Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment

Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili & Michael Rosenblum

To cite this article: Bingkai Wang, Ryoko Susukida, Ramin Mojtabai, Masoumeh Amin-Esmaeili & Michael Rosenblum (2021): Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1981338

To link to this article: https://doi.org/10.1080/01621459.2021.1981338

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2021.1981338

Taylor & Francis
Taylor & Francis Group

Model-Robust Inference for Clinical Trials that Improve Precision by Stratified Randomization and Covariate Adjustment

Bingkai Wang[a], Ryoko Susukida[b], Ramin Mojtabai[b], Masoumeh Amin-Esmaeili[b,c], and Michael Rosenblum[a]

[a]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, MD; [b]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, MD; [c]Iranian National Center for Addiction Studies (INCAS), Tehran University of Medical Sciences, Tehran, Iran

ABSTRACT

Two commonly used methods for improving precision and power in clinical trials are stratified randomization and covariate adjustment. However, many trials do not fully capitalize on the combined precision gains from these two methods, which can lead to wasted resources in terms of sample size and trial duration. We derive consistency and asymptotic normality of model-robust estimators that combine these two methods, and show that these estimators can lead to substantial gains in precision and power. Our theorems cover a class of estimators that handle continuous, binary, and time-to-event outcomes; missing outcomes under the missing at random assumption are handled as well. For each estimator, we give a formula for a consistent variance estimator that is model-robust and that fully captures variance reductions from stratified randomization and covariate adjustment. Also, we give the first proof (to the best of our knowledge) of consistency and asymptotic normality of the Kaplan–Meier estimator under stratified randomization, and we derive its asymptotic variance. The above results also hold for the biased-coin covariate-adaptive design. We demonstrate our results using data from three trials of substance use disorder treatments, where the variance reduction due to stratified randomization and covariate adjustment ranges from 1% to 36%. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received December 2020
Accepted September 2021

KEYWORDS
Covariate-adaptive randomization; Generalized linear model; Robustness

1. Introduction

A joint guidance document from the U.S. Food and Drug Administration and the European Medicines Agency (FDA and EMA 1998) states that "Pretrial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis to improve precision and to compensate for any lack of balance between treatment groups." More recent regulatory guidance documents also encourage consideration of baseline variables in order to improve precision in randomized trials (EMA 2015; FDA 2020, 2021). There is a rich literature on model-robust, statistical methods to adjust for baseline variables and improve precision in randomized trials that use simple randomization, for example, Koch et al. (1998), Yang and Tsiatis (2001), Rubin and van der Laan (2008), Tsiatis et al. (2008), Moore and van der Laan (2009a), Moore and van der Laan (2009b), Zhang (2015), and Jiang et al. (2018). However, less is known for trials that use other forms of randomization. This is a practical concern since, as discussed below, many clinical trials use other forms of randomization.

"Covariate-adaptive randomization" refers to randomization procedures that take baseline variables into account when assigning participants to study arms. The goal is to achieve better balance across study arms in preselected strata of the baseline variables compared to simple randomization (which ignores baseline variables). For example, balance on disease severity, a genetic marker, or another variable thought to be correlated with the primary outcome could be sought. The simplest and most commonly used type of covariate-adaptive randomization is stratified permuted block randomization (Zelen 1974), referred to as "stratified randomization" throughout, for conciseness.

Compared with simple randomization, covariate-adaptive randomization can be advantageous in minimizing imbalance and improving efficiency (Efron 1971; Pocock and Simon 1975; Wei 1978). Due to these benefits, covariate-adaptive randomization has become a popular approach in clinical trials. According to a survey by Lin, Zhu, and Su (2015), 183 out of their sample of 224 randomized clinical trials published in 2014 in leading medical journals used some form of covariate-adaptive randomization. Stratified randomization was implemented by 70% of trials in this survey. Another method for covariate-adaptive randomization is the biased-coin design by Efron (1971), which we call "biased-coin randomization" throughout. Other examples include Wei's urn design (Wei 1978) and rerandomization (Morgan and Rubin 2012). We only consider the following two types of covariate-adaptive randomization: stratified randomization and biased-coin randomization.

Concerns have been raised regarding how to perform valid statistical analyses at the end of trials that use covariate-adaptive randomization. Adjusting for stratification variables is recommended (Lachin, Matts, and Wei 1988; Kahan and Morris 2012; EMA 2015). However, this recommendation is not reliably carried out. Kahan and Morris (2012) sampled 65 published trials from major medical journals from March to May 2010 and found that 41 implemented covariate-adaptive randomization

CONTACT Michael Rosenblum mrosenbl@jhsph.edu Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Room E3616, Baltimore, MD 21205, USA

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

(among which 29 used stratified randomization), but only 14 adjusted in the primary analysis for the variables used in the randomization procedure. Furthermore, many results on how to conduct the primary efficacy analysis in trials that use stratified randomization require one to assume a correctly specified regression model for the outcome given study arm assignment and baseline variables, for example, Shao, Yu, and Zhong (2010), Shao and Yu (2013), Ma, Hu, and Zhang (2015), Ma et al. (2018), and Yang et al. (2020). Our focus is on model-robust estimators, that is, estimators that do not require such an assumption when there is no missing data or when outcome data are missing completely at random. We also consider a related property when data are missing at random, for estimators that involve multiple working models.

Yang and Tsiatis (2001) showed that the analysis of covariance (ANCOVA) estimator (both with and without treatment by baseline variable interactions) is consistent and asymptotically normal under simple randomization, and that this holds under arbitrary misspecification of the linear regression model used to construct the estimator. Analogous results for the ANCOVA estimator were shown by Bugni, Canay, and Shaikh (2018, 2019) under a variety of covariate-adaptive randomization procedures that include stratified and biased-coin randomization; however, their results only allow adjustment for the variables used in the randomization procedure. The proofs of our results build on key ideas from their work as described below. Ye, Shao, and Zhao (2020a) and Ye, Yi, and Shao (2020b) extended the results of Bugni, Canay, and Shaikh (2018, 2019) to allow adjustment for additional baseline variables. The results of Li and Ding (2020) and Liu and Yang (2020) for the ANCOVA estimator are robust to arbitrary misspecification of the linear regression model; however, they use the randomization inference framework while many clinical trials are analyzed using the superpopulation inference framework (as done here); see Robins (2002) for a comparison of these frameworks. All of the results in this paragraph are for the ANCOVA estimator, and so do not apply to logistic regression models for binary outcomes nor to commonly used models for time-to-event outcomes. Ye and Shao (2020) derived asymptotic distributions for log-rank and score tests in survival analysis under covariate-adaptive randomization; however, estimation was not addressed.

For trials using stratified or biased-coin randomization, to the best of our knowledge, it was an open problem to determine (in the commonly used superpopulation inference framework and without making parametric model assumptions) the large sample properties of estimators such as the following: covariate adjustment for binary outcomes using logistic regression, the mixed-effects model for repeated measures (MMRM) estimator, and the Kaplan-Meier estimator for survival outcomes. We derived the large sample properties for these (and other) estimators, which we think can be important for the analysis of clinical trials. For example, binary and time-to-event outcomes are commonly used in clinical trials. According to a survey by Austin et al. (2010) of trials published in leading medical journals in 2007, 74 out of 114 trials involved binary or time-to-event outcomes. As we show in our data analyses, the addition of baseline variables beyond those used for stratified randomization can lead to substantial precision gains.

Under regularity conditions, we prove that a large class of estimators is consistent and asymptotically normally distributed in randomized trials that use stratified or biased-coin randomization, and we give a formula for computing their asymptotic variance. This class of estimators consists of all M-estimators that are consistent under simple randomization. Examples are listed in Section 4. We prove analogous results for the Kaplan–Meier (K-M) estimator (Kaplan and Meier 1958) of the survival function. Underlying these results is our general technique for characterizing the large sample behavior of asymptotically linear estimators under stratified or biased-coin randomization, described in Section 7.

Our theorems imply that under standard regularity conditions, whenever an estimator in our class is consistent and asymptotically normally distributed under simple randomization, then it is consistent and asymptotically normally distributed under stratified (or biased-coin) randomization. Also, its influence function is the same regardless of whether data are generated under simple, stratified or biased-coin randomization. This can be advantageous since for many estimators used to analyze randomized trials, their influence functions have already been derived under simple randomization. An estimator's influence function can be input into our formula (4) to produce a consistent variance estimator under stratified and biased-coin randomization.

As in the aforementioned work, we assume that the randomization procedure and analysis method have been completely specified before the trial starts, as is typically required by regulators (FDA and EMA 1998; EMA 2015; FDA 2020, 2021).

In the next section, we describe three trial examples to which we apply our methods. In Section 3, we describe our setup, notation and assumptions. We present our main results in Section 4. In Section 5, we give example estimators for continuous and binary outcomes to which our general results apply. In Section 6, we present asymptotic results for the Kaplan–Meier estimator for time-to-event outcomes. Trial applications are provided in Section 7. Practical recommendations and future directions are discussed in Section 8.

2. Three Completed Trials That Used Stratified Randomization

In some cases, the outcomes in our analyses differ from the primary outcomes in the corresponding trials. This is because we wanted similar outcomes across trials for illustration.

2.1. *Buprenorphine Tapering and Illicit Opioid Use (NIDA-CTN-0003)*

The trial of "Suboxone (Buprenorphine/Naloxone) Taper: A Comparison of Two Schedules" trial in the National Drug Abuse Treatment Clinical Trials Network (NIDA-CTN-0003), is a phase-3 randomized trial completed in 2005 (Ling et al. 2009). The goal was to compare the effects of a short or long taper schedule after buprenorphine stabilization of patients with opioid use disorder. Patients were randomized into two arms: 28-day taper (control, 259 patients, 36% missing outcomes) and 7-day taper (treatment, 252 patients, 21%

missing outcomes), stratified by maintenance dose (3 levels) measured at randomization. The outcome of interest is a binary indicator of whether a participant's urine tested at the end of the study is opioid-free (encoded as 0) or not (encoded as 1). In addition to the stratification variable, we adjust for the following baseline variables: sex, opioid urine toxicology results, the Adjective Rating Scale for Withdrawal (ARSW), the Clinical Opiate Withdrawal Scale (COWS) and the Visual Analog Scale (VAS).

2.2. Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling (NIDA-CTN-0030)

The Two-Phase Randomized Controlled Clinical Trial of Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence (NIDA-CTN-0030) is a phase-3 randomized trial completed in 2013 (Weiss et al. 2011). The goal was to determine whether adding individual drug counseling to the prescription of buprenorphine/naloxone would improve outcomes for patients with prescription opioid use disorder. Though this study adopted a 2-phase adaptive design, we focus on the first phase, in which patients were randomized into standard medical management (control, 330 patients, 10% missing outcomes) or standard medical management plus drug counseling (treatment, 335 patients, 13% missing outcomes). Randomization was stratified by the presence or absence of (i) a history of heroin use and (ii) current chronic pain, resulting in 4 strata. The outcome of interest is the proportion of positive urine laboratory results among all tests (treated as a continuous outcome between 0 and 1). Among all 5 urine laboratory tests during the first 4 weeks of phase I, if a patient missed two consecutive visits, then the outcome is regarded as missing. We included the following baseline variables in the analysis: randomization stratum, age, sex and urine laboratory results.

2.3. Internet-Delivered Treatment for Substance Use Disorders (NIDA-CTN-0044)

The phase-3 randomized trial Web-delivery of Evidence-Based, Psychosocial Treatment for Substance Use Disorders (NIDA-CTN-0044) was completed in 2012 (Campbell et al. 2014). The goal was to evaluate the effectiveness of a web-delivered behavioral intervention, Therapeutic Education System (TES), in the treatment of substance use disorder. Participants were randomly assigned to two arms: treatment as usual (control, 252 participants, 19% missing outcomes) and treatment as usual plus TES (treatment, 255 participants, 18% missing outcomes).

Randomization was stratified by site, patient's primary substance of use (stimulant or nonstimulant) and abstinence status at baseline. Unfortunately, the available dataset for this trial did not include the site variable. Our analyses and claims in Section 7 assume that the only randomization strata are the patient's primary substance of use and abstinence status at baseline (4 levels overall). Our theorems imply that ignoring one or more randomization stratum variables leads to conservative variance estimates when using our variance formulas, as explained in Section 8.

After randomization, each participant was followed for 12 weeks with 2 urine laboratory tests per week. The outcome of interest is the proportion of positive urine lab results among all tests (treated as a continuous outcome between 0 and 1). If a participant missed visits of more than 6 weeks, then the outcome is regarded as missing. We adjust for randomization stratum and the following additional baseline variables: age, sex, and urine laboratory result.

We also analyze a second outcome: time to abstinence, defined as the time to first two consecutive negative urine tests during the study. Censoring time is defined as the first missing visit. We used the data from the first 6 weeks of follow-up in our data analysis of this time-to-event outcome, during which 99% of the events occurred.

3. Definitions and Assumptions

3.1. Data-Generating distributions

We focus on two-arm randomized trials that use simple, stratified or biased-coin randomization. Let n denote the sample size. For each participant $i = 1, \ldots, n$, let Y_i denote the primary outcome, M_i denote whether Y_i is observed ($M_i = 1$) or missing ($M_i = 0$), A_i denote study arm assignment ($A_i = 1$ if assigned to treatment and $A_i = 0$ if assigned to control), and $\boldsymbol{X}_i$ denote a vector of baseline covariates. This notation is for real-valued outcomes, for example, continuous or binary outcomes. Modified definitions, assumptions, and results for time-to-event outcomes are in Section 6.

We use the Neyman–Rubin potential outcomes framework (Neyman, Dabrowska, and Speed 1990), which assumes the existence of potential outcomes $Y_i(0)$ and $Y_i(1)$ for each participant i. These represent the outcome that would be observed under assignment to study arm 0 or 1, respectively. We make the following consistency assumption linking the observed outcome Y_i to the potential outcomes: $Y_i = Y_i(A_i) = Y_i(1)A_i + Y_i(0)(1 - A_i)$ for each participant i. Also, let $M_i(a)$ be the indicator of whether participant i would have a nonmissing outcome if they get assigned to study arm $a \in \{0, 1\}$. We assume, analogous to the consistency assumption above, that $M_i = M_i(A_i) = M_i(1)A_i + M_i(0)(1 - A_i)$.

For each participant i, we define the full data vector (including potential outcomes, some of which are not observed) $\boldsymbol{W}_i = (Y_i(1), Y_i(0), M_i(1), M_i(0), \boldsymbol{X}_i)$ and the observed data vector $\mathbf{O}_i = (A_i, \boldsymbol{X}_i, Y_i M_i, M_i)$. The reason that the product $Y_i M_i$ appears in $\mathbf{O}_i$ is to encode that whenever the outcome is missing ($M_i = 0$), the outcome value Y_i is not available in $\mathbf{O}_i$; also, including $Y_i M_i$ in $\mathbf{O}_i$ is useful in simplifying the estimating equations for some of our examples, as described in the supplementary materials.

We make the following assumptions on the distribution of $\{\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n\}$:

Assumption 1. (i) $\boldsymbol{W}_i, i = 1, \ldots, n$ are independent, identically distributed samples from an unknown joint distribution P on $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$.

(ii) Missing at random: $M(a) \perp\!\!\!\perp Y(a) | \boldsymbol{X}$ for each arm $a \in \{0, 1\}$, where $\perp\!\!\!\perp$ denotes independence.

Throughout, we use E to denote the expectation with respect to distribution P.

3.2. Randomization Procedures: Simple, Stratified, and Biased-Coin

First consider simple randomization, which assigns study arms $A_1, \ldots, A_n$ by independent Bernoulli draws each with fixed probability π of being 1, for example, using a random number generator. By design, the draws are independent of each other and of all participant characteristics measured before randomization or not impacted by randomization. Therefore, we have that $(A_1, \ldots, A_n)$ is independent of $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$, and that the observed data $\mathbf{O}_1, \ldots, \mathbf{O}_n$ are independent, identically distributed.

Next consider stratified or biased-coin randomization, where treatment allocation depends on predefined baseline strata, such as gender, age, site, disease severity, or combinations of these. We refer to the baseline strata that are used in the randomization procedure as "randomization strata." The baseline stratum of participant i is denoted by the single, categorical variable S_i taking K possible values. For example, if randomization strata are defined by 4 sites and a binary indicator of high disease severity, then S has $K = 8$ possible values. Let S_i denote the stratification variable for participant i and let $\mathcal{S} = \{1, \ldots, K\}$ denote the set of all K randomization strata. The goal of stratified or biased-coin randomization is to achieve balance in each stratum; that is, the proportion of participants assigned to the treatment arm is targeted to the prespecified proportion $\pi \in (0, 1)$, for example, $\pi = 0.5$. Throughout, the stratification variable S is encoded in the baseline covariate vector $\boldsymbol{X}$ using $K - 1$ dummy variables that make up the first $K-1$ components of $\boldsymbol{X}$ (which can include additional baseline variables).

Stratified randomization uses permuted blocks to assign treatment. For each randomization stratum, a randomly permuted block with fraction π 1's (representing treatment) and $(1 - \pi)$ 0's (representing control) is used for sequential allocation. When a block is exhausted, a new block is used.

Biased-coin randomization can be applied when $\pi = 0.5$ and it allocates participants sequentially by the following rule for $k = 1, \ldots, n$:

$$P(A_k = 1 | S_1, \ldots, S_k, A_1, \ldots, A_{k-1}) = \begin{cases} 0.5, & \text{if } \sum_{i=1}^{k-1}(A_i - 0.5) I\{S_i = S_k\} = 0 \\ \lambda, & \text{if } \sum_{i=1}^{k-1}(A_i - 0.5) I\{S_i = S_k\} < 0 \\ 1 - \lambda, & \text{if } \sum_{i=1}^{k-1}(A_i - 0.5) I\{S_i = S_k\} > 0 \end{cases}$$

where $\lambda \in (0.5, 1]$, $I\{Z\}$ is the indicator function that has value 1 if Z is true and 0 otherwise, and by convention the first participant is assigned with probability 0.5 to each arm. Our results for biased-coin randomization assume that $\pi = 0.5$.

When comparing the three types of randomization procedures (simple, stratified, or biased-coin), we assume that all use the same value of π. For the stratified randomization and biased-coin designs, it follows by construction (and was shown by Bugni, Canay, and Shaikh 2018) that the study arm assignments $(A_1, \ldots, A_n)$ are conditionally independent of the participant baseline variables and potential outcomes $(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n)$ given the randomization strata $(S_1, \ldots, S_n)$. Intuitively, this is because the study arm assignment procedure only has access to the participants' randomization strata. Under stratified or biased-coin randomization, the observed data vectors $\mathbf{O}_1, \ldots, \mathbf{O}_n$ are not independent.

Under any of the three randomization procedures, the observed data vectors $\mathbf{O}_1, \ldots, \mathbf{O}_n$ are identically distributed; that is, the distribution of $\mathbf{O}_1$ is the same as that of $\mathbf{O}_2$, etc. Let P^* denote this distribution, that is, the distribution of a generic, observed data vector $\mathbf{O} = (A, \boldsymbol{X}, YM, M)$. This distribution is the same for each of the three randomization procedures, and is that induced by first drawing a single realization $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$ from the distribution P (see Assumption 1), then drawing A as an independent Bernoulli draw with probability π of being 1, and lastly applying the consistency assumptions $Y = Y(1)A + Y(0)(1 - A)$ and $M = M(1)A + M(0)(1 - A)$ to construct Y, the (non)-missingness indicator M, and their product YM. The corresponding expectation with respect to P^* is denoted E^*, which is used below. The claims in this paragraph are proved in the supplementary material.

3.3. Targets of Inference (Estimands) and Estimators

For continuous and binary outcomes, our goal is to estimate a population parameter $\underline{\Delta}^*$, which is a contrast between the marginal distributions of $Y(1)$ and $Y(0)$. For example, Δ^* can be defined as the population average treatment effect $E[Y(1)] - E[Y(0)]$.

We consider M-estimators of Δ^* (van der Vaart 1998, chap. 5). Let $\boldsymbol{\theta} = (\Delta, \boldsymbol{\beta}^t)^t$ denote a column vector of $p + 1$ parameters where $\Delta \in \mathbb{R}$ is the parameter of interest and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a column vector of p nuisance parameters. We define the M-estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\Delta}, \widehat{\boldsymbol{\beta}}^t)^t$ to be the solution to the following estimating equations:

$$\sum_{i=1}^{n} \boldsymbol{\psi}(A_i, \boldsymbol{X}_i, Y_i, M_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{1}$$

where $\boldsymbol{\psi}$ is a column vector (with $p + 1$ components) of known functions. We define $\widehat{\Delta}$ to be the estimator of Δ^*. We assume that $\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})$ does not depend on the outcome Y when $M = 0$ (since then Y is missing). Many estimators used in clinical trials can be expressed as solutions to estimating equations (1) for an appropriately chosen estimating function $\boldsymbol{\psi}$; see Sections 4 and 5 for examples.

For time-to-event outcomes, the K-M estimator of the survival curve is commonly used. Since it is not an M-estimator, our general result (Theorem 1) for M-estimators below does not apply. We separately prove analogous results for the K-M estimator; see Section 6.

We assume regularity conditions similar to the classical conditions that are used for proving consistency and asymptotic linearity of M-estimators for independent, identically distributed data, as given in (van der Vaart 1998, sec. 5.3). One of the conditions is that $E^*[\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})] = \mathbf{0}$ has a unique solution in $\boldsymbol{\theta}$, which is denoted as $\underline{\boldsymbol{\theta}} = (\underline{\Delta}, \underline{\boldsymbol{\beta}}^t)^t$. The other regularity conditions are given in the supplementary material.

We assume that the estimating equations $\boldsymbol{\psi}$ were chosen to ensure that the property $\Delta^* = \underline{\Delta}$ holds. This property is generally needed to show consistency of the M-estimator $\widehat{\Delta}$ for Δ^* under simple randomization, and has previously been proved for all of the estimators in Section 5.2. In general, whether the property $\Delta^* = \underline{\Delta}$ holds does not depend on the randomization

procedure (simple, stratified, or biased-coin randomization); this is because the property depends only on $\boldsymbol{\psi}$, P and P^*.

Results in (van der Vaart 1998, sec. 5.3) imply that under simple randomization, given Assumption 1 and the regularity conditions in the supplementary material, $\widehat{\Delta}$ converges in probability to $\underline{\Delta}$ and is asymptotically normally distributed with asymptotic variance that we denote by $\widetilde{V}$. We focus on determining what happens under stratified or biased-coin randomization, where our main result (Section 4) is that consistency and asymptotic normality still hold but the asymptotic variance may be smaller (and a consistent variance estimator is given).

4. Main Result for M-estimators

Consider the setup in Section 3.3, where the M-estimator $\widehat{\Delta}$ is defined. The proof of the following theorem (and all results in the paper) is given in the supplementary material:

Theorem 1. Assume the regularity conditions in the supplementary material, $\Delta^* = \underline{\Delta}$, and Assumption 1. Then under simple, stratified, or biased-coin randomization, we have consistency, that is, $\widehat{\Delta} \to \Delta^*$ in probability, and asymptotic linearity, that is,

$$\sqrt{n}(\widehat{\Delta} - \Delta^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(A_i, \boldsymbol{X}_i, Y_i, M_i) + o_p(1), \quad (2)$$

where the influence function $IF(A, \boldsymbol{X}, Y, M)$ is the first entry of $-\boldsymbol{B}^{-1}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \underline{\boldsymbol{\theta}})$ for $\boldsymbol{B} = E^*\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\underline{\boldsymbol{\theta}}}\right]$.

For stratified and biased-coin randomization, $\sqrt{n}(\widehat{\Delta} - \Delta^*) \xrightarrow{d} N(0, V)$ for

$$V = \widetilde{V} - \frac{1}{\pi(1-\pi)} E^*\left[E^*\{(A-\pi)IF(A, \boldsymbol{X}, Y, M)|S\}^2\right], \quad (3)$$

where $\widetilde{V} = E^*\{IF(A, \boldsymbol{X}, Y, M)^2\}$ is the asymptotic variance under simple randomization. The asymptotic variance V can be consistently estimated by formula (4).

Theorem 1 implies that whenever an M-estimator $\widehat{\Delta}$ is consistent and asymptotically normally distributed under simple randomization, then it is consistent and asymptotically normally distributed under stratified (or biased-coin) randomization with equal or smaller asymptotic variance. Also, its influence function is the same regardless of whether data are generated under simple, stratified, or biased-coin randomization.

For the unadjusted estimator $\widehat{\Delta} = \sum_{i=1}^n Y_i A_i / \sum_{i=1}^n A_i - \sum_{i=1}^n Y_i(1-A_i) / \sum_{i=1}^n (1-A_i)$, our Theorem 1 is equivalent to (Bugni, Canay, and Shaikh 2018, theor. 4.1) under stratified or biased-coin randomization. In the special case of continuous outcomes, if the ANCOVA estimator is used with $X = S$, then Theorem 1 is equivalent to the result in (Bugni, Canay, and Shaikh 2018, sec. 4.2) under stratified or biased-coin randomization, though their results also handle other types of covariate-adaptive randomization.

Theorem 1 extends the results of Bugni, Canay, and Shaikh (2018) to handle the class of M-estimators, that is, estimators calculated by solving estimating equations (1). This includes, for example, the standardized logistic regression estimator for binary outcomes (Example 1 of Section 5.2), the DR-WLS estimator (Example 2 of Section 5.2), and the maximum likelihood (ML) estimator corresponding to an MMRM (Example 3 in Section 5.4). This class of estimators also includes the inverse-probability-weighted estimator (IPW, Robins, Rotnitzky, and Zhao 1994), the augmented inverse probability weighted (AIPW) estimator (Robins, Rotnitzky, and Zhao 1994; Scharfstein, Rotnitzky, and Robins 1999), and targeted maximum likelihood estimators (TMLE) that converge in 1-step (van der Laan and Gruber 2012), among others. Thus, Theorem 1 covers estimators that handle various outcome types, repeated measures outcomes, missing outcome data, and covariate adjustment. Our proof relies on key ideas from Lemmas B.1 and B.3 in the supplement of Bugni, Canay, and Shaikh (2018).

In order to construct confidence intervals (CIs) and perform hypothesis tests, one can use the following estimator for the asymptotic variance V, which is the empirical counterpart of the right-hand side of Equation (3):

$$\widehat{V} = \widetilde{V}_n - \frac{1}{\pi(1-\pi)} E_n\left[E_n\{(A-\pi)IF(A, \boldsymbol{X}, Y, M)|S\}^2\right], \quad (4)$$

where $\widetilde{V}_n$ is the sandwich variance estimator of $\widehat{\Delta}$ (Tsiatis 2007, sec. 3.2), defined as the first-row first-column entry of

$$\left\{E_n\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]\right\}^{-1} \left\{E_n\left[\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \widehat{\boldsymbol{\theta}})\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \widehat{\boldsymbol{\theta}})^t\right]\right\} \left\{E_n\left[\frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}\right]\right\}^{-1,t},$$

and E_n denotes expectation with respect to the empirical distribution of the observed data $\mathbf{O}_1, \ldots, \mathbf{O}_n$. Then a CI for Δ^* can be constructed based on the normal approximation with variance $\widehat{V}/n$. We show in the supplementary material that $\widehat{V}$ is a consistent estimator of V.

5. Example Estimators for Continuous and Binary Outcomes

5.1. Definition of Model-Robustness

We consider three examples of M-estimators, to which Theorem 1 can be applied. Each estimator uses working models, that is, models used to compute the estimator, but that are not assumed to be correctly specified. Each estimator involves an outcome working model, defined as a model of the outcome given study arm assignment and baseline variables. One of the estimators (DR-WLS) also uses a working model for the probability of the outcome being missing given study arm assignment and baseline variables.

We call an estimator "model-robust" if it is consistent and asymptotically normal under arbitrary misspecification of the outcome working model, when there are no missing data or the outcome data are missing completely at random. The estimators in all three examples have this model-robustness property under each of the randomization procedures considered in this paper (simple, stratified, and biased-coin). The DR-WLS estimator has the further "double robustness" property of being consistent and

asymptotically normal if at least one of its working models is correctly specified, when outcome data are missing at random. The proofs of the above results are in the supplementary material.

5.2. Standardized Logistic Regression and DR-WLS

For estimators defined in Examples 1 and 2, the parameter of interest, that is, Δ^*, is the average treatment effect defined as $E[Y(1)] - E[Y(0)]$, and we denote $\boldsymbol{Z} = (1, A, \boldsymbol{X}^t)^t$. In Example 1, we assume no missing data.

Example 1. For binary outcomes, the standardized logistic regression estimator $\widehat{\Delta}_{\text{logistic}}$ is calculated by first fitting a working model: $P(Y = 1|A, \boldsymbol{X}) = \text{expit}(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X})$, where $\text{expit}(x) = 1/(1 + e^{-x})$, and getting the maximum likelihood estimates $(\widehat{\beta}_0, \widehat{\beta}_A, \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t)^t$. Then define $\widehat{\Delta}_{\text{logistic}} = \frac{1}{n}\sum_{i=1}^{n}\{\text{expit}(\widehat{\beta}_0 + \widehat{\beta}_A + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i) - \text{expit}(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i)\}$. Equivalently, the estimator $\widehat{\Delta}_{\text{logistic}}$ is the solution to estimating Equations (1) letting

$$\boldsymbol{\psi}(A, \boldsymbol{X}, Y, M; \boldsymbol{\theta}) = \begin{pmatrix} \text{expit}(\beta_0 + \beta_A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}) - \text{expit}(\beta_0 + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}) - \Delta \\ \{Y - \text{expit}(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X})\}\boldsymbol{Z} \end{pmatrix}.$$

This estimator is mentioned as potentially useful in COVID-19 treatment and prevention trials in a recent FDA guidance (FDA 2020).

Example 2. When some outcomes are missing and the missing at random assumption holds, then one can estimate Δ^* by the DR-WLS estimator, which generalizes the estimator in Example 1. The DR-WLS estimator can be used with binary or continuous outcomes. The estimator is calculated by first fitting the logistic regression working model:

$$P(M = 1|A, \boldsymbol{X}) = \text{expit}(\alpha_0 + \alpha_A A + \boldsymbol{\alpha}_{\boldsymbol{X}}^t \boldsymbol{X}) \tag{5}$$

and getting the ML estimates $(\widehat{\alpha}_0, \widehat{\alpha}_A, \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}^t)^t$ of parameters $(\alpha_0, \alpha_A, \boldsymbol{\alpha}_{\boldsymbol{X}}^t)^t$. Next, fit the following working model for the outcome given study arm and baseline variables (from the generalized linear model family):

$$E[Y|A, \boldsymbol{X}] = g^{-1}(\beta_0 + \beta_A A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}), \tag{6}$$

with weights $1/\text{expit}(\widehat{\alpha}_0 + \widehat{\alpha}_A A_i + \widehat{\boldsymbol{\alpha}}_{\boldsymbol{X}}^t \boldsymbol{X}_i)$ using only the data with $M_i = 1$. Here the inverse link function is $g^{-1}(x) = x$ for continuous outcomes and $g^{-1}(x) = \text{expit}(x)$ for binary outcomes. Third, the DR-WLS estimator is

$$\widehat{\Delta}_{\text{DR-WLS}} = \frac{1}{n}\sum_{i=1}^{n}\{g^{-1}(\widehat{\beta}_0 + \widehat{\beta}_A + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i) - g^{-1}(\widehat{\beta}_0 + \widehat{\boldsymbol{\beta}}_{\boldsymbol{X}}^t \boldsymbol{X}_i)\}.$$

The DR-WLS estimator can be expressed as the solution to estimating equations (given in the supplementary material) of the general form (1). For the DR-WLS estimator, we assume that at least one of the two working models (5) and (6) is correctly specified, and $\inf_{(a,\mathbf{x})\in(\mathcal{A},\mathcal{X})} P(M = 1|a, \mathbf{x}) > 0$, where $(\mathcal{A}, \mathcal{X})$ is the support of $(A, \boldsymbol{X})$.

The ANCOVA estimator and the standardized logistic regression estimator are special cases of the DR-WLS estimator. If there are no missing data, which means $M_i = 1$ for $i = 1, \ldots, n$, and the regression weights used to fit (6) are constant, then $\widehat{\Delta}_{\text{DR-WLS}}$ reduces to the ANCOVA estimator for continuous outcomes and to $\widehat{\Delta}_{\text{logistic}}$ for binary outcomes. (The ANCOVA estimator for Δ^* involves first fitting a linear regression working model $E[Y|A, \boldsymbol{X}] = \beta_0 + \Delta A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X}$ using ordinary least squares and then letting $\widehat{\Delta}$ be the estimate of Δ.) The DR-WLS estimator can be generalized to allow the addition of interaction terms in the model (6).

5.3. Asymptotic Results for Estimators in Examples 1 and 2

Under simple randomization and assuming that $\Delta^* = \underline{\Delta}$, consistency and asymptotic normality for the estimators in Examples 1 and 2 have been proved by Scharfstein, Rotnitzky, and Robins (1999) and Robins et al. (2007), respectively. Under stratified or biased-coin randomization, Theorem 1 applies to these estimators since each is an M-estimator. In particular, under the conditions in the theorem, each of the three estimators is consistent and asymptotically normal with asymptotic variance that is consistently estimated by use disorder (4).

Under the additional conditions (a)–(c) listed in the corollary below, for each estimator in Examples 1 and 2, its asymptotic variance is the same regardless of whether simple, stratified, or biased-coin randomization is used; also, the asymptotic variance is consistently estimated by the sandwich variance estimator $\widetilde{V}_n$. Under such conditions, the estimators and their corresponding sandwich variance estimators can be used to perform hypothesis tests and construct CIs that are asymptotically correct.

Recall that we assume throughout that S is encoded by dummy variables in $\boldsymbol{X}$.

Corollary 1. Assume that $\Delta^* = \underline{\Delta}$, the regularity conditions in the Supplementary Material, and Assumption 1. Consider the standardized logistic regression estimator. If any of the conditions (a)–(c) below holds, then under simple, stratified, or biased-coin randomization, the estimator is consistent and asymptotically normally distributed with asymptotic variance $V = \widetilde{V}$; furthermore, the sandwich variance estimator is consistent. Conditions:

(a) $\pi = 0.5$;
(b) the outcome regression model (6) includes indicators for the randomization strata and also treatment-by-randomization-strata interaction terms;
(c) the outcome regression model (6) is correctly specified.

The claims in Corollary 1 also hold for the DR-WLS estimator if at least one of the two working models (5) and (6) is correctly specified and $\inf_{(a,\mathbf{x})\in(\mathcal{A},\mathcal{X})} P(M = 1|a, \mathbf{x}) > 0$, where $(\mathcal{A}, \mathcal{X})$ is the support of $(A, \boldsymbol{X})$.

5.4. MMRM for Repeatedly Measured Continuous Outcomes

Example 3. Consider the scenario where the outcome is continuous and repeatedly measured at $K > 1$ visits, and the mixed-effects model for repeated measures (MMRM) is used (Mallinckrod et al. 2008). We define $\boldsymbol{Y}(a) = (Y_1(a), \ldots, Y_K(a))$ and $\boldsymbol{M}(a) = (M_1(a), \ldots, M_K(a))$ for $a = 0, 1$, where for each $t = 1, \ldots, K$, $Y_t(a)$ is the potential outcome and $M_t(a)$

is an indicator of whether $Y_t(a)$ is nonmissing at visit t under assignment to study arm a. The notation in Section 3 is modified by substituting $\boldsymbol{Y}(a)$ and $\boldsymbol{M}(a)$ for $Y(a)$ and $M(a)$, respectively.

We make Assumption 1 (i) but change Assumption 1 (ii) from missing at random to missing completely at random, that is, we assume that $\boldsymbol{M}(a) \perp\!\!\!\perp (\boldsymbol{Y}(a), \boldsymbol{X})$ for $a = 0, 1$, and that $\boldsymbol{M}(1)$ and $\boldsymbol{M}(0)$ are identically distributed. We further assume that $P(M_1(a) = 1, \ldots, M_K(a) = 1) > 0$ for $a = 0, 1$. The parameter of interest is the average treatment effect at the last visit, that is, $\Delta^* = E[Y_K(1)] - E[Y_K(0)]$.

The MMRM working model (whose correct specification is not assumed when establishing the model-robustness property in the next paragraph) is defined as follows:

$$Y_t = \beta_{0t} + \beta_{At} A + \boldsymbol{\beta}_{\boldsymbol{X}}^t \boldsymbol{X} + \varepsilon_t, \text{ for each } t = 1, \ldots, K, \quad (7)$$

where $\boldsymbol{\epsilon} = (\varepsilon_1, \ldots, \varepsilon_K)$ is independent of $(A, \boldsymbol{X})$ and has a multivariate normal distribution with mean $\boldsymbol{0}$ and covariance Σ. Under the working model, $\boldsymbol{\epsilon}_i, i = 1, \ldots, n$, are assumed to be independent, identically distributed draws from $N(\boldsymbol{0}, \Sigma)$. The working model uses an unstructured covariance matrix Σ, that is, the only assumption is that Σ is positive-definite (and that Σ does not depend on A or $\boldsymbol{X}$). Such model specifications were also used by Mallinckrod et al. (2008), and Lane (2008). The coefficients $\beta_{0t}, \beta_{At}, \boldsymbol{\beta}_{\boldsymbol{X}}$ in (7) are fixed effects. As described by Mallinckrod et al. (2008), the working model (7) is called MMRM because it is derived from a linear mixed-effects model that is similar to (7) but with random effects as well; the model (7) results from marginalizing over the random components, which then implicitly get represented in Σ. Our results below also hold when a different correlation structure, for example, lag 1 autoregressive structure, is used in the working model.

The ML estimator $\widehat{\beta}_{AK}$ of the parameter β_{AK} in the outcome working model (7) is a model-robust M-estimator for Δ^*, that is, consistent and asymptotically normal under arbitrary misspecification of model (7), under the regularity conditions for M-estimators in the supplementary material; this holds under simple, stratified or biased-coin randomization. The MMRM working model assumptions in the previous paragraph, which include for example, that $E^*[\boldsymbol{Y}|A, \boldsymbol{X}]$ is linear in $(A, \boldsymbol{X})$ and $\boldsymbol{\epsilon} = \boldsymbol{Y} - E^*[\boldsymbol{Y}|A, \boldsymbol{X}]$ is independent of $(A, \boldsymbol{X})$, are not needed for this model-robustness property to hold. We give the corresponding estimating equations, influence function, and proofs in the supplementary material.

An alternative and commonly used estimator for Δ^* is the restricted maximum likelihood (REML) estimator. Under simple randomization and a correctly specified model (7), the ML estimator and REML estimator are asymptotically equivalent on the $n^{1/2}$–scale (Das 1979; Jiang 2017, p.17). To the best of our knowledge, it remains an open question whether this holds under model misspecification and/or other randomization procedures (such as stratified or biased-coin randomization).

If we consider a larger working model than (7) where we allow Σ to depend on the baseline variables, then it is an open question whether the ML estimator $\widehat{\beta}_{AK}$ is model-robust. Under a correctly specified MMRM model (7), Cnaan, Laird, and Slasor (1997) stated that the MMRM estimator is consistent and asymptotically normal under the missing at random assumption when simple randomization is used. Whether this claim extends to stratified or biased-coin randomization is an open question.

6. Estimators Involving Time-to-Event Outcomes

6.1. Notation and Assumptions

For time to event outcomes, we use slightly modified notation and assumptions compared to above. We assume that the outcome is right-censored. Let Y_i denote the failure time and M_i denote the censoring time. Other variables including $A_i, \boldsymbol{X}_i$ and the potential outcomes $Y_i(a), M_i(a)$ for $a = 0, 1$ are defined analogously as in Section 3. For each participant $i \in \{1, \ldots, n\}$, we observe $(A_i, \boldsymbol{X}_i, U_i, \delta_i)$, where $U_i = \min\{Y_i, M_i\}$ and $\delta_i = I\{Y_i \leq M_i\}$. We further define a restriction time τ such that the time window $t \in [0, \tau]$ is of interest. We define P^* and E^* analogously as in Section 3.2, except here they represent the distribution and expectation, respectively, for a single observed data vector $(A, \boldsymbol{X}, U, \delta)$.

The following assumption is made in place of Assumption 1: Assumption 1'.

(i) $\boldsymbol{W}_i, i = 1, \ldots, n$ are independent, identically distributed samples from an unknown joint distribution P on $\boldsymbol{W} = (Y(1), Y(0), M(1), M(0), \boldsymbol{X})$.
(ii) Censoring completely at random: $M(a) \perp\!\!\!\perp Y(a)$ for each arm $a \in \{0, 1\}$.
(iii) $P(\min\{Y(a), M(a)\} > \tau) > 0$ for each $a = 0, 1$.

Compared with Assumption 1, Assumption 1'(i) is the same as Assumption 1(i), and Assumption 1'(ii) assumes censoring completely at random instead of missing at random. This modification of the assumption on missing data is because we consider the K-M estimator and its consistency generally requires Assumption 1'(ii). Assumption 1'(iii) is often made in survival analysis, which states that there is a positive probability that both the failure time and censoring time occur after τ (under each study arm assignment).

6.2. Kaplan–Meier Estimator Under Simple, Stratified, and Biased-Coin Randomization

One commonly used method for survival analysis is the K-M estimator. The goal is to estimate the survival curve $\{S_0^{(a)}(t) : t \in [0, \tau]\}$ for each $a = 0, 1$, where $S_0^{(a)}(t) = P(Y(a) > t)$. This represents the survival curve if everyone in the study population were assigned to study arm a. The K-M estimator is defined as follows:

$$\widehat{S}_n^{(a)}(t) = \prod_{j: T_j \leq t} \left(1 - \frac{\sum_{i=1}^n \delta_i I\{A_i = a\} I\{U_i = T_j\}}{\sum_{i=1}^n I\{A_i = a\} I\{U_i \geq T_j\}}\right),$$

where $\{T_j, j = 1, \ldots, m_n\}$ is the list of unique observed failure times.

While the K-M estimator does not adjust for any baseline variable, its variance under simple randomization is typically different than under stratified or biased-coin randomization, and this is not accounted for by standard methods for estimating its variance. Specifically, the standard method for variance estimation will typically overestimate the K-M variance under stratified or biased-coin randomization, leading to wasted power.

Our variance estimator below avoids this problem. Since the K-M estimator estimates a survival function rather than a real number or a vector, our Theorem 1 on M-estimators does not apply. The following theorem gives the asymptotic distribution of the K-M estimator under our three different types of randomization. It involves the influence function $\mathrm{IF}^{(a)}(A_i, U_i, \delta_i; t)$ for the K-M estimator under simple randomization (Kosorok 2008, Section 4.2), which is also given in the supplementary material.

Theorem 2. Given Assumption 1', under simple, stratified, or biased-coin randomization, we have for each $t \in [0, \tau]$ that

$$\sqrt{n}(\widehat{S}_n^{(a)}(t) - S_0^{(a)}(t)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF^{(a)}(A_i, U_i, \delta_i; t) + o_{p^*}(1), \quad (8)$$

where $o_{p^*}(1)$ represents a sequence of random variables converging to 0 in probability uniformly over $t \in [0, \tau]$.

For stratified and biased-coin randomization, the process $\{\sqrt{n}(\widehat{S}_n^{(a)}(t) - S_0^{(a)}(t)) : t \in [0, \tau]\}$ converges weakly to a mean 0, tight Gaussian process with covariance function $V^{(a)}(t, t')$ defined in the supplementary material, which has the following property: for any $t \le \tau$,

$$V^{(a)}(t,t) = \widetilde{V}^{(a)}(t,t) - \frac{1}{\pi(1-\pi)} E^* \left[E^* \left\{ (A - \pi) IF^{(a)}(A, U, \delta; t) | S \right\}^2 \right], \quad (9)$$

where $\widetilde{V}^{(a)}(t, t)$ is the asymptotic variance under simple randomization. $V^{(a)}(t, t)$ can be consistently estimated as described in the supplementary material.

Analogous to Theorem 1, Theorem 2 implies that the influence function of the K-M estimator is the same under simple, stratified, and biased-coin randomization. The above theorem implies that under stratified or biased-coin randomization, the K-M estimator is consistent and asymptotically normally distributed with equal or smaller asymptotic variance than under simple randomization. The asymptotic covariance function of the K-M estimator under stratified or biased-coin randomization is given in Appendix C of the supplementary material. It can be used to construct pointwise CIs and a simultaneous confidence band.

The challenge in proving Theorem 2 is that the traditional tool for deriving asymptotic normality in survival analysis, that is, martingale central limit theorems such as (Andersen et al. 2012, theor. II.5.1) or (Fleming and Harrington 2011, theor. 5.1.1), is not applicable here because of the dependence among data points introduced by stratified or biased-coin randomization. To overcome the above difficulty, in the proof of Theorem 2, we first developed a central limit theorem for sums of random functions under stratified randomization (Lemma 5 in the supplementary material) based on the empirical process results of Shorack and Wellner (2009) combined with generalizations of the techniques from Bugni, Canay, and Shaikh (2018). We then proved Theorem 2 by generalizing the arguments in our proof of Theorem 1 to handle random functions. We conjecture that, using our central limit theorem, Theorem 2 can be generalized to apply to other estimators of survival functions, such as the covariate-adjusted estimators proposed by Lu and Tsiatis (2011) and Zhang (2015), which may improve precision even further.

6.3. Other Estimators for Time-to-Event Outcomes

Another parameter of interest is the restricted mean survival time, defined as $\Delta^* = E[\min\{Y(1), \tau\} - \min\{Y(0), \tau\}]$. One covariate adjusted estimator of the restricted mean survival time is the AIPW estimator of Moore and van der Laan (2009b). This estimator is an M-estimator, to which our Theorem 1 applies. When the survival probability at a given time point is the parameter of interest, one can use the K-M estimator or the method from Moore and van der Laan (2009b).

7. Clinical Trial Applications

7.1. Binary and Continuous Outcomes

Table 1 summarizes our data analyses involving binary and continuous outcomes. The outcome is binary for NIDA-CTN-0003 and is continuous for NIDA-CTN-0030 and NIDA-CTN-0044. In all cases, the target of inference is the average treatment effect defined as $E[Y(1)] - E[Y(0)]$.

All missing baseline values were imputed by the median for continuous variables and mode for binary or categorical variables. The only estimator in Table 1 that adjusts for missing outcomes is the DR-WLS estimator; all other estimators omit data from the participants with missing outcomes. Negative (positive) estimates are in the direction of clinical benefit (harm). For all estimators presented in Table 1, the 95% CI is constructed using the normal approximation with variance calculated from formula (4).

For NIDA-CTN-0003, the outcome is binary and "adjusted estimator" in Table 1 refers to the standardized logistic regression estimator. The unadjusted point estimate is -0.104 with 95% CI $(-0.204, -0.004)$. If randomization strata and additional baseline variables are adjusted for (as in the row "Adjusted estimator ($\boldsymbol{X}$)" in Table 1), the point estimate is unchanged but the 95% CI $(-0.184, -0.024)$ is substantially smaller. The corresponding variance reduction due to covariate adjustment, defined as one minus the variance ratio of "Adjusted estimator ($\boldsymbol{X}$)" to the unadjusted estimator, is 36%. This is equivalent to needing 36% fewer participants to achieve the same power as a trial that uses the unadjusted estimator, asymptotically.

NIDA-CTN-0030 and NIDA-CTN-0044 had continuous-valued outcomes and "Adjusted estimator" in Table 1 refers to

Table 1. Summary of clinical trial data analyses with each cell giving the point estimate and 95% CI of an estimator.

	Clinical Trial:		
	NIDA-CTN-0003	NIDA-CTN-0030	NIDA-CTN-0044
Unadjusted estimator	−0.104(−0.204, −0.004)	0.015(−0.023, 0.052)	−0.093(−0.149, −0.038)
Adjusted estimator (S)	−0.110(−0.209, −0.009)	0.015(−0.022, 0.052)	−0.089(−0.145, −0.033)
Adjusted estimator ($\boldsymbol{X}$)	−0.104(−0.184, −0.024)	0.012(−0.022, 0.046)	−0.087(−0.142, −0.032)
DR-WLS estimator ($\boldsymbol{X}$)	−0.099(−0.180, −0.019)	0.012(−0.022, 0.045)	−0.091(−0.148, −0.035)

NOTES: Each row is for a different estimator. "Adjusted estimator" refers to the standardized logistic estimator for the trial with binary outcome (Column 2) and to the ANCOVA estimator for the trials with continuous outcomes (Columns 3 and 4). The variable in parentheses after the estimator name indicates which variables (if any) are adjusted for, with S denoting the randomization strata only and $\boldsymbol{X}$ denoting the randomization strata and additional baseline variables.

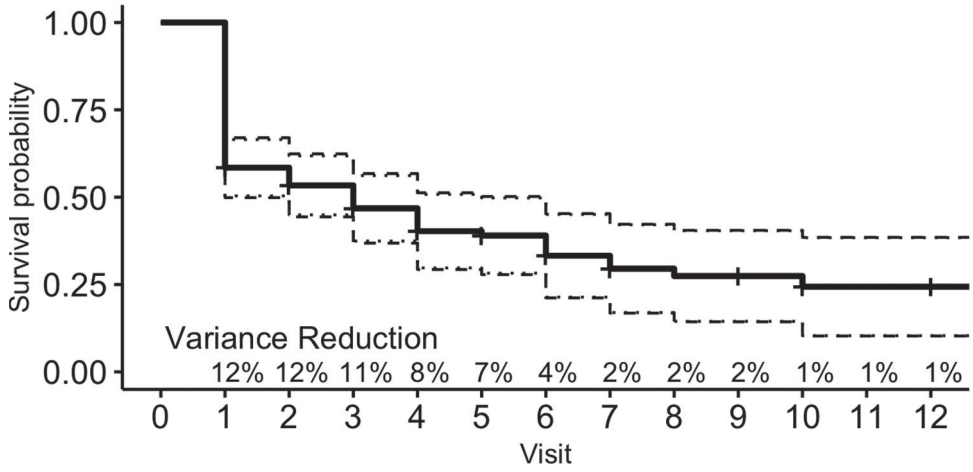

Figure 1. The K-M estimator of survival function for NIDA-CTN-0044 treatment group. The solid line is the estimated survival function. Dashed and dotted lines, respectively, represent CIs using the standard method and CIs accounting for randomization strata using (9); the dashed and dotted lines are very similar and almost coincide. "Variance Reduction" and the associated percentages represent the variance reduction due to accounting for stratified randomization using (9).

the ANCOVA estimator. Covariate adjustment brings 17% and 3% variance reduction for NIDA-CTN-0030 and NIDA-CTN-0044, respectively, compared to the unadjusted estimator. In all cases, the variance reduction from covariate adjustment is larger when the baseline variables are more strongly prognostic for (i.e., more strongly correlated with) the outcome.

In all three trials, the variance reduction due to adjusting for baseline variables beyond S, defined by one minus the variance ratio of "adjusted estimator ($\boldsymbol{X}$)" and "adjusted estimator (S)," is the same (to the nearest percent) as the corresponding variance reduction comparing "adjusted estimator ($\boldsymbol{X}$)" to the unadjusted estimator. This is expected for the ANCOVA estimator since Bugni, Canay, and Shaikh (2018) showed that "adjusted estimator (S)" and the unadjusted estimator are asymptotically equivalent when the randomization probability $\pi = 0.5$. Also, in all three trials, the "DR-WLS estimator ($\boldsymbol{X}$)," which handles missing outcomes under the missing at random assumption, has a similar point estimate and 95% CI compared to "adjusted estimator ($\boldsymbol{X}$)," which omits missing outcomes. We recommend using the DR-WLS estimator when outcomes have missing values and the missing at random assumption is plausible.

We next compare the estimated variance (and resulting CIs) based on the sandwich variance estimator versus the variance estimator (4). For the unadjusted estimator, using the sandwich variance estimator instead of formula (4) may lead to conservative variance estimates, as implied by Theorem 1. For example, for NIDA-CTN-0044, the 95% CI of the unadjusted estimator constructed by formula (4) is $(-0.149, -0.038)$, while the 95% CI calculated using the sandwich variance formula is $(-0.162, -0.025)$, which is 23% wider. The former 95% CI is asymptotically correct assuming outcomes are missing completely at random, an assumption that is generally needed for the unadjusted estimator to be consistent. Furthermore, the variance of the unadjusted estimator calculated by formula (4) (which is consistent) is 34% smaller than the variance calculated by the sandwich variance estimator (which is conservative). In contrast, for the adjusted estimator or the DR-WLS estimator, since all three trials have randomization probability $\pi = 0.5$, the sandwich variance estimator is not conservative; this follows from Corollary 1.

7.2. Time-to-Event Outcome

Figure 1 presents the K-M estimator for time-to-abstinence in the treatment group as defined in Section 2.3 for study NIDA-CTN-0044. We estimated the variance of the K-M estimator in two different ways: one ignored the stratification variable and was the estimated variance returned by the "survfit" function in R; the other used our proposed variance formula that takes the stratification into account. For each of the two variance estimators, we constructed corresponding point-wise CIs for the K-M estimator.

While Figure 1 shows that CIs based on different variance estimators are very close to each other, there are variance reductions due to accounting for stratification, which can be translated into sample size reduction needed to achieve the desired power. The variance reduction ranges from 1% to 12% as we consider the survival function at different time points. Among all time points, the first time point (one week after randomization) has the greatest variance reduction. The variance formula (9) from Theorem 2 accounts for the improved precision due to stratified randomization (unlike standard methods that ignore stratification variables); this can be used to construct more powerful hypothesis tests based on the K-M estimator divided by its standard error. The corresponding figure and results for the control group are given in the supplementary material and are qualitatively similar to those described above for the treatment group.

8. Discussion

The primary efficacy analysis in confirmatory randomized trials is typically based on a treatment effect estimator that is asymptotically linear under simple randomization; that is, for an appropriately chosen influence function IF, the estimator has the form (2) when estimating a scalar/vector or (8) when estimating a function such as a survival curve. All estimators in

this article have this property, and we proved for each estimator covered by Theorems 1 and 2 that under stratified and biased-coin randomization, it is asymptotically linear with the same influence function as under simple randomization. We then gave formulas (3) and (9) for the asymptotic variance under stratified (and biased-coin) randomization in terms of the influence function.

Though our theorems cover a variety of estimators used to analyze randomized trials, they do not handle every estimator. However, our results point to a general approach for deriving the asymptotic behavior under stratified and biased-coin randomization of any estimator that is known to be asymptotically linear under simple randomization. The approach is to (a) conjecture that under stratified and biased-coin randomization it is asymptotically linear with the same influence function as under simple randomization; (b) prove this, which may need to be tailored to the estimator, for example, using techniques as shown in the supplementary material for M-estimators and the K-M estimator; (c) apply results from the supplementary material (Proposition 1 or Lemma 5) to show that the asymptotic variance is given by (3) for scalar/vector parameters or by (9) for functions. An area of future research is to apply this approach to the estimators of Lu and Tsiatis (2011) and Zhang (2015) that use covariate adjustment to improve precision of the K-M estimator.

Our asymptotic results, just as many asymptotic results under the commonly used superpopulation inference framework for randomized trials, assume that the number of randomization strata is fixed and the number of participants in each stratum goes to infinity. This may be a reasonable approximation when no stratum has a small number of participants. In our data examples, the smallest stratum has 49 participants. An area of future research is to consider cases where some randomization strata have few participants.

In our data analyses of NIDA-CTN-0044, the stratification variable "site" was not available in our dataset. It was therefore neither used in the estimators nor in the corresponding variance estimates. The variance formulas (3) and (9) in this case are asymptotically conservative. This is because the outer expectation in the rightmost terms of these formulas are unchanged or decreased if S is replaced by a coarsening of S (defined as merging several randomization strata together in a preplanned way, in the analysis); this follows from the conditional Jensen's inequality. This result may be useful more generally, for example, when some strata are so small compared to the sample size that stratum-specific evaluation of the empirical means E_n in (4) and the corresponding estimator for (9) cannot be reliably done. In such cases a pre-planned, coarsened stratum indicator could be used and the resulting hypothesis test would still control Type I error, asymptotically.

Stratified randomization is related to stratified sampling designs, also called "two-phase sampling" (Sen 1988; Breslow and Wellner 2007; Bai, Tsiatis, and O'Brien 2013). To the best of our knowledge, asymptotic results for these designs do not directly apply to our problem; a key difference is that asymptotic results for stratified sampling designs often involve finite population inference (commonly used in survey sampling), while here we use superpopulation inference (commonly used in analyzing randomized trials).

We provide R functions to calculate the variance for estimators including those in Examples 1 and 2 and the K-M estimator which are available on Github at *https://github.com/BingkaiWang/covariate-adaptive*

Supplementary Materials

The Supplementary Material contains the regularity conditions for Theorem 1, consistent estimators for the asymptotic variances in Theorems 1 and 2, proofs of all results, and additional data analysis for the K-M estimator.

Funding

This project was supported by a research award from Arnold Ventures. The information reported here results from secondary analyses of data from clinical trials conducted by the National Institute on Drug Abuse (NIDA). Specifically, data from NIDA–CTN-0003 (Suboxone (Buprenorphine/Naloxone) Taper: A Comparison of Two Schedules), NIDA-CTN-0030 (A Two-Phase Randomized Controlled Clinical Trial of Buprenorphine/Naloxone Treatment Plus Individual Drug Counseling for Opioid Analgesic Dependence) and NIDA-CTN-0044 (Web-delivery of Evidence-Based, Psychosocial Treatment for Substance Use Disorders) were included. NIDA databases and information are available at (*https://datashare.nida.nih.gov*). MR was supported by the Johns Hopkins Center of Excellence in Regulatory Science and Innovation, which is funded by the Food and Drug Administration (FDA) of the U.S. Department of Health and Human Services (HHS) as part of a financial assistance award (U01FD005942). The contents are those of the author(s) and do not necessarily represent the official views of, nor an endorsement, by any of the aforementioned organizations, the FDA/HHS, nor the U.S. Government.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (2012), *Statistical Models Based on Counting Processes*, Springer Science & Business Media. [8]

Austin, P., Manca, A., Zwarenstein, M., Juurlink, D., and Stanbrook, M. (2010), "A Substantial and Confusing Variation Exists in Handling of Baseline Covariates in Randomized Controlled Trials: A Review of Trials Published in Leading Medical Journals," *Journal of Clinical Epidemiology*, 63, 142 – 153. [2]

Bai, X., Tsiatis, A. A., and O'Brien, S. M. (2013), "Doubly-Robust Estimators of Treatment-Specific Survival Distributions in Observational Studies With Stratified Sampling," *Biometrics*, 69, 830–839. [10]

Breslow, N. E., and Wellner, J. A. (2007), "Weighted Likelihood for Semiparametric Models and Two-Phase Stratified Samples, With Application to Cox Regression," *Scandinavian Journal of Statistics*, 34, 86–102. [10]

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018), "Inference Under Covariate-Adaptive Randomization," *Journal of the American Statistical Association*, 113, 1784–1796. [2,4,5,8,9]

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2019), "Inference Under Covariate-Adaptive Randomization With Multiple Treatments," *Quantitative Economics*, 10, 1747–1785. [2]

Campbell, A. N., Nunes, E. V., Matthews, A. G., Stitzer, M., Miele, G. M., Polsky, D., Turrigiano, E., Walters, S., McClure, E. A., Kyle, T. L., Wahle, A., Van Veldhuisen, P., Goldman, B., Babcock, D., Stabile, P. Q., Winhusen, T., and Ghitza, U. E. (2014), "Internet-Delivered Treatment for Substance Abuse: A Multisite Randomized Controlled Trial," *American Journal of Psychiatry*, 171, 683–690. [3]

Cnaan, A., Laird, N. M., and Slasor, P. (1997), "Using the General Linear Mixed Model to Analyse Unbalanced Repeated Measures and Longitudinal Data," *Statistics in Medicine*, 16, 2349–2380. [7]

Das, K. (1979), "Asymptotic Optimality of Restricted Maximum Likelihood Estimates for the Mixed Model," *Calcutta Statistical Association Bulletin*, 28, 125–142. [7]

Efron, B. (1971), "Forcing a Sequential Experiment to be Balanced," *Biometrika*, 58, 403–417. [1]

EMA (2015), "European Medicines Agency Guideline on Adjustment for Baseline Covariates in Clinical Trials," *European Medicines Agency: CPMP/295050/2013*. [1,2]

FDA (2020), "COVID-19: Developing Drugs and Biological Products for Treatment or Prevention. Guidance for Industry." Available at: *https://www.fda.gov/media/137926/download*. [1,2,6]

FDA. (2021), "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products. Draft Guidance for Industry." Available at: *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products*. [1,2]

FDA and EMA. (1998), "E9 Statistical Principles for Clinical Trials," U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96. [1,2]

Fleming, T. R., and Harrington, D. P. (2011), *Counting Processes and Survival Analysis*, Volume 169. New York: Wiley. [8]

Jiang, F., Tian, L., Fu, H., Hasegawa, T., and Wei, L. J. (2018), "Robust Alternatives to ANCOVA for Estimating the Treatment Effect Via a Randomized Comparative Study," *Journal of the American Statistical Association*, 114, 1854–1864. [1]

Jiang, J. (2017). *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*. Boca Raton: CRC Press. [7]

Kahan, B. C., and Morris, T. P. (2012), "Improper Analysis of Trials Randomised Using Stratified Blocks or Minimisation," *Statistics in Medicine*, 31, 328–340. [1]

Kaplan, E. L., and Meier, P. (1958), "Nonparametric Estimation From Incomplete Observations," *Journal of the American Statistical Association*, 53, 457–481. [2]

Koch, G. G., Tangen, C. M., Jung, J.-W., and Amara, I. A. (1998), "Issues for Covariance Analysis of Dichotomous and Ordered Categorical Data From Randomized Clinical Trials and Non-Parametric Strategies for Addressing Them," *Statistics in Medicine*, 17, 1863–1892. [1]

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer. [8]

Lachin, J., Matts, J., and Wei, L. (1988), "Randomization in Clinical Trials: Conclusions and Recommendations," *Controlled Clinical Trials*, 9, 365 – 374. [1]

Lane, P. (2008), "Handling Drop-Out in Longitudinal Clinical Trials: A Comparison of the LOCF and MMRM Approaches," *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 7, 93–106. [7]

Li, X., and Ding, P. (2020), "Rerandomization and Regression Adjustment," *Journal of the Royal Statistical Society*, Series B, 82, 241–268. [2]

Lin, Y., Zhu, M., and Su, Z. (2015), "The Pursuit of Balance: An Overview of Covariate-Adaptive Randomization Techniques in Clinical Trials," *Contemporary Clinical Trials*, 45, 21 – 25. 10th Anniversary Special Issue. [1]

Ling, W., Hillhouse, M., Domier, C., Doraimani, G., Hunter, J., Thomas, C., Jenkins, J., Hasson, A., Annon, J., Saxon, A., Selzer, J., Boverman, J., and Bilangi, R. (2009), "Buprenorphine Tapering Schedule and Illicit Opioid Use," *Addiction*, 104, 256–265. [2]

Liu, H., and Yang, Y. (2020), "Regression-Adjusted Average Treatment Effect Estimates in Stratified Randomized Experiments," *Biometrika*, 107, 935–948. [2]

Lu, X., and Tsiatis, A. A. (2011), "Semiparametric Estimation of Treatment Effect With Time-Lagged Response in the Presence of Informative Censoring," *Lifetime Data Analysis*, 17, 566–593. [8,10]

Ma, W., Hu, F., and Zhang, L. (2015), "Testing Hypotheses of Covariate-Adaptive Randomized Clinical Trials," *Journal of the American Statistical Association*, 110, 669–680. [2]

Ma, W., Qin, Y., Li, Y., and Hu, F. (2018), "Statistical Inference of Covariate-Adjusted Randomized Experiments," available at: *https://arxiv.org/abs/1807.09678*. [2]

Mallinckrod, C. H., Lane, P. W., Schnell, D., Peng, Y., and Mancuso, J. P. (2008), "Recommendations for the Primary Analysis of Continuous Endpoints in Longitudinal Clinical Trials," *Drug Information Journal*, 42, 303–319. [6,7]

Moore, K., and van der Laan, M. (2009a), "Covariate Adjustment in Randomized Trials With Binary Outcomes: Targeted Maximum Likelihood Estimation," *Statistics in Medicine*, 28, 39–64. [1]

——— (2009b), "Increasing Power in Randomized Trials With Right Censored Outcomes Through Covariate Adjustment," *Journal of Biopharmaceutical Statistics*, 19, 1099–1131. PMID: 20183467. [1,8]

Morgan, K. L., and Rubin, D. B. (2012), "Rerandomization to Improve Covariate Balance in Experiments," *The Annals of Statistics*, 40, 1263–1282. [1]

Neyman, J. S., Dabrowska, D. M., and Speed, T. (1990), "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," *Statistical Science*, 5, 465–472. [3]

Pocock, S. J., and Simon, R. (1975), "Sequential Treatment Assignment With Balancing for Prognostic Factors in the Controlled Clinical Trial," *Biometrics*, 31, 103–115. [1]

Robins, J. M. (2002), "Covariance Adjustment in Randomized Experiments and Observational Studies: Comment," *Statistical Science*, 17, 309–321. [2]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [5]

Robins, J. M., Sued, M., Lei-Gomez, Q., and Rotnitzky, A. (2007), "Comment: Performance of Double-Robust Estimators When 'Inverse Probability' Weights Are Highly Variable," *Statistical Science*, 22, 544–559. [6]

Rubin, D. and van der Laan, M. (2008), "Covariate Adjustment for the Intention-To-Treat Parameter With Empirical Efficiency Maximization," *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 229. Available at: *https://biostats.bepress.com/ucbbiostat/paper229*. [1]

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999), "Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models," *Journal of the American Statistical Association*, 94, 1096–1120. [5,6]

Sen, P. K. (1988), "Asymptotics in Finite Population Sampling," in *Sampling*, ed. C. R. Rao, Volume 6 of *Handbook of Statistics*, pp. 291–331. Amsterdam, Netherlands: Elsevier. [10]

Shao, J., and Yu, X. (2013), "Validity of Tests Under Covariate-Adaptive Biased Coin Randomization and Generalized Linear Models," *Biometrics*, 69, 960–969. [2]

Shao, J., Yu, X., and Zhong, B. (2010), "A Theory for Testing Hypotheses Under Covariate-Adaptive Randomization," *Biometrika*, 97, 347–360. [2]

Shorack, G. R., and Wellner, J. A. (2009), *Empirical Processes With Applications to Statistics*. Philadelphia: Society for Industrial and Applied Mathematics. [8]

Tsiatis, A. (2007), *Semiparametric Theory and Missing Data*, New York: Springer Science & Business Media. [5]

Tsiatis, A., Davidian, M., Zhang, M., and Lu, X. (2008), "Covariate Adjustment for Two-Sample Treatment Comparisons in Randomized Clinical Trials: A Principled Yet Flexible Approach," *Statistics in Medicine*, 27, 4658–4677. [1]

van der Laan, M. J., and Gruber, S. (2012), "Targeted Minimum Loss Based Estimation of Causal Effects of Multiple Time Point Interventions," *The International Journal of Biostatistics*, 8, Article 9. [5]

van der Vaart, A. (1998), *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press. [4,5]

Wei, L. J. (1978), "The Adaptive Biased Coin Design for Sequential Experiments," *The Annals of Statistics*, 6, 92–100. [1]

Weiss, R. D., Potter, J. S., Fiellin, D. A., Byrne, M., Connery, H. S., Dickinson, W., Gardin, J., Griffin, M. L., Gourevitch, M. N., Haller, D. L., Hasson, A. L., Huang, Z., Jacobs, P., Kosinski, A. S., Lindblad, R., McCance-Katz, E. F., Provost, S. E., Selzer, J., Somoza, E. C., Sonne, S. C., and Ling, W. (2011), "Adjunctive Counseling During Brief and Extended Buprenorphine-Naloxone Treatment for Prescription Opioid Dependence: A 2-Phase Randomized Controlled Trial," *JAMA Psychiatry*, 68, 1238–1246. [3]

Yang, L., Ma, W., Qin, Y., and Hu, F. (2020), "Testing for Treatment Effect in Covariate-Adaptive Randomized Clinical Trials With Generalized Linear Models and Omitted Covariates," available at: *https://arxiv.org/abs/2009.04136*. [2]

Yang, L., and Tsiatis, A. (2001), "Efficiency Study of Estimators for a Treatment Effect in a Pretest-Posttest Trial," *The American Statistician*, 55, 314–321. [1,2]

Ye, T., and Shao, J. (2020), "Robust Tests for Treatment Effect in Survival Analysis Under Covariate-Adaptive Randomization," *Journal of the Royal Statistical Society*, Series B, 82, 1301–1323. [2]

Ye, T., Shao, J., and Zhao, Q. (2020a), "Principles for Covariate Adjustment in Analyzing Randomized Clinical Trials," available at: *https://arXiv:2009.11828*. [2]

Ye, T., Yi, Y., and Shao, J. (2020b), "Inference on Average Treatment Effect Under Minimization and Other Covariate-Adaptive Randomization Methods," available at: *https://arxiv.org/abs/2007.09576*. [2]

Zelen, M. (1974), "The Randomization and Stratification of Patients to Clinical Trials," *Journal of Chronic Diseases*, 27, 365 – 375. [1]

Zhang, M. (2015), "Robust Methods to Improve Efficiency and Reduce Bias in Estimating Survival Curves in Randomized Clinical Trials," *Lifetime Data Analysis*, 21, 119–137. [1,8,10]