tag{B.26}$$

$$G^{\omega}_{N,3}(y) = \frac{\lambda(1-\lambda)}{\sqrt{N}} \sum_{i=1}^{N} (\omega_i - 1) \Big(\mathbb{E}\left(m_1(y, Z_i) | S_i\right) - \mathbb{E}\left(m_0(y, Z_i) | S_i\right) \Big) \ . \tag{B.27}$$

We break the proof of the lemma into three steps as used in the proof of Lemma B.2.

Step 1. We begin by showing the follow asymptotic expansion

$$\left(G^{\omega}_{N,1}, G^{\omega}_{N,2}, G^{\omega}_{N,3}\right)(\cdot) \overset{\mathrm{d}}{=} \left(G^{*,\omega}_{N,1}, G^{\omega}_{N,2}, G^{\omega}_{N,3}\right)(\cdot) + o_p(1) \tag{B.28}$$

holds uniformly over $y \in \mathbb{R}$, and the process $G^{*,\omega}_{N,1}$—to be defined shortly—*i)* is independent of both $G^{\omega}_{N,2}$ and $G^{\omega}_{N,3}$, and *ii)* converges weakly to $\mathbb{G}_1(\cdot)$ conditionally on data and its covariance structure is given in (7). We break down the proof of the asymptotic representation (B.28) into two steps.

Step 1.a. For fixed y, set $\varepsilon^{s}_{i,1}(y)$ and $\varepsilon^{s}_{i,0}(y)$ as in Step 1.a in the proof of Lemma B.2, and

define the auxiliary process $\tilde{G}^{\omega}_{N,1}$ by

$$\tilde{G}_{N,1}(y) \equiv \sum_{s\in\mathscr{S}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=\mathcal{N}_N(s)+1}^{\mathcal{N}_N(s)+m(s)} (1-\lambda)(\omega_i - 1)\varepsilon^s_{i,1}(y) - \frac{1}{\sqrt{N}} \sum_{i=\mathcal{N}_N(s)+m(s)+1}^{\mathcal{N}_N(s)+m(s)+n(s)} \lambda(\omega_i - 1)\varepsilon^s_{i,0}(y) \right\} . \tag{B.29}$$

This construction—combined with the i.i.d. assumption on data and Assumption 2 *i)*—ensures the distribution of $G^{\omega}_{N,1}$ is the same as $\tilde{G}^{\omega}_{N,1}$ since

$$\left\{ G^{\omega}_{N,1}(y) | \{A_i, S_i\}_{i=1}^N \right\} \overset{\mathrm{d}}{=} \left\{ \tilde{G}^{\omega}_{N,1}(y) | \{A_i, S_i\}_{i=1}^N \right\} .$$

Moreover, since $G^{\omega}_{N,2}(\cdot)$ and $G^{\omega}_{N,3}(\cdot)$ are both functions of $\{A_i, S_i\}_{i=1}^N$ then

$$\left(G^{\omega}_{N,1}(y), G^{\omega}_{N,2}(y), G^{\omega}_{N,3(y)} \right) \overset{\mathrm{d}}{=} \left(\tilde{G}^{\omega}_{N,1}(y), G^{\omega}_{N,2}(y), G^{\omega}_{N,3}(y) \right) .$$

Step 1.b. We further define the process $G^{*,\omega}_{N,1}$ as

$$G^{*,\omega}_{N,1}(y) \equiv \sum_{s\in\mathscr{S}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=\lfloor N\mathcal{N}(s)\rfloor+1}^{\lfloor N(\mathcal{N}(s)+\lambda p(s))\rfloor} (1-\lambda)(\omega_i - 1)\varepsilon^s_{i,1}(y) - \frac{1}{\sqrt{N}} \sum_{i=\lfloor N(\mathcal{N}(s)+\lambda p(s))\rfloor+1}^{\lfloor N(\mathcal{N}(s)+p(s))\rfloor} \lambda(\omega_i - 1)\varepsilon^s_{i,0}(y) \right\} , \tag{B.30}$$

where $\mathscr{N}(s) \equiv P\{S_i < s\}$ for all $s \in \mathscr{S}$, and observe that it *1)* converges weakly to a tight Gaussian process with mean 0 and covariance structure as in (7) given data, in probability; and, *2)* satisfies

$$\sup_y \left| \tilde{G}^{\omega}_{N,1}(y) - G^{*,\omega}_{N,1}(y) \right| \xrightarrow{\mathrm{P}} 0 \tag{B.31}$$

$$G^{*\omega}_{N,1}(\cdot) \perp\!\!\!\perp (G^{\omega}_{N,2}, G^{\omega}_{N,3})(\cdot) . \tag{B.32}$$

Conditional weak convergence in probability of $G^{*,\omega}_{N,1}$ follows from Lemma B.7, whereas condition (B.31) holds by Lemma B.8. Finally, the independence condition in (B.32) holds because $G^{*,\omega}_{N,1}(\cdot)$ depends on $\{Y^s_i(1), Y^s_i(0)\}_{i=1}^N$ only, which is independent of $\{A_i, S_i\}_{i=1}^N$ by construction.

Combining Steps 1.a–1.b imply the asymptotic representation (B.28) holds.

Step 2. We now show the conditional weak limits of $G^{\omega}_{N,2}$ and $G^{\omega}_{N,3}$. More specifically, we will

show that $G^{\omega}_{N,2}$ and $G^{\omega}_{N,3}$ weakly converge to zero-mean Gaussian processes $\mathbb{G}_2$ and $\mathbb{G}_3$ with covariance structure as in (8)–(9). Consider $G^{\omega}_{N,2}$ first, and note that

$$G^{\omega}_{N,2}(y) = \sum_{s\in\mathscr{S}} \frac{D^{\omega}_N(s)}{\sqrt{N}} \Big((1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_i=s\right) + \lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_i=s\right) \Big) \ ,$$

where $m_j(y,Z_i)$, $j\in\{0,1\}$ is given by equations (B.2)–(B.3), and $D^{\omega}_N(s)$ is

$$D^{\omega}_N(s) = \sum_{i=1}^{N} (\omega_i - 1)(A_i - \lambda)\mathbb{1}_{\{S_i=s\}}, \quad s\in\mathscr{S} \ .$$

Fix y and observe that assumptions A.2 and A.3 imply that $G^{\omega}_{N,2}(y)|\mathbf{S}_N$ converges in distribution, conditionally on data, to a multivariate normal distribution with mean zero and covariance

$$\sum_{s\in\mathscr{S}} p(s)\,\tau(s) \Big((1-\lambda)\,\mathbb{E}\left(m_1(y,Z_i)|S_1=s\right) + \lambda\,\mathbb{E}\left(m_0(y,Z_i)|S_1=s\right) \Big)^2$$

for almost all sequences $(Y_1,A_1,Z_1),(Y_2,A_2,Z_2),\ldots$, by Van der Vaart and Wellner (1996, Lemma 2.9.5). The preceding result takes care of conditional marginal convergence in probability.

For conditional weak convergence, it suffices to check asymptotic equicontinuity in terms of conditional laws. In view of Van der Vaart and Wellner (1996, Theorem 2.9.6), we need to verify the assumptions of the conditional multiplier central limit theorem. Let $\mathscr{E} = \{F_0(y|S) : y\in\mathbb{R}\}$ with constant envelope function C and bounded L_1-bracketing numbers of size $2\varepsilon\|C\|$, so $\mathscr{E}$ is Donsker. Moreover, under our assumptions, the weights $\{\omega_i - 1 : 1\le i\le N\}$ are i.i.d. random variables with mean 0 and variance 1, satisfying condition A.3. Then $G^{\omega}_{N,2}(\cdot)$, seen as a random function on $\ell^{\infty}(\mathscr{E})$, converges weakly to a process $\mathbb{G}_2$ given data, in probability (see also Van der Vaart and Wellner (1996, Theorem 3.6.13)).

By the same reasoning, $G^{\omega}_{N,3}(\cdot)$ given in (B.27) converges weakly in $\ell^{\infty}(\mathscr{E})$ to a process $\mathbb{G}_3$ given data, in probability. Here $\mathbb{G}_3$ is a tight Gaussian process with mean zero and covariance structure given by (9).

Step 3. Lastly, we show that $\left(G^{*,\omega}_{N,1}, G^{\omega}_{N,2}, G^{\omega}_{N,3}\right)(\cdot)$ weakly converges to a process $(\mathbb{G}_1,\mathbb{G}_2,\mathbb{G}_3)(\cdot)$ where its marginals $\mathbb{G}_1(\cdot)$, $\mathbb{G}_2(\cdot)$, and $\mathbb{G}_3(\cdot)$ are mutually independent. This step is proved by the same arguments as used in the proof of Step 3, Lemma B.2, and the fact that the weights

$(\omega_1, \ldots, \omega_N)$ are independent of data. This finishes the proof of the Lemma.

□

Lemma B.7. *Suppose assumptions A.1 and A.2 hold. If $(\omega_1, \ldots, \omega_N)$ are weights satisfying Assumption 3, then $G^{*,\omega}_{N,1}(\cdot)$ defined in (B.30) converges weakly in $\ell^\infty([0,1] \times \mathscr{F})$ to $\mathbb{G}_1$ given data, in probability. Here $\mathbb{G}_1$ is a tight Gaussian process with mean zero and covariance structure given by*

$$\mathbb{C}(\mathbb{G}_1(y_1), \mathbb{G}_1(y_2)) = \lambda(1-\lambda)\left(F_0(y_1 \wedge y_2) - F_0(y_1)F_0(y_2)\right) \ .$$

Proof. The proof resembles that of Lemma (B.3), so we omit some details. Set $\mathscr{F}_s$ as in the proof of Lemma (B.3), and consider the process

$$\left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{N} (1-\lambda)(\omega_i - 1)(\varepsilon^s_{i,1}(y)) \ : \ y \in \mathbb{R} \right\} \ . \tag{B.33}$$

Under our assumptions, the marginals of (B.33) converge weakly, conditional on data, to the marginals of normally distributed random variable with mean zero and covariance $(1-\lambda)^2 F_0(y)(1-F_0(y))$ by the conditional multiplier central limit theorem (Van der Vaart and Wellner, 1996, Lemma 2.9.5). For conditional weak convergence in $\ell^\infty(\mathscr{F}_s)$ in probability, it suffices to check asymptotic equicontinuity of (B.33). This follows automatically under our assumptions and Van der Vaart and Wellner (1996, Theorem 2.9.6).

With this in mind, we can now establish conditional weak convergence of $G^{*,\omega}_{N,1}(\cdot)$ in probability. As in Lemma (B.3), the proof is based on its partial-sum representation,

$$\mathbb{Z}^{\omega}_{N,1}(t, \varepsilon^s_1(y), s) = (1-\lambda)\sqrt{\frac{\lfloor Nt \rfloor}{N}} \left\{ \frac{1}{\sqrt{\lfloor Nt \rfloor}} \sum_{i=1}^{\lfloor Nt \rfloor} (\omega_i - 1)\varepsilon^s_{i,1}(y) \right\} \ , \quad \tfrac{\lfloor N\lambda p(s) \rfloor}{N} \leq t \leq \tfrac{\lfloor N\lambda p(s) \rfloor + 1}{N} \ .$$

Since the class $\mathscr{F}_s$ is Donsker, then Van der Vaart and Wellner (1996, Theorem 2.12.1) implies that $\mathbb{Z}^{\omega}_{N,1}$ converges weakly in $\ell^\infty([0,1] \times \mathscr{F}_s)$, conditional on data, to the Kiefer–Müller process of Lemma B.3.

Repeating an analogous argument for $\varepsilon^s_{i,0}(y)$ and using the fact that the finite union of Donsker classes—across $s \in \mathscr{S}$, and experimental groups—is Donsker too, we can conclude that $G^{*,\omega}_{N,1}(\cdot)$ converges weakly, conditional on data, to a tight Gaussian process $\mathbb{G}_1$ in probability,

where $\mathbb{G}_1$ has mean zero and covariance structure as in (7). This finishes the proof. □

Lemma B.8. *Suppose assumptions A.1 and A.2 hold. If $(\omega_1, \ldots, \omega_N)$ are weights satisfying Assumption 3, then*

$$\sup_y \left| \tilde{G}^{\omega}_{N,1}(y) - G^{*,\omega}_{N,1}(y) \right| \xrightarrow{\mathrm{P}} 0 \ ,$$

conditionally on data, where the processes $\tilde{G}^{\omega}_{N,1}$ and $G^{*,\omega}_{N,1}$ are given by equations (B.29) and (B.30), respectively.

Proof. In view of Markov's inequality, it suffices to verify

$$\mathbb{E}\left(\sup_y \left| \tilde{G}^{\omega}_{N,1}(y) - G^{*,\omega}_{N,1}(y) \right| \right) \to 0 \ , \tag{B.34}$$

conditionally on data. Arguing as in the proof of Lemma B.4, we can show that

$$\tilde{G}^{\omega}_{N,1}(y) - G^{*,\omega}_{N,1}(y) = \sum_{s \in \mathcal{S}} \left\{ \frac{1}{\sqrt{N}} \sum_{i=1}^{r_{1,m}(s)} (1-\lambda)(\omega_i - 1)\varepsilon^s_{i,1}(y) - \frac{1}{\sqrt{N}} \sum_{i=1}^{r_{0,n}(s)} \lambda(\omega_i - 1)\varepsilon^s_{i,0}(y) \right\} \ ,$$

with $r_{1,m}(s)$ and $r_{0,m}(s)$ two integer-valued random variables such that $r_{1,m}(s) \xrightarrow{\mathrm{P}} 0$, $r_{0,n}(s) \xrightarrow{\mathrm{P}} 0$, as $N \to \infty$. Combine the conditional weak convergence in probability result of Lemma B.7 with the fact that $r_{1,m}(s) \xrightarrow{\mathrm{P}} 0$ and $r_{0,n}(s) \xrightarrow{\mathrm{P}} 0$ to conclude that the process $\left(\tilde{G}^{\omega}_{N,1} - G^{*,\omega}_{N,1} \right)(\cdot)$ weakly converges to zero by Durrett and Resnick (1977, Theorem 3). Then (B.34) follows by Van der Vaart and Wellner (1996, Lemma 2.3.11). This finishes the proof. □

Lemma B.9. *Suppose assumptions A.1 and A.2 hold. If $(\omega_1, \ldots, \omega_N)$ are weights satisfying Assumption 3, then*

$$\sup_y \left| \frac{1}{N} \sum_{i=1}^{N} (\omega_i - 1) \left(\left(\mathbb{1}_{\{Y_i(1) \le y\}} - F_1(y) \right) A_i + \left(\mathbb{1}_{\{Y_i(0) \le y\}} - F_0(y) \right) (1 - A_i) \right) \right| \xrightarrow{\mathrm{P}} 0 \ . \tag{B.35}$$

Proof. The proof is essentially the same as the proof of Lemma (B.5), so we omit some details.

Consider

$$\Psi_N^{\omega}(y) = \frac{1}{N}\sum_{i=1}^{N}(\omega_i - 1)\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)A_i$$
$$\stackrel{\mathrm{d}}{=} \sum_{s\in\mathcal{S}}\frac{m(s)}{N}\left(\frac{1}{m(s)}\sum_{i=1}^{m(s)}(\omega_i - 1)\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)\right) .$$

The rest of the proof is the same as Lemma (B.5), except that the convergence result in (B.23) now becomes

$$\sup_y\left|\frac{1}{m}\sum_{i=1}^{m}(\omega_i - 1)\left(\mathbb{1}_{\{Y_i(1)\leq y\}} - F_1(y)\right)\right| \xrightarrow{\mathrm{P}} 0 ,$$

and follows by Van der Vaart and Wellner (1996, 2.9.2) instead. This finishes the proof.

□

C Examples of CAR Schemes

Example 1. (Simple Randomization) Simple randomization refers to the case when every sequence of treatment assignments is equally likely. More formally, $\mathbf{A}_N$ consists of N i.i.d. random variables with

$$\mathbb{P}\{A_k = 1|\mathbf{S}_k, \mathbf{A}_{k-1}\} = \mathbb{P}\{A_k = 1\} = \lambda$$

for $1 \leq k \leq N$. Note that $\mathbb{E}(\mathcal{D}_N(s)) = 0$ for all $s \in \mathscr{S}$.

Example 2. (Covariate-Adaptive Biased-coin Design) (Efron, 1971)'s biased-coin design is given by:

$$\mathbb{P}\{A_k = 1|\mathbf{S}_k, \mathbf{A}_{k-1}\} = \begin{cases} 1/2 & \mathcal{D}_{k-1}(S_k) = 0 \\ \gamma & \mathcal{D}_{k-1}(S_k) < 0 \\ 1-\gamma & \mathcal{D}_{k-1}(S_k) > 0 \end{cases} ,$$

where $\mathcal{D}_{k-1}(S_k) = \sum_{i=1}^{k-1}(A_i - 1/2)\,\mathbb{1}_{\{S_i = S_k\}}$, $\gamma > 1/2$, $\mathcal{D}_0(S_1) = 0$. Efron's original biased-coin design does not make use of any covariate Z_i, so it cannot be considered a CAR scheme in strict sense. However, one may ensure it is a CAR scheme if we apply it within each stratum S_i (Shao, Yu, and Zhong, 2010; Pocock and Simon, 1975). It improves balance relative to simple

randomization.

Example 3. (Adaptive Biased-coin Design) Due to Wei (1978), this CAR scheme is similar to Efron's biased-coin design and it is given by

$$\mathbb{P}\{A_k = 1|\mathbf{S}_k, \mathbf{A}_{k-1}\} = \varphi\left(\frac{\mathcal{D}_{k-1}(S_k)}{k-1}\right) ,$$

where $\varphi(x) : [0,1] \to [0,1]$ is a pre-specified function satisfying $\varphi(-x) = 1 - \varphi(x)$.

Example 4. (Stratified Block Randomization) For $s \in \mathscr{S}$, denote $N(s) = \sum_{1\le i\le N} \mathbb{1}\{S_i = s\}$, and $m(s)$ and $n(s)$ as before. In this randomization scheme all possible assignments in stratum $s \in \mathscr{S}$,

$$\binom{N(s)}{m(s)} ,$$

are equally likely, and treatment across strata are independent. If $m(s) = \lfloor \lambda\, N(s) \rfloor$, then $|\mathcal{D}_N(s)| \le 1$ for all $s \in \mathscr{S}$. See (Zelen, 1974; Rosenberger and Lachin, 2015) for discussion.

Example 5. (Sequential Randomization Algorithm) This design, due to Hu and Hu (2012), allows for dependence on multiple strata, and assigns treatment in a recurrent fashion:

$$\mathbb{P}\{A_k = 1|\mathbf{S}_k, \mathbf{A}_{k-1}\} = \begin{cases} \lambda & \text{Imb}_k = 0 \\ \gamma & \text{Imb}_k < 0 \\ 1-\gamma & \text{Imb}_k > 0 \end{cases} ,$$

where Imb_k =Imb$(\mathbf{S}_k, \mathbf{A}_{k-1})$ is defined in the main text, and $\gamma > \lambda$. Imb_k is a weighted average of three types of discrepancies: overall, marginal, and within-stratum.

D Coupling Construction under CAR

Let $\mathcal{S} = |\mathscr{S}| < \infty$ be the total number of strata. Denote $Y_{1,i} = Y_i$ among the treated, and $Y_{0,i} = Y_i$ among the non-treated, and collect all these outcomes in one vector as $\tilde{\boldsymbol{X}} = (X_1, \ldots, X_{\mathcal{S}})$, where each X_s, $1 \le s \le \mathcal{S}$ represents the observed data for stratum s and is given by

$$X_s = \left(X_{s,1}, \ldots, X_{s,N(s)}\right) = \left(Y_{1,s_1}, \ldots, Y_{1,s_{m(s)}}, Y_{0,s_1}, \ldots, Y_{0,s_{n(s)}}\right) .$$

The idea behind the coupling argument in Chung and Romano (2013) is that the behavior of the permutation distribution based on $\tilde{\boldsymbol{X}}$ should behave approximately like the permutation distribution based on a sample of N i.i.d. observations $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_N)$ from the mixture distribution $\bar{P}$, where $\bar{P}(y) = \sum_{s \in \mathscr{S}} p(s)\{\lambda F_1(y|S = s) + (1 - \lambda)F_0(y|S = s)\}$.

Except for ordering, we can construct $\bar{\boldsymbol{X}}$ to include almost the same set of observations as in $\tilde{\boldsymbol{X}}$ as follows. First draw an index j from $\{1, \ldots, \mathcal{S}\}$ with probability $\mathrm{P}(j = s) = p(s)$, $1 \leq s \leq \mathcal{S}$. Next, conditionally on the outcome being $j = s$, draw an index $l \in \{0, 1\}$ with probability $\mathrm{P}(l = 1) = \lambda$. Then conditionally on the outcome being $l = 1$, set $\bar{X}_{s,1} = Y_{1,s_1}$.

Next, draw another index j from $\{1, \ldots, \mathcal{S}\}$ with probability $\mathrm{P}(j = s') = p(s')$. If $s' \neq s$, then conditionally on outcome being s', repeat the same step as before, that is, conditionally on the outcome being $j = s'$, draw an index $l \in \{0, 1\}$ with probability $\mathrm{P}(l = 1) = \lambda$. Then conditionally on the outcome being $l = 1$, set $\bar{X}_{s',1} = Y_{1,s'_1}$. However, if $s' = s$, then conditionally on this outcome, draw an index $i \in \{0, 1\}$ with probability $\mathrm{P}(i = 1) = \lambda$. If $i = 0$, then $\bar{X}_{s,2} = Y_{0,s_1}$; otherwise, set $\bar{X}_{s,2} = Y_{1,s_2}$.

We iterate the previous steps to "couple" our original data—draw another index j from $\{1, \ldots, \mathcal{S}\}$ with probability $\mathrm{P}(j = s'') = p(s'')$. If $s'' \neq s$ and $s'' \neq s'$, then conditionally on outcome being s'', repeat the same step as before, that is, conditionally on the outcome being $j = s''$, draw an index $l \in \{0, 1\}$ with probability $\mathrm{P}(l = 1) = \lambda$. Then conditionally on the outcome being $l = 1$, set $\bar{X}_{s'',1} = Y_{1,s''_1}$. However, if either $s'' = s$ or $s'' = s'$, then conditionally on this outcome, draw an index $k \in \{0, 1\}$ with probability $\mathrm{P}(k = 1) = \lambda$. If $k = 0$ and $s'' = s$, then either $\bar{X}_{s,3} = Y_{0,s_2}$ if $i = 0$ in the previous step, or $\bar{X}_{s,3} = Y_{0,s_3}$ if $i = 1$ in the previous step. Analogously, if $k = 0$ and $s'' = s'$, then $\bar{X}_{s',2} = Y_{0,s'_1}$ if $l = 0$, or $\bar{X}_{s',2} = Y_{1,s'_2}$ if $l = 1$ in the previous step.

Keep repeating this process, noting that there will probably be a point in which you exhaust all the $m(s)$ observations governed by the distribution of $Y_1|S = s$ for some $s \in \{1, \ldots, \mathcal{S}\}$. If this happens and another index s is drawn, then conditionally on s, if $l = 1$ is drawn again, then just sample a new observation from the distribution of $Y_1|S = s$, and analogously if you have exhausted all the $n(s)$ from the distribution of $Y_0|S = s$.

Continue this way so that as many as possible of the original X_s observations are used in the construction of $\bar{X}_s$ for all $s \in \mathcal{S}$. However, it is possible that for some s, we have used the observations from X_s to fill all the $N(s)$ observations in $\bar{X}_s$ and another index s is drawn. If

this happens, then conditionally on s, draw another index $l \in \{0,1\}$, with $\mathrm{P}(l=1)=\lambda$. If $l=1$, then sample a new observation $\bar{X}_{s,N(s)+1}$ from the distribution of $Y_1|S=s$, otherwise sample $\bar{X}_{s,N(s)+1}$ from the distribution of $Y_0|S=s$. Continue this way so that as many as possible of the original $\tilde{\boldsymbol{X}}$ observations are used in the construction of $\bar{\boldsymbol{X}} = \left(\bar{X}_1, \ldots, \bar{X}_{\mathcal{S}}\right)$.

We can reorder the observations in $\bar{\boldsymbol{X}}$ by a permutation π_0 so that for each s, $X_{s,i}$ and $\bar{X}_{s,\pi_0(i)}$ agree for all i except for some hopefully small (random) number $\mathcal{C}_s$. To see how this reordering works, we proceed sequentially, *i.e.*, by stratum. Recall that for the first stratum, X_1 has the observations in order, that is, the first $m(1)$ observations arose from the distribution of $Y_1|S=1$, the next $n(1)$ observations from the distribution of $Y_0|S=1$, and so on and so forth. Thus, to couple $\bar{X}_1$ with X_1, put all observations in $\bar{X}_1$ that came from the distribution of $Y_1|S=1$ in the first $m(1)$ positions. If the number of observations from the distribution of $Y_1|S=1$ is *greater than or equal to* $m(1)$ (recall that this is a possibility), then $\bar{X}_{1,\pi(i)}$ for $i=1,\ldots,m(1)$ are filled according to the observations in $\bar{X}_1$ which came from the distribution of $Y_1|S=1$, and if the number is greater, put them aside for now. On the other hand, if the number of observations in $\bar{X}_1$ which came from the distribution of $Y_1|S=1$ is *less* than $m(1)$, fill up as many of $\bar{X}_1$ from the distribution of $Y_1|S=1$ as possible, and leave the rest of the blank spots for now.

Next, move onto the observations in $\bar{X}_1$ that came from the distribution of $Y_0|S=1$ and repeat the above procedure for $m(1)+1, m(1)+2, \ldots, m(1)+n(s)$ spots in order to complete the observations in $\bar{X}_{1,\pi(i)}$; simply fill up the empty spots with the remaining observations which were put aside (at this point the order does not matter, but chronological order is an option).

Repeat this reordering steps for each subsequent stratum. In the end, this permutation of the observations in $\bar{X}$ corresponds to a permutation π_0 and satisfies, for each $1 \le s \le \mathcal{S}$, $X_{s,i} = \bar{X}_{s,\pi_0(i)}$ for indexes i, except for $\mathcal{C}_s$ of them. The number of observations $\mathcal{C} = \sum_{s\in\mathscr{S}} \mathcal{C}_s$ where $\tilde{\boldsymbol{X}}$ and $\bar{\boldsymbol{X}}_{\pi_0} = \left(\bar{X}_{1,\pi_0}, \ldots, \bar{X}_{\mathcal{S},\pi_0}\right)$ differ is random. We can apply the same reasoning as in the proof of Chung and Romano (2013, (5.8)) and write for each s

$$\mathcal{C}_s = \max\{m(s) - \tilde{m}(s), 0\} + \max\{n(s) - \tilde{n}(s), 0\} \ ,$$

where $\tilde{m}(s)$ denotes the number of observations in $\bar{\boldsymbol{X}}$ which are generated from $Y_1|S=s$, and similarly, $\tilde{n}(s)$ denotes the total number of observations in $\bar{\boldsymbol{X}}$ which are generated from $Y_0|S=s$. Indeed, $(\tilde{m}(1), \tilde{n}(1), \ldots, \tilde{m}(\mathcal{S}), \tilde{n}(\mathcal{S}))$ follows the multinomial distribution based

on N trials and success probabilities $(\lambda\, p(1), (1-\lambda)p(1), \ldots, \lambda\, p(\mathcal{S}), (1-\lambda)p(\mathcal{S}))$. For fixed $1 \leq s \leq \mathcal{S}$

$$\begin{aligned}
\tilde{m}(s) - m(s) &= \{\tilde{m}(s) - N\lambda\, p(s)\} - \{m(s) - N(s)\lambda\} - \{\lambda\, N(s) - N\lambda\, p(s)\} \\
&= \{\tilde{m}(s) - N\lambda\, p(s)\} - \mathcal{D}_N(s) - \lambda\, \{N(s) - N\, p(s)\} \qquad \text{(D.1)} \\
\tilde{n}(s) - n(s) &= \{\tilde{n}(s) - N(1-\lambda)\, p(s)\} + \mathcal{D}_N(s) - (1-\lambda)\, \{N(s) - N\, p(s)\} \ , \qquad \text{(D.2)}
\end{aligned}$$

where $N(s) = m(s) + n(s) = \sum_{1 \leq i \leq N} \mathbb{1}_{\{S_i = s\}}$. The usual central limit theorem implies

$$\begin{aligned}
\tilde{m}(s) - N\lambda\, p(s) &= \mathcal{O}_p\left(N^{1/2}\right) \\
\tilde{n}(s) - N(1-\lambda)\, p(s) &= \mathcal{O}_p\left(N^{1/2}\right) \\
N(s) - N\, p(s) &= \mathcal{O}_p\left(N^{1/2}\right) \ .
\end{aligned}$$

Assumption A.2 *ii)* implies $\mathcal{D}_N(s) = \mathcal{O}_p\left(N^{1/2}\right)$. Then, we conclude that $\mathcal{C} = \mathcal{O}_p\left(N^{1/2}\right)$, and $\mathcal{C}/N \xrightarrow{\mathrm{P}} 0$. Moreover, by arguing as in the proof of Chung and Romano (2013, (5.8)), we have that

$$\mathbb{E}(\mathcal{C}) \leq \mathcal{O}\left(N^{1/2}\right) \ . \qquad \text{(D.3)}$$

To see why, plug Eqs. (D.1)–(D.2) into $\mathcal{C}$ to obtain

$$\begin{aligned}
\mathbb{E}(\mathcal{C}) &\leq \sum_{s \in \mathscr{S}} \mathbb{E}\Big(|\tilde{m}(s) - m(s)|\Big) + \sum_{s \in \mathscr{S}} \mathbb{E}\Big(|\tilde{n}(s) - n(s)|\Big) \\
&\leq \sum_{s \in \mathscr{S}} \Big\{ \Big(\mathbb{E}\Big(\tilde{m}(s) - N\lambda p(s)\Big)^2\Big)^{1/2} + \Big(\mathbb{E}\Big(\tilde{n}(s) - N(s)(1-\lambda)\Big)^2\Big)^{1/2} \\
&+ 2\Big(\mathbb{E}\Big(\mathcal{D}_N(s)\Big)^2\Big)^{1/2} + \Big(\mathbb{E}\Big(N(s) - N\, p(s)\Big)^2\Big)^{1/2} \Big\} \\
&\leq \sum_{s \in \mathscr{S}} \Big\{ \Big(N\lambda p(s)(1 - \lambda p(s)\Big)^{1/2} + \Big(N(1-\lambda)p(s)(1 - p(s) - \lambda p(s)\Big)^{1/2} \\
&+ 2\Big(N\, p(s)\tau(s)\Big)^{1/2} + \Big(N\, p(s)(1 - p(s)\Big)^{1/2} \Big\} \\
&= \mathcal{O}\left(N^{1/2}\right) \ ,
\end{aligned}$$

where $0 \leq \tau(s) \leq \lambda(1-\lambda)$ for all s.

References

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69(1):61–67.

Bai, Y. (2019). Optimality of matched-pair designs in randomized controlled trials. *Available at SSRN 3483834*.

Baldi Antognini, A. and Giovagnoli, A. (2004). A new 'biased coin design' for the sequential allocation of two treatments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(4):651–664.

Baldi Antognini, A. and Zagoraiou, M. (2011). The covariate-adaptive biased coin design for balancing clinical trials in the presence of prognostic factors. *Biometrika*, 98(3):519–535.

Baldi Antognini, A. and Zagoraiou, M. (2015). On the almost sure convergence of adaptive allocation procedures. *Bernoulli*, 21(2):881–908.

Basse, G. W. and Airoldi, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika*, 105(4):849–858.

Begg, C. B. and Iglewicz, B. (1980). A treatment allocation procedure for sequential clinical trials. *Biometrics*, pages 81–90.

Beran, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika*, 74(3):457–468. [DOI].

Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83(403):687–697. [DOI].

Beran, R. and Millar, P. (1986). Confidence sets for a multivariate distribution. *The Annals of Statistics*, pages 431–443. [DOI].

Bertsimas, D., Johnson, M., and Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876.

Birkett, N. J. (1985). Adaptive allocation in randomized controlled trials. *Controlled clinical trials*, 6(2):146–155.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods: Theory and Methods*. Springer Science & Business Media.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics*, 1(4):200–232.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796. [DOI].

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, 10(4):1747–1785.

Butler, D. M. and Broockman, D. E. (2011). Do politicians racially discriminate against constituents? a field experiment on state legislators. *American Journal of Political Science*, 55(3):463–477. [DOI].

Chung, E. and Olivares, M. (2020). Permutation test for heterogeneous treatment effects with a nuisance parameter. *(accepted) Journal of Econometrics*, pages 1–49. Latest version.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41(2):484–507. [DOI].

Chung, E. and Romano, J. P. (2016). Multivariate and multiple permutation tests. *Journal of Econometrics*, 193(1):76–91. [DOI].

Cohen, P. L. and Fogarty, C. B. (2020). Gaussian prepivoting for finite population causal inference. *arXiv preprint arXiv:2002.06654*.

Drew, J. H., Glen, A. G., and Leemis, L. M. (2000). Computing the cumulative distribution function of the kolmogorov–smirnov statistic. *Computational statistics & data analysis*, 34(1):1–15.

Duflo, E. and Banerjee, A. (2017). *Handbook of field experiments*. Elsevier.

Durrett, R. T. and Resnick, S. I. (1977). Weak convergence with random indices. *Stochastic Processes and their Applications*, 5(3):213–220.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417.

Fisher, R. A. (1934). Statistical methods for research workers. *Statistical methods for research workers.*, (5th Ed).

Forsythe, A. B. (1987). Validity and power of tests when groups have been balanced for prognostic factors. *Computational Statistics & Data Analysis*, 5(3):193–200.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, pages 65–70. [Stable URL].

Hu, F., Hu, Y., Ma, Z., and Rosenberger, W. F. (2014). Adaptive randomization for balancing over covariates. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4):288–303. [DOI].

Hu, F. and Rosenberger, W. F. (2006). *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons.

Hu, F., Zhang, L.-X., and He, X. (2009). Efficient randomized-adaptive designs. *The Annals of Statistics*, 37(5A):2543–2560.

Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, 40(3):1794–1815.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters*, 36(1):9–21. [DOI].

Janssen, A. (1999). Testing nonparametric statistical functionals with applications to rank tests. *Journal of Statistical Planning and Inference*, 81(1):71–93. [DOI].

Kuznetsova, O. M. and Tymofyeyev, Y. (2013). Shift in re-randomization distribution with conditional randomization test. *Pharmaceutical Statistics*, 12(2):82–91.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Science & Business Media.

Ma, W., Hu, F., and Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, 110(510):669–680.

Ma, W., Qin, Y., Li, Y., and Hu, F. (2020). Statistical inference for covariate-adaptive randomization procedures. *Journal of the American Statistical Association*, 115(531):1488–1497. [DOI].

Marsaglia, G., Tsang, W. W., and Wang, J. (2003). Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18):1–4.

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

Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *The Annals of Statistics*, 21(4):1760–1779. [DOI].

Pew Research Center (2018). Wide gender gap, growing educational divide in voters' party identification. Pew Research Center. U.S Politics & Policy. Retrieved from https://www.pewresearch.org/politics/2018/03/20/1-trends-in-party-affiliation-among-demographic-groups/.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Science & Business Media.

Qin, Y., Li, Y., Ma, W., and Hu, F. (2018). Pairwise sequential randomization and its properties. *arXiv preprint arXiv:1611.02802*.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, pages 141–159. [DOI].

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Rosenberger, W. F. and Sverdlov, O. (2008). Handling covariates in the design of clinical trials. *Statistical Science*, 23(3):404–419.

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics*, pages 130–134.

Shao, J., Yu, X., and Zhong, B. (2010). A theory for testing hypotheses under covariate-adaptive randomization. *Biometrika*, 97(2):347–360.

Simard, R. and L'Ecuyer, P. (2011). Computing the two-sided kolmogorov-smirnov distribution. *Journal of Statistical Software*, 39(11):1–18.

Simon, R. and Simon, N. R. (2011). Using randomization tests to preserve type i error with response adaptive and covariate adaptive randomization. *Statistics & Probability Letters*, 81(7):767–772.

Smith, R. L. (1984a). Properties of biased coin designs in sequential clinical trials. *The Annals of Statistics*, 12(3):1018–1034.

Smith, R. L. (1984b). Sequential treatment allocation using biased coin designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(3):519–543.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van der Vaart, A. W. and Wellner, J. (1996). *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media.

Wei, L.-J. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, pages 92–100.

Williams, D. (1991). *Probability with martingales*. Cambridge university press.

Ye, T. (2018). Testing hypotheses under covariate-adaptive randomisation and additive models. *Statistical Theory and Related Fields*, 2(1):96–101.

Ye, T., Yi, Y., and Shao, J. (2020). Inference on average treatment effect under minimization and other covariate-adaptive randomization methods. *arXiv preprint arXiv:2007.09576*.

Zelen, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, 27(7):365–375.

Zhang, L.-X., Hu, F., Cheung, S. H., and Chan, W. S. (2007). Asymptotic properties of covariate-adjusted response-adaptive designs. *The Annals of Statistics*, 35(3):1166–1182. [DOI].

Zhang, Y. and Zheng, X. (2020). Quantile treatment effects and bootstrap inference under covariate-adaptive randomization. *Quantitative Economics*, 11(3):957–982.

Zhou, Z., Li, P., and Hu, F. (2020). Adaptive randomization in network data. *arXiv preprint arXiv:2009.01273*.