Leveraging prognostic baseline variables to gain precision in randomized trials

Elizabeth Colantuoni and **Michael Rosenblum**[*]

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, U.S.A., 21205

Abstract

We focus on estimating the average treatment effect in a randomized trial. If baseline variables are correlated with the outcome, then appropriately adjusting for these variables can improve precision. An example is the analysis of covariance (ANCOVA) estimator, which applies when the outcome is continuous, the quantity of interest is the difference in mean outcomes comparing treatment versus control, and a linear model with only main effects is used. ANCOVA is guaranteed to be at least as precise as the standard unadjusted estimator, asymptotically, under no parametric model assumptions and also is locally semiparametric efficient. Recently, several estimators have been developed that extend these desirable properties to more general settings that allow any real-valued outcome (e.g., binary or count), contrasts other than the difference in mean outcomes (such as the relative risk), and estimators based on a large class of generalized linear models (including logistic regression). To the best of our knowledge, we give the first simulation study in the context of randomized trials that compares these estimators. Furthermore, our simulations are not based on parametric models; instead, our simulations are based on resampling data from completed randomized trials in stroke and HIV in order to assess estimator performance in realistic scenarios. We provide practical guidance on when these estimators are likely to provide substantial precision gains and describe a quick assessment method that allows clinical investigators to determine whether these estimators could be useful in their specific trial contexts.

Keywords

prognostic variables; randomized trial; relative efficiency

---

1. Introduction

We focus on estimating the average treatment effect in a randomized trial that compares an experimental treatment versus control. The average treatment effect can be defined, for example, as a difference between population means, a relative risk (for binary outcomes), or any smooth contrast between the population means under treatment and control.

The unadjusted estimator of the average treatment effect is constructed using only the sample means in the treatment group and control group. Advantages of the unadjusted estimator include that it is simple to describe and implement, and it is unbiased. However, it ignores information in baseline variables that can improve the precision of treatment effect estimates.

There has been much debate over the proper use of estimators that adjust for baseline variables in randomized trials [1,2]. If baseline variables are strongly correlated with the outcome, then estimators that adjust for these variables can increase precision compared with the unadjusted estimator, as described in the succeeding paragraphs. This increased precision can translate into a reduction in the required sample size to achieve a desired power.

Limitations of some adjusted estimators are the following: they can be inconsistent if certain parametric model assumptions do not hold, they can have lower asymptotic precision than the unadjusted estimator, they can be challenging to implement, and they can require solving difficult optimization problems. The analysis of covariance estimator has none of these limitations (as shown by Yang and Tsiatis [3]) but is limited to cases where the outcome is continuous, the quantity of interest is the difference between mean outcomes comparing treatment versus control, and a linear model with only main effects is used.

Recently, estimators have been developed that extend the desirable properties of analysis of covariance, described earlier, to much more general settings [4–6]. These estimators have been studied in terms of theoretical properties, and each estimator has been separately studied in its own finite sample simulation. (An exception is the PLEASE estimator defined in Section 4.2, which had not yet been studied in any finite sample context.) However, to the best of our knowledge, these recent estimators have not been compared against one another in any simulation study. We conduct simulation studies comparing these estimators based on resampling data from two completed randomized trials.

In the first set of simulations, the efficiency gained from adjusting for prognostic baseline variables is equivalent to a 22–30% reduction in the required sample size to achieve a desired power. In the second example, the baseline variables are weakly correlated with the outcome, resulting in a very small efficiency gain. We also simulated scenarios where baseline variables are independent of the outcome to determine if adjusting for uninformative variables causes a loss in efficiency compared with the unadjusted estimator. The observed efficiency losses are quite small.

We describe a quick assessment method and simulation approach that an investigator planning a phase III trial could apply to determine if an adjusted estimator is likely to

provide substantial precision gains. This method can be used as a practical guide to help inform what estimator to prespecify in a trial protocol. We provide R and sas code (SAS Institute, Cary, NC, USA) for one of the adjusted estimators that is a practical compromise between computational complexity and statistical efficiency.

In the next section, we describe the randomized trial data on which our simulations are based. In Sections 3 and 4, we present estimators of the risk difference (for binary outcomes) that leverage baseline variable information. In Sections 5 and 6, we compare the performance of these estimators in simulations. A modified version of the quick assessment method of Rubin and van der Laan [7] and Moore and van der Laan [8] is applied to these estimators in Section 7. Generalizations to estimate the relative risk and log odds ratio, and to handle non-binary outcomes, are given in Section 8. Practical issues related to implementing these estimators and areas for future research are discussed in Section 9.

2. Data examples

2.1. MISTIE II trial

MISTIE II is a phase II, multicenter, randomized, prospective trial completed in 2013. Participants were randomized to the treatment arm (surgical) or control arm (standard medical care). We define the treatment arm as those assigned to one of the following surgical treatments: MISTIE (minimally invasive surgery and clot lysis with rt-PA to remove intracerebral hemorrhage (ICH)), or ICES (intraoperative stereotactic CT-guided endoscopic surgery) [9]. The primary outcome is the participant's score on the modified Rankin Scale (mRS), which measures functional disability. A mRS score of 3 or less is defined to be a successful outcome. The following baseline variables are strongly associated with the primary outcome: age, ICH volume, and National Institutes of Health Stroke Scale (NIHSS). The strength of the association was determined by calculating a modified version of R^2 in the ordinary least squares sense defined in Section 7, which represents the proportion of the outcome variance accounted for by treatment and baseline variables (age, ICH volume, and NIHSS) beyond that accounted for by treatment alone. The modified R^2 was 22% (Section 7).

The average treatment effect of interest is the risk difference, that is, the difference between the population proportion of successes under assignment to treatment versus control. The unadjusted estimator is the difference between the observed proportion of successes in treatment versus control. Its value is 12.0% (95% CI: −5.9 to 30.2%) comparing the 66 MISTIE II participants to the 37 standard medical care participants. (The randomization ratio was 2 : 1 treatment to control.) The estimator we describe in Section 4.2, which adjusts for the aforementioned baseline variables, results in the estimate and confidence interval: 14.4% (95% CI: 1.3 to 32.8%). The width of this confidence interval is 12.7% smaller than that of the unadjusted estimator. This gives an initial indication that there may be gains in precision from appropriately leveraging prognostic information in baseline variables. We show this holds in simulation studies in Section 5.

2.2. PEARLS trial

The Prospective Evaluation of Anti-retroviral Combinations for Treatment Naive, HIV Infected Persons in Resource-limited Settings (PEARLS) trial was completed in 2010. It is a phase IV, randomized, open-label multinational clinical trial in HIV-1 infected, antiretroviral-naive participants comparing three different drug combinations [10]. We considered the 1044 participants assigned to either of the first two therapies (called A and B). Poor response to therapy is defined as at least one of the following occurring during follow-up: virologic failure (two consecutive plasma HIV-1 RNAs > 1000 copies/mL at week 16 onwards), HIV-1 disease progression (AIDS), or death. Baseline variables that are weakly associated (modified R^2 of 3.6%) with the primary outcome include gender, plasma viral load, and CD4+ cell count. The unadjusted and adjusted estimators of the risk difference of poor response comparing treatment A with B are −5.9% (95% CI: −10.7 to −1.4%) and −6.0% (95% CI: −10.5 to −1.3%), respectively, using the aforementioned baseline variables. There was essentially no reduction in the confidence interval width from adjusting for baseline variables. This is not surprising given the weak association between the primary outcome and baseline variables within the PEARLS trial. A goal of this paper is to investigate the practical benefits and costs of adjusting for weakly prognostic variables, as well as strongly prognostic variables.

3. Notation, definitions, and desired estimator properties

3.1. Notation

Consider a randomized trial where each participant's data is a vector (W, A, Y), where W is a vector of baseline variables, A is the treatment arm indicator ($A = 1$ for treatment, and $A = 0$ for control), and Y is the primary outcome. We assume the variables included in the vector W are determined before the trial starts. Throughout, we focus on the case of binary Y with no missing outcomes, that is, $Y \in \{0, 1\}$ and is observed for all participants. We provide generalizations for non-binary Y in Section 8 and describe extensions to handle missing outcomes in Section 1 of the Supplementary Material.

We consider 1 : 1 randomization, that is, each participant is randomized to treatment or control with probability 1/2, independent of the participant's baseline variables. However, the estimators we consider can also be applied to data from trials with unequal randomization probabilities. We consider a nonparametric model $\mathcal{M}$ for the data generating distribution P for (W, A, Y), which includes the following assumptions: $P \in \mathcal{M}$ is dominated by a common, σ-finite measure, and A is independent of W because of randomization. Although A and W are independent in the data generating distribution P, there may be chance imbalances between study arms in any given realization of study data $\{(W_i, A_i, Y_i)\}_{i=1}^{n}$.

The aim is to estimate the population average treatment effect

$$\psi = E(Y|A=1) - E(Y|A=0) = P(Y=1|A=1) - P(Y=1|A=0), \quad (1)$$

based on n independent, identically distributed realizations $\{(W_i, A_i, Y_i)\}_{i=1}^{n}$ of the random vector (W, A, Y), each drawn from the (unknown) probability distribution P. Although most of the paper focuses on estimating the difference between $P(Y=1|A=1)$ and $P(Y=1|A=0)$, a similar approach can be used to estimate other contrasts between these quantities such as the relative risk $P(Y=1|A=1)/P(Y=1|A=0)$ or the log odds ratio

$$logit\left\{P\left(Y=1|A=1\right)\right\}-logit\left\{P\left(Y=1|A=0\right)\right\}=log\left[\frac{P\left(Y=1|A=1\right)/\left\{1-P\left(Y=1|A=1\right)\right\}}{P\left(Y=1|A=0\right)/\left\{1-P\left(Y=1|A=0\right)\right\}}\right].$$

(2)

where $logit(x) = log\{x/(1-x)\}$. Estimation of these contrasts is described in Section 8.

The primary analysis in a confirmatory randomized trial typically involves estimation of the average (also called marginal) treatment effect, which is our goal. We do not aim to estimate conditional treatment effects such as $P(Y=1|A=1, W) - P(Y=1|A=0, W)$, that is, the risk difference comparing treatment with control within strata of the baseline variables W. An advantage of the average treatment effect is that it be consistently estimated from a randomized trial without any parametric model assumptions; in contrast, the conditional treatment effect generally requires model assumptions when baseline variables are continuous or are discrete with many levels.

To illustrate the difference between conditional and marginal effects, consider an example where the true distribution is $P(Y=1|A, W) = logit^{-1}(1 + A + W)$ for $W \sim N(0,1)$, A having probability 1/2 of being 0 or 1, and A and W are independent by randomization. Then the conditional treatment effect of A given W on the logit scale is

$$logit\left\{P\left(Y=1|A=1, W\right)\right\}-logit\left\{P\left(Y=1|A=0, W\right)\right\}=logit\left\{logit^{-1}\left(1+1+W\right)\right\}-logit\left\{logit^{-1}\left(1+W\right)\right\}=1,$$

that is, a conditional log odds ratio of 1. In contrast, the average (unconditional) treatment effect of A is a log odds ratio of 0.86 (rounded to two decimal places) computed by numerical integration and (2) using

$$P\left(Y=1|A=a\right)=\int P\left(Y=1|A=a, W=w\right)dP_W\left(w\right)=\int logit^{-1}\left(1+a+w\right)dP_W\left(w\right),$$

where $P_W(w)$ is the marginal distribution of W; the first equality in the previous display follows from A and W being independent. Consider the maximum likelihood estimator $\hat{\beta}_1$ of β_1 in the logistic regression model $logit^{-1}(\beta_0 + \beta_1 A + \beta_2 W)$. This estimator converges to 1, i.e., the conditional effect, as sample size goes to infinity; it is therefore not a consistent estimator of the average treatment effect of 0.86. This example involved a distribution that satisfies the assumptions of a logistic regression model. In cases where the logistic regression model is misspecified, $\hat{\beta}_1$ will generally converge to a limit that is neither a

conditional nor an average effect (or it may fail to converge). For more discussion of average versus conditional effects, see, for example, [11,12].

3.2. Definitions

An estimator $\hat{\psi}_n$ of the population parameter ψ is called consistent if $lim_{n \to \infty} \hat{\psi}_n = \psi$ with probability 1, for any data generating distribution $P \in \mathcal{M}$. It is called asymptotically normal if for any $P \in \mathcal{M}$, $\sqrt{n}\left(\hat{\psi}_n - \psi\right)$ converges to a Gaussian distribution. Similar to [5], we assume throughout that the regularity conditions of Theorems 5.9 and 5.21 of [13] hold. These conditions and the randomization assumption from Section 3.1 imply that the estimators we compare are consistent and asymptotically normal.

Consider two estimators $\hat{\psi}_{1,n}, \hat{\psi}_{2,n}$ of ψ at sample size n, with variances $\sigma_{1,n}^2, \sigma_{2,n}^2$, respectively. Their relative efficiency (also called average relative efficiency) is defined as $RE = lim_{n \to \infty} \sigma_{2,n}^2 / \sigma_{1,n}^2$ [13]. If both estimators are consistent and asymptotically normal, then the relative efficiency equals the asymptotic ratio of sample sizes required by each estimator to achieve a desired power (e.g., 80%) at local alternatives; equivalently, the relative sample size reduction from using $\hat{\psi}_{1,n}$ instead of $\hat{\psi}_{2,n}$, to achieve a desired power at local alternatives, converges to $1 - (1/RE)$.

For a given submodel $\mathcal{M}_{sub} \subset \mathcal{M}$, an estimator $\hat{\psi}_n$ is called semiparametric, locally efficient (or just locally efficient) with respect to $\mathcal{M}_{sub}$ if $\hat{\psi}_n$ achieves the semiparametric efficiency bound for $\mathcal{M}$ at each $P \in \mathcal{M}_{sub}$. Intuitively, this means that, when the true distribution P satisfies the model assumptions of $\mathcal{M}_{sub}$, the estimator $\hat{\psi}_n$ has the best possible (asymptotic) precision among all estimators for ψ that are regular and asymptotically linear in the full model $\mathcal{M}$. In our context, the submodel $\mathcal{M}_{sub}$ is defined by a generalized linear model for Y given A and W. We require our estimators to be consistent regardless of whether the assumptions of the generalized linear model are true; that is, we require our estimators to be consistent for any $P \in \mathcal{M}$, even if $P \notin \mathcal{M}_{sub}$. If the assumptions of the submodel do hold (i.e., if $P \in \mathcal{M}_{sub}$), we would like our estimators to achieve the semiparametric efficiency bound.

The unadjusted estimator of ψ, which ignores the baseline variables W, is defined as

$$\frac{\sum_{i=1}^n Y_i A_i}{\sum_{i=1}^n A_i} - \frac{\sum_{i=1}^n Y_i \left(1 - A_i\right)}{\sum_{i=1}^n \left(1 - A_i\right)}. \quad (3)$$

It is consistent and asymptotically normal. A disadvantage is that it ignores information in the baseline variables W. In doing so, it may sacrifice considerable precision compared with estimators discussed in the succeeding paragraphs.

All of the estimators we consider, except the unadjusted estimator, use at least one working model. A working model, in our context, is a parametric model used in constructing an estimator, which we do not assume to be correctly specified. That is, the population

distribution P is not assumed to obey the constraints of the working model. For example, for a binary outcome, if the working model for $P(Y = 1 | A, W)$ is $\text{logit}^{-1}(\beta_0 + \beta_1 A + \beta_2 W)$, we neither assume that the correct terms are included in the model nor assume that $P(Y = 1 | A, W)$ has this functional form.

The estimators we consider require two types of working models: an outcome regression model for $P(Y | A, W)$ and a propensity score model for the study arm assignment given baseline variables $P(A | W)$. For binary Y, logistic regression working models are used for both $P(Y | A, W)$ and $P(A | W)$. By randomization, A and W are independent, which implies $P(A = a | W)$ does not depend on W. Therefore, the working propensity score model will be correctly specified as long as it contains at least an intercept term.

3.3. Desired estimator properties

Our goal is to identify a consistent estimator for the risk difference ψ with all the following properties:

(A)   It can lead to substantial gains in precision compared to the unadjusted estimator, when baseline variables are correlated with the outcome.

(B)   It has been been proved that the estimator has equal or greater asymptotic precision than the unadjusted estimator, for any $P \in \mathcal{M}$.

(C)   It is locally, semiparametric efficient with respect to a given logistic regression model (denoted by Q in Section 4).

(D)   It is simple to implement using standard statistical software and does not require solving a non-convex optimization problem.

(E)   It is always in the interval $[-1, 1]$ (and so always lies in the parameter space determined by the risk difference).

Lacking property B means that, to the best of our knowledge, it has not been proved that the estimator has equal or better asymptotic precision than the unadjusted estimator for each $P \in \mathcal{M}$.

4. Estimators

A multitude of estimators have been developed that achieve some or all of the aforementioned properties. All of the estimators in this section are consistent for any $P \in \mathcal{M}$. These estimators vary in their other asymptotic properties and whether they involve solving challenging optimization problems. We describe some of these estimators in the succeeding paragraphs and focus on properties B–E because these are theoretical properties that are straightforward to check. The performance of a subset of these estimators is compared through simulations in Sections 5 and 6. Table I summarizes which properties among A–E are achieved by this subset of estimators.

4.1. Estimators achieving a subset of properties B–E

We first define the inverse probability weighted (IPW) estimator. Let $\widetilde{W}$ be a column vector of prespecified functions of the baseline variables W in which the first component is the

constant 1, for example, $\tilde{W} = (1, W_1, W_2, W_1 W_2, W_3)^T$, where the superscript T denotes the transpose. Let a be a column vector the same length as $\tilde{W}$ and define $\text{expit}(x) = \text{logit}^{-1}(x) = \exp(x)/\{1 + \exp(x)\}$. Let $g(\tilde{W}, a) = \text{expit}(a^T \tilde{W})$ denote a logistic regression working model for $P(A = 1|W)$, and let $\hat{\alpha}$ be the corresponding maximum likelihood estimator of a.

The IPW estimator of ψ is

$$\left\{ \sum_{i=1}^{n} \frac{A_i}{g\left(\tilde{W}_i, \hat{\alpha}\right)} \right\}^{-1} \sum_{i=1}^{n} \frac{Y_i A_i}{g\left(\tilde{W}_i, \hat{\alpha}\right)} - \left\{ \sum_{i=1}^{n} \frac{1 - A_i}{1 - g\left(\tilde{W}_i, \hat{\alpha}\right)} \right\}^{-1} \sum_{i=1}^{n} \frac{Y_i (1 - A_i)}{1 - g\left(\tilde{W}_i, \hat{\alpha}\right)}. \tag{4}$$

The IPW estimator satisfies properties B (as shown by Shen *et al.* [14] building on results of Robins *et al.* [15] and Rotnitzky *et al.* [16]) and D, but neither C nor E. The practical impact, as shown in [14], is that, although these estimators can improve efficiency compared with the unadjusted estimator, they may fail to fully leverage information in baseline variables compared with locally efficient estimators; our simulation studies are consistent with this finding.

The model standardization approach [8, 17, 18] utilizes a working outcome regression model to estimate ψ. Let $\beta^{(0)}$ and $\beta^{(1)}$ each be column vectors of the same length as $\tilde{W}$. For each study arm $a \in \{0, 1\}$, define a logistic regression working model $Q^{(a)}(\tilde{W}, \beta^{(a)})$ for $P(Y = 1|A = a, W)$, for example, $Q^{(a)}(\tilde{W}, \beta^{(a)}) = \text{expit}(\beta^{(a)T}\tilde{W})$. Define $Q = (Q^{(0)}, Q^{(1)})$. The terms in these models and in $g(\tilde{W}, a)$ need not be the same, but we use the same terms here for simplicity. Fit the logistic regression model $Q^{(1)}(\tilde{W}, \beta^{(1)})$ for $P(Y = 1|A = 1, W)$ by maximum likelihood estimation using only data from participants with $A = 1$; similarly, fit the logistic regression model $Q^{(0)}(\tilde{W}, \beta^{(0)})$ for $P(Y = 1|A = 0, W)$ by maximum likelihood estimation using only data from participants with $A = 0$. Let $\beta_{mle}^{(1)}$ and $\beta_{mle}^{(0)}$, respectively, denote the coefficient vectors corresponding to these model fits. The model standardization estimator of ψ is defined as

$$\frac{1}{n}\sum_{i=1}^{n} Q^{(1)}\left(\tilde{W}_i, \beta_{mle}^{(1)}\right) - \frac{1}{n}\sum_{i=1}^{n} Q^{(0)}\left(\tilde{W}_i, \beta_{mle}^{(0)}\right). \tag{5}$$

Each sum in (5) is taken over all n participants, regardless of their study arm assignment A. Expression (5) is an outcome regression estimator because it is computed by taking empirical means of a regression model fit. Outcome regression estimators are guaranteed to be within the range of the regression function, for example, in the range $[-1, 1]$ when $Q^{(a)}$ is a logistic regression model as aforementioned. Moore and van der Laan [8] prove that this estimator is consistent for any $P \in \mathscr{M}$ (even under arbitrary misspecification of the working models $Q^{(a)}$) and has properties C, D, and E. However, it lacks property B.

The doubly-robust weighted least squares (DR-WLS) estimator is attributed to Marshall Joffe by Robins *et al.* [19]. It is similar to the model standardization estimator, except that each working model $Q^{(a)}$ for $P(Y = 1|A = a, W)$, is fit using weighted logistic regression.

The model $Q^{(1)}$ is fit using weights $1/g\left(\tilde{W}, \hat{\alpha}\right)$ and uses only data from participants with A = 1; the model $Q^{(0)}$ is fit using weights $1/\{1 - g(\tilde{W}, _{\hat{\alpha}})\}$ and uses only data from participants with $A = 0$. Denote the corresponding estimates of $\beta^{(0)}$, $\beta^{(1)}$ by $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$, respectively. The DR-WLS estimator of ψ is (5) except that each $\beta_{mle}^{(a)}$ is replaced by $\hat{\beta}^{(a)}$. The DR-WLS estimator has properties C, D, and E, but not B.

Additional estimators exist that achieve some, but not all, of properties B–E. The 'direct' implementation of the method by Zhang *et al.* [20] using a logistic regression working model fit with iteratively reweighted least squares has properties C, D. and E, but not B. The estimators of Scharfstein *et al.* [21, Section 3.2], Moore and van der Laan [8], Tsiatis *et al.* [22], and Rosenblum and van der Laan [18] have properties C, D, and E, but not B. The estimators of Rubin and van der Laan [7], and Cao *et al.* [23] have properties B, C, and E, but not D. A variety of estimators are presented in Kang and Schafer [24], and Robins *et al.* [19] that have subsets of properties B–E; Tan [4] and Rotnitzky *et al.* [5] provide detailed comparisons of the theoretical properties of these estimators.

4.2. Estimators that achieve all properties A–E

We define an estimator based on the general class of estimators in Section 3 of Rotnitzky *et al.* [5], which we modified to fit our context of a randomized trial. The estimator is a simplified version of a special case from this general class. It uses the quantities calculated in the DR-WLS estimator and is designed to achieve property B while preserving properties C, D, and E. We refer to this estimator as PLEASE, which stands for 'precise, locally efficient, augmented, simple estimator.' First, $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$ are computed as described earlier for the DR-WLS estimator. For each study arm $a \in \{0, 1\}$, define the initial estimator of $E(Y|A = a)$ (equivalently, of $P(Y = 1|A = a)$) to be

$$\hat{\mu}_a = \frac{1}{n}\sum_{i=1}^{n}Q^{(a)}\left(\tilde{W}_i, \hat{\beta}^{(a)}\right). \tag{6}$$

Let $\gamma = (\gamma_0, \gamma_1)$. PLEASE consists of two additional steps:

Step 1: For each $a \in \{0, 1\}$, using $\hat{\beta}^{(a)}$ and $\hat{\mu}_a$ computed as described earlier, define the following new variable:

$$u_a\left(\tilde{W}\right) = Q^{(a)}\left(\tilde{W}, \hat{\beta}^{(a)}\right) - \hat{\mu}_a.$$

Fit the following augmented logistic regression model for $P(A = 1|W)$:

$$g_{aug}\left(\tilde{W}, \alpha, \gamma\right) = expit\left(\alpha^T\tilde{W} + \gamma_0 u_0\left(\tilde{W}\right) + \gamma_1 u_1\left(\tilde{W}\right)\right),$$

to obtain estimated coefficients $\tilde{\alpha}$, $\tilde{\gamma}$.

Step 2: Recompute the DR-WLS estimator, using $g_{aug}\left(\tilde{W}, \tilde{\alpha}, \hat{\gamma}\right)$ in place of $g\left(\tilde{W}, \hat{\alpha}\right)$ in the weights.

This is the PLEASE estimator.

The R and SAS codes for the aforementioned estimator are provided in the Supplementary Materials. We generalize the aforementioned procedure to estimate any smooth contrast between $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$, for example, the log odds ratio (2) or the relative risk and to handle non-binary outcomes such as continuous-valued or count-valued Y, in Section 8.

We next discuss the intuition behind the IPW, model standardization, DR-WLS, and PLEASE estimators. All except the model standardization estimator involve the propensity score model g for $P(A = 1|W)$. Fitting such a model may appear odd because, in a randomized trial, A and W are independent by design, so that $P(A = 1|W) = P(A = 1)$. Why fit a model for a relationship that is already known? Intuitively, the model fit $g\left(\tilde{W}, \hat{\alpha}\right)$ is designed to capture chance imbalances of prognostic covariates between study arms in a given data set. As a simple example, let W be a binary indicator of severe disease at baseline, and let $\tilde{W} = (1, W)^T$; if, proportionally, more participants with $W = 1$ are in the treatment arm $A = 1$ in a given data set, the logistic regression model fit

$g\left(\tilde{W}, \hat{\alpha}\right) = expit\left(\hat{\alpha}_0 + \hat{\alpha}_1 W\right)$ will have $\hat{\alpha}_1 > 0$. The IPW estimator uses $g\left(\tilde{W}, \hat{\alpha}\right)$ to upweight outcomes of participants who are under-represented in a given arm, and to downweight outcomes that are over-represented. The result is a rebalanced estimator of the average treatment effect. The model standardization estimator achieves a similar goal by first estimating the relationship between Y and W within each arm, and then standardizing to the empirical distribution of W pooled across arms. The DR-WLS estimator combines these two methods of rebalancing into a single estimator.

The only property among A–E that DR-WLS lacks is property B. The PLEASE estimator builds on DR-WLS to achieve property B by augmenting the propensity score model g with the carefully selected covariates u_0, u_1. These covariates are simplifications of the covariates proposed by Rotnitzky *et al.* [5]; the simplification is possible because we are in the context of a randomized trial, while [5] handled the general case of observational studies. As discussed by [5], adding any baseline covariate to the propensity score model in the DR-WLS estimator is guaranteed to improve or leave unchanged the asymptotic variance of this estimator. This was a key insight of Robins *et al.* [15]. By examining the influence function of the DR-WLS estimator, [5] cleverly deduced covariates that ensure property B.

Intuitively, the covariate $u_a\left(\tilde{W}\right) = Q^{(a)}\left(\tilde{W}, \hat{\beta}^{(a)}\right) - \hat{\mu}_a$ is intended to approximate the difference $E(Y|A = a, W) - E(Y|A = a)$. This difference represents how much the population mean of Y within strata of W in arm $A = a$ differs from the population mean of Y in arm $A = a$. The difference $E(Y|A = a, W) - E(Y|A = a)$ roughly characterizes the mean influence of adding one more participant with covariate value W to arm a on the unadjusted estimator of $E(Y|A = a)$. Because of this, $u_a(\tilde{W})$ is a particularly appealing covariate to adjust for.

The PLEASE is closely related to the estimator from Section 3.1 of the Harvard University technical report of Robins *et al.* [25], denoted there as $\hat{\mu}_{DR}\left(\hat{m}_{t,ITER-WLS}; \hat{\pi}^{(t)}_{ITER-WLS}\right)$; that estimator has properties A–E for separately estimating $P(Y = 1|A = a)$ for each $a \in \{0, 1\}$ but (unlike PLEASE) does not have property B when it is used to estimate the risk difference ψ by taking the difference of its estimates of $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$). This is not surprising because the estimator of [25] was not designed to achieve this latter goal.

Another estimator that achieves all of properties B–E is the estimator $\tilde{\mu}_{LIK2}$ of Tan [4], who provides software using the R package **iWeigReg**. This estimator can be modified to also achieve improved local efficiency (described in the succeeding paragraphs) as shown in Section 5.4 of [4], but this requires solving a non-convex optimization problem; we use the version without this modification (called ate.clik in the package **iWeigReg**) in our simulation study, and call it the Tan estimator.

4.3. Estimators with enhanced efficiency properties that require solving a non-convex optimization problem

The general classes of estimators of Rotnitzky *et al.* [5], and of Gruber and van der Laan [6] have properties B, C, and E, but lack property D because each requires solving a non-convex optimization problem. We consider an example from each of these classes of estimators. Let Rotnitzky *et al.*$_{K=1}$ denote an estimator from the class of Rotnitzky *et al.* [5] with $K = 1$, which involves similar augmenting covariates as in PLEASE except u_0 and u_1 are divided by $1 - g\left(\tilde{W}, \tilde{\alpha}\right)$ and $g\left(\tilde{W}, \tilde{\alpha}\right)$, respectively, and there are two additional covariates included in the model for $P(A = 1|W)$ in Step 1. These additional covariates are defined in [5, Section 2.2] as a solution to a non-convex optimization problem given there. By adding these covariates, the estimator Rotnitzky *et al.*$_{K=1}$ of ψ achieves improved local efficiency as defined by Tan [4], which means the estimators are as or more precise, asymptotically, as each estimator in a certain class of locally efficient estimators. We give these additional covariates in Section 2 of the Supplementary Material.

Rotnitzky *et al.* [5] constructed these covariates by first deriving a formula for the asymptotic variance of a large class of estimators that augment DR-WLS as in Steps 1 and 2 of PLEASE. This formula is a function of the augmenting variables. They next define a parametric class of potential augmenting covariates indexed by the parameter vector η. They then minimize an empirical estimate of the resulting asymptotic variance over all possible values of η, using the general technique of empirical efficiency maximization from Rubin and van der Laan [7]. Denote the minimizer by η^*. The last step is that Rotnitzky *et al.* [5] augment the propensity score model using the covariate corresponding to η^*.

The aforementioned minimization problem is not convex in η, and therefore, existing solvers (such as optim or nlm in R) are not guaranteed to converge to the true optimum (or even to a local optimum). This leads to practical concerns if convergence is not achieved for a given dataset. We note that the estimator is still well defined even in the case of non-convergence (using the final value output by the optimization algorithm). A second concern is that these estimators are only guaranteed to be consistent if the computed solution to the optimization

problem converges in probability to some limit as the sample size goes to infinity; as pointed out by Rubin and van der Laan [26], this convergence is not guaranteed.

In exchange for added computational complexity, the estimators of Rotnitzky *et al.* [5], and Gruber and van der Laan [6] achieve improved local efficiency. In our simulations, we used the R code of [5, Supplementary Material], which we modified to target the parameter ψ, as described in Section 2 of our Supplementary Material. We use the R code of Gruber and van der Laan [6, Appendix] to implement the estimator that we call the Gruber and van der Laan estimator.

5. Simulation study based on the MISTIE II trial

5.1. Simulation design

We construct data generating mechanisms based on resampling in order to closely mimic the relationships between baseline variables and outcomes observed in the MISTIE II trial described in Section 2.1. The baseline variables for each participant are $W = (W_1, W_2, W_3)$ = (age, ICH volume, and NIHSS). We compare the following estimators of the risk difference: unadjusted, IPW, model standardization, DR-WLS, Tan, PLEASE, Rotnitzky *et al.*$_{K=1}$, and Gruber and van der Laan. We refer to all but the first of these as adjusted estimators because they adjust for baseline variables.

Simulations are conducted under the following four types of data generating distributions (called scenarios):

> Scenario 1: Y and W dependent; zero average treatment effect;
>
> Scenario 2: Y and W dependent; positive average treatment effect;
>
> Scenario 3: Y and W independent; zero average treatment effect; and
>
> Scenario 4: Y and W independent; positive average treatment effect.

In scenarios 1 and 2, W is prognostic for Y, and there is potential for gaining precision by adjusting for W; the modified R^2 (defined in Section 7) due to adjustment is 27% in scenario 1 and is 22% in scenario 2.

In scenarios 3 and 4, baseline variables are not prognostic for the outcome; therefore, the estimators that adjust for chance imbalances in W are adjusting for noise. We examine this case to determine how much efficiency is lost because of adjusting for variables with no prognostic value. Although such losses disappear as sample size goes to infinity, it is important to determine the magnitude of such potential efficiency losses at realistic sample sizes.

For each scenario, we generated 100,000 simulated randomized trials, each with $n = 412$ participants. This sample size was selected to approximately mimic the projected size of the planned phase III trial that is a follow-up to the MISTIE II trial.

A simple way to define a data generating distribution P based on the MISTIE II trial data, would be to use the empirical distribution. This distribution can be simulated by resampling

triples (W, A, Y) with replacement from the MISTIE II trial data. Unfortunately, A and W are dependent under this distribution because A and W have non-zero correlations in the MISTIE II trial data (as would be expected for essentially any trial data set). Generating simulated trials from this empirical distribution would violate the randomization assumption from Section 3.1 that, under P, A is assigned independent of W. We next describe modifications to the empirical distribution that were performed in order to satisfy the randomization assumption.

For scenarios 1 and 2, we resampled pairs (W, Y) with replacement from the MISTIE II trial data. This preserves the relationship between baseline variables and the outcome from the MISTIE II trial data. In scenario 1, we generated A independent of (W, Y), with probability 1/2 of being treatment or control. For this data generating distribution, we have $P(Y = 1|A = 1) = P(Y = 1|A = 0) = 0.32$, which implies that the average treatment effect is $\psi = 0$.

For scenario 2, we construct a data generating distribution similar to scenario 1, except with average treatment effect $\psi = 0.12$; we chose this value of ψ because it is the average treatment effect observed in the MISTIE II trial based on the unadjusted estimator. The initial step in scenario 2 is that data is generated as in scenario 1. Next, to induce the positive average treatment effect $\psi = 0.12$, for each simulated participant who was initially assigned $A = 1$ and $Y = 0$, we randomly replaced Y by 1 with probability $q = 0.18$ (by an independent Bernoulli draw for each such participant). The value $q = 0.18$ was determined by the formula $q = 0.12/(1 - 0.32)$ (rounded to two decimal places), that is, the target average treatment effect 0.12 divided by the probability of $Y = 0$ given $A = 1$ based on the initial step (where data was generated as in scenario 1). The combination of the aforementioned two steps results in $P(Y = 1|A = 0) = 0.32$ and $P(Y = 1|A = 1) = 0.32 + (1 - 0.32)q = 0.44$, which implies $\psi = 0.12$. The value of q was set before running any of the simulations.

For scenarios 3 and 4, baseline variables W for each participant were randomly drawn with replacement from the MISTIE II trial data. This results in the marginal distribution of W being the empirical distribution of the MISTIE II trial data. Study arm assignment A was generated independent of W, with probability 1/2 of being treatment or control. In scenario 3, Y is a random draw from a Bernoulli distribution with probability 0.32 (the marginal probability of $Y = 1$ in the MISTIE II trial, pooling all participants). Because A is generated independent of (W, Y), this results in $\psi = 0$. In scenario 4, the conditional distribution of Y, given $A = a$ and W is set to be Bernoulli with probability p_a of $Y = 1$, where p_a is the observed proportion of successes in each treatment group in the MISTIE II trial ($p_0 = 0.24$, and $p_1 = 0.36$). This results in an average treatment effect of $\psi = 0.36 - 0.24 = 0.12$.

Each estimator uses the same initial working models in each scenario (though the augmenting variables differ by estimator). The propensity score working model and the outcome regression working model each have an intercept and a main term for each component of W. Because treatment is randomized in all scenarios, we have $P(A = 1|W) = 1/2$; so the propensity score working model is correctly specified. The working outcome regression model is misspecified in scenarios 1 and 2 because the true joint distribution of (W, Y) is based on resampling from the MISTIE II trial data, which does not obey the constraints imposed by the logistic regression working model. Nonetheless, the adjusted

estimators using this working model can still gain substantial efficiency compared with the unadjusted estimator, as we show in the succeeding paragraphs. In scenarios 3 and 4, (A, Y) is independent of W, which implies for each $a \in \{0, 1\}$, $P(Y = 1 | A = a, W) = P(Y = 1 | A = a)$. Because the working outcome regression model has separate components $Q^{(a)}$ for each $a \in \{0, 1\}$ it is correctly specified in scenarios 3 and 4.

5.2. Simulation results

Table II displays the results of the simulation study for scenarios 1–4. The bias of all estimators is quite small and is similar for all estimators and scenarios. In scenarios 1 and 2, where W is prognostic for Y, there were large gains in relative efficiency for all the adjusted estimators relative to the unadjusted estimator. Unsurprisingly, the locally efficient estimators (model standardization, DR-WLS, Tan, PLEASE, Rotnitzky $et\ al._{K=1}$, and Gruber and van der Laan) were all more precise than the IPW estimator. The efficiency gains for PLEASE are slightly larger than those for the DR-WLS estimator, and this difference is due to the augmentation Steps 1 and 2. The relative efficiency was generally similar for the model standardization, DR-WLS, Tan, PLEASE and Gruber and van der Laan estimators, while Rotnitzky $et\ al._{K=1}$ had greater relative efficiency than all of these. A possible explanation for the increased efficiency gains for Rotnitzky $et\ al._{K=1}$ is that it uses two additional covariates in the augmented propensity score model.

To illustrate how efficiency gains in scenarios 1 and 2 translate into reductions in sample size to achieve a desired power at a given alternative, consider PLEASE. Its relative efficiency compared with the unadjusted estimator in scenario 1 is 1.43, which is equivalent (asymptotically) to a $1 - (1/1.43) \approx 30\%$ reduction in the sample size required to achieve a desired power. Similarly, in scenario 2, the relative efficiency 1.29 translates into a $1 - (1/1.29) \approx 22\%$ sample size reduction.

In scenarios 3 and 4, where baseline variables are not prognostic, the relative efficiency loss of the adjusted estimators ranged from 0.7% to 2.8% (compared with the unadjusted estimator). The efficiency loss of the PLEASE and Rotnitzky $et\ al._{K=1}$ estimators, though small, exceeded that of the other adjusted estimators.

Two types of convergence problems occurred in some estimators. First, for estimators that involve solving a convex or non-convex optimization problem, the algorithm computing this may fail to converge. In this case, the optimization algorithm may still output a solution (e.g., the best value achieved by the optimization algorithm during its iterations) or may output no solution. If a solution is provided, the estimator can still be computed using this solution. In our simulations, the *optim* algorithm used by Rotnitzky $et\ al._{K=1}$ always converged when we set the maximum number of iterations at 2000. For the estimators of Gruber and van der Laan, and of Tan, the optimization algorithm did not output any solution if the algorithm did not converge. For the Gruber and van der Laan estimator, this occurred with frequency 1.4% and 0.1% of the simulated trials for scenarios 1 and 2, respectively. For the Tan estimator, this occurred with frequency 0.1% of the simulated trials for each scenario 1 and 2.

The second type of convergence problem is that the maximum likelihood estimator for the augmented propensity score model for Rotnitzky *et al.*$_{K=1}$ failed to converge because of quasi-complete separation [27] in 0.3% and 0.03% of the simulated studies for scenarios 1 and 2, respectively; this caused the weighted logistic regression in Step 2 to fail to converge as well. In such cases, we set the final output of Rotnitzky *et al.*$_{K=1}$ to be the estimator obtained prior to augmentation, that is, the DR-WLS estimator.

6. Simulation study based on PEARLS trial

6.1. Simulation design

We used baseline variables $W = (W_1, W_2, W_3) =$ (gender, RNA viral load, CD4+ cell count) from the PEARLS trial. Similar scenarios as 1–4 from Section 5 were used, except replacing all empirical distributions by those in the PEARLS trial. Details are given in Sections 3 and 4 of the Supplementary Material. In each scenario, we generated 100,000 trials of size 1044 participants, which is the sample size of the PEARLS trial counting those in treatment arms A and B.

6.2. Simulation results

Table III displays the results of the simulation study. All estimators performed similarly. In scenarios 1 and 2, the efficiency gains of the adjusted estimators range from 2–3%. This set of simulations shows that, when baseline variables are weakly prognostic for the outcome, the efficiency gains from adjusting for these variables is small, as may be expected. This raises the question of how prognostic the baseline variables need to be in order to result in substantial efficiency gains. We discuss this in the next section.

In scenarios 3 and 4, the loss of efficiency ranged from 0% to 0.9%. These smaller efficiency losses compared with the MISTIE II trial simulations from Section 5 are not surprising because of the larger sample size in the PEARLS simulations.

7. Quick assessment of whether adjusting for baseline variables can give substantial gains

The primary statistical analysis for a confirmatory trial must be fully specified before the trial is started. If using an adjusted estimator, all details of its implementation need to be specified, for example, the working models to be used and the variables to be included in them. When there are many baseline variables, identifying which to include in the working models is a practical concern. In planning a phase III trial, data from the corresponding phase II trial could be used in a quick assessment of whether a few baseline variables have potential to substantially improve efficiency, using a method of [7,8].

Given data from a completed trial, the procedure involves computing the modified R^2 (in the ordinary least squares sense), which we define next. The modified R^2 compares an outcome regression model that includes A and the baseline variables W versus an outcome regression model including only A. First, one computes the sample mean among those assigned to treatment (denoted m_1) and among those assigned to control (denoted m_0). Second, for each

study arm $a \in \{0, 1\}$, one fits an outcome regression model $Q^{(a)}(\tilde{W}, \beta^{(a)})$ for $E(Y|A = a, W)$. Lastly, one computes

$$1 - \frac{\Sigma_i \left\{ Y_i - Q^{(A_i)} \left(\tilde{W}, \hat{\beta}^{(A_i)} \right) \right\}^2}{\Sigma_i \left(Y_i - m_{A_i} \right)^2}. \tag{7}$$

The intuition for (7) is that it captures how much additional variance in the outcome is accounted for by the baseline variables, above what is explained by A alone. The expression (7) gives a rough approximation of how much relative efficiency may be gained by using an adjusted estimator based on an outcome regression working model, compared with the unadjusted estimator. However, (7) only reflects the precision gains from using the outcome regression model and does not account for additional precision gains from augmenting the propensity score model as in the estimators: Tan, PLEASE, Rotnitzky *et al.*$_{K=1}$, and Gruber and van der Laan. Also, expression (7) does not account for the correlation between the adjusted estimators of $E(Y|A = 1, W)$ and of $E(Y|A = 0, W)$, which can impact the variance of the risk difference estimator.

We propose to compute a modified version of (7) using leave-one-out cross-validation. The purpose is to avoid being fooled by a model overfit; for example, even independent baseline variables will explain a portion of the outcome variance because of chance variation in a given data set. The more variables one adds to the outcome regression working model, the more variance will be explained, but this may fail to translate into precision gains if it is because of overfitting. The cross-validation is implemented by, for each data point (W_i, A_i, Y_i), fitting the outcome regression model for $P(Y = 1|A = A_i, W)$ on all data in arm $A = A_i$ except observation i, then computing the squared difference between Y_i and the model-based prediction for Y_i; this is done for each i and then summed to compute the cross-validated version of the numerator in (7). For the denominator, for each i the sample mean in arm $A = A_i$ is computed using all data from this arm except observation i, and then the squared difference between Y_i and this sample mean is computed; this is done for each i and then summed to compute the cross-validated version of the denominator in (7).

The output of this quick assessment method can help determine whether it is worth the effort to perform a simulation study to evaluate if an adjusted estimator is likely to provide substantial precision gains. A rule of thumb is to conduct such a simulation study if the cross-validated version of (7) is 10% or greater. The value of (7) without cross-validation was 34% and 5% for the MISTIE and PEARLS trial data, respectively, and with cross-validation was 22% and 3.6%, respectively. We demonstrated in the simulations the potential large efficiency gains from using an adjusted estimator in the MISTIE trial data and modest efficiency gains within the PEARLS trial data.

8. Generalization of PLEASE to estimate relative risk and log odds ratio, and to handle non-binary outcomes

Consider any smooth contrast $f(P(Y = 1|A = 1), P(Y = 1|A = 0))$ between $P(Y = 1|A = 1)$ and $P(Y = 1|A = 0)$. The following generalization of PLEASE can be used to estimate this contrast: replace Step 2 by

1. **Step 2':** For each $a \in \{0, 1\}$, let $\bar{\beta}^{(a)}$ denote the estimated coefficient vector in the weighted logistic regression fit of $Q^{(a)}$ as defined for the DR-WLS estimator, except using $g_{aug}\left(\tilde{W}, \tilde{\alpha}, \hat{\gamma}\right)$ in place of $g\left(\tilde{W}, \hat{\alpha}\right)$ in the weights. Define $\bar{\mu}_a$ to be (6) except with $\hat{\beta}^{(a)}$ replaced by $\bar{\beta}^{(a)}$. The estimator of the contrast $f(P(Y = 1|A = 1), P(Y = 1|A = 0))$ is $f\left(\bar{\mu}_1, \bar{\mu}_0\right)$.

For example, the contrast $f(x, y) = x - y$ corresponds to the risk difference (1); in which case, Step 2' reduces to Step 2. The contrast $f(x, y) = \text{logit}(x) - \text{logit}(y)$ corresponds to the log odds ratio (2). The contrast $f(x, y) = x/y$ corresponds to the relative risk. If f is a continuously differentiable function (as is the case for the examples earlier as long as each $P(Y = 1|A = a)$ is neither 0 nor 1), then properties B–D and an extension of property E (where we replace the interval $[-1, 1]$ by the range of f) hold for the aforementioned generalization of PLEASE, as proved in Section 5 of the Supplementary Material.

Consider the case where Y is non-binary valued, for example, Y is continuous-valued or count-valued. Let the parameter of interest be any smooth contrast $f(E(Y|A = 1), E(Y|A = 0))$, for example, the difference in population means $f(x, y) = x - y$. The PLEASE estimator can be generalized to handle this case by replacing Step 2 by Step 2' as aforementioned and replacing the working logistic regression model $Q^{(a)}$ by a corresponding generalized linear model with canonical link function. For example, a standard linear regression model (with identity link function) could be used if Y is continuous-valued, or a Poisson regression model (with log link function) could be used if Y is count-valued. For the resulting estimator, generalizations of properties B–E hold, where we replace 'logistic regression' in C by the corresponding generalized linear model, and we replace the interval $[-1, 1]$ in E by the range of f.

9. Discussion

If the quick assessment method from Section 7 indicates potential for a substantial precision gain (e.g., at least 10%) from adjusting for prognostic baseline variables, we recommend considering one of the following adjusted estimators: Tan, PLEASE, Rotnitzky *et al.*$_{K=1}$, and Gruber and van der Laan. Trade-offs to consider when selecting an estimator include complexity of implementation, theoretical properties, and ease of communicating the approach. The estimators of Rotnitzky *et al.*, and Gruber and van der Laan require non-convex optimization so are computationally more complex than PLEASE and the Tan estimator; however, the estimators of Rotnitzky *et al.* and Gruber and van der Laan have enhanced efficiency properties, as described in Section 4.2. PLEASE is the simplest

estimator to describe among those recommended. The greatest efficiency gain was from Rotnitzky *et al.*, which outperformed all other estimators in scenarios 1 and 2 of the MISTIE simulations; it also had the largest efficiency loss when the baseline variables were not prognostic (i.e., independent) of the outcome. If one is planning a phase III trial, the aforementioned trade-offs for different estimators could be informed by conducting a simulation study based on phase II trial data using the general approach of Section 5.

The decision of which estimator to use depends on the trial designer's preferences regarding the aforementioned trade-offs. These preferences may be influenced by prior knowledge of the likelihood of different scenarios (e.g., how likely is it that the prognostic variables identified from a phase II trial will be similarly prognostic in a future phase III trial), and on the relative value of different magnitudes of precision gains/losses. In principle, this problem could be formalized using decision theory, where one specifies a loss function representing the penalty for different types of mistakes (e.g., selecting the unadjusted estimator when in truth an adjusted estimator would have provided much more precision or selecting an adjusted estimator when the baseline variables are not prognostic) and then determines an optimality criterion such as the weighted average of the loss function over likely scenarios. This is an area of future research and could be informed by examining pairs of phase II/phase III trial data to determine how often prognostic variables in the phase II trial are similarly prognostic in the phase III trial.

In practice, we recommend *a priori* specification of a few key baseline variables to be included in the adjusted estimator. We caution against trying many combinations of variables to see which leads to the maximum efficiency gain; the danger is that one may inadvertently be selecting for noise if too many comparisons are performed. It is an area of future work to consider a variety of potential variables and working models and to use cross-validation to select the most promising. Another area of future research is to explore the impact of selecting which baseline variables to use from the trial data itself (rather than from a prior dataset). This has been considered by Tian *et al.* [28] and Yuan *et al.* [29].

We focused on improving precision for estimation of the average treatment effect, but the same ideas can be used to increase power in hypothesis tests. For example, a Wald statistic can be constructed by dividing an adjusted estimator by its estimated standard error computed, for example, using the nonpara-metric bootstrap. The efficiency gains for estimation then directly translate into power gains for testing the null hypothesis of zero average treatment effect.

In planning a trial in which adjusted estimators will be used, several options are available. A conservative approach is to plan as if there will be no efficiency gain from adjustment and setting the total sample size accordingly. Then precision gains from using an adjusted estimator, if they occur, will translate into improved power compared with what is planned for. A less conservative approach would be to plan the sample size assuming a relative efficiency gain will be achieved; however, this is risky because there is no guarantee that the correlation structure in, for example, a phase II trial will be identical to that in a follow-up phase III trial.

It is straightforward to incorporate adjusted estimators into group sequential designs that involve preplanned interim analyses where the trial may be stopped early for efficacy or futility. One would simply use an adjusted estimator at each interim analysis. Because the joint distribution of its value computed at each interim analysis has a canonical joint distribution [30], standard group sequential methods can be used.

We focused on the case where the trial data consists of n-independent, identically distributed realizations $\{(W_i, A_i, Y_i)\}_{i=1}^n$ of the random vector (W, A, Y) as defined in Section 3.1. This corresponds to randomization that is independent for each participant. The estimators in this paper can also be applied in the case where block randomization is used, either with or without stratification by a baseline covariate (that we assume can take a finite set of values). For the case of block randomization stratified by a baseline covariate S (e.g., study site), any of the estimators from Section 4.2 could be applied separately within each stratum $S = s$ to estimate $E(Y|A = a, S = s)$ for each arm $a \in \{0, 1\}$; the weighted combination of stratum-specific estimates, with weights equal to the empirical value of each proportion $P(S = s)$, can be used as an adjusted estimator of $E(Y|A = a)$ for each arm $a \in \{0, 1\}$. This takes advantage of both the stratification by design, and the potential for precision gains by adjusting for prognostic baseline variables other than S. The reason we recommend applying an adjusted estimator for each stratum of S separately, and then combining the estimates, is that the resulting estimator is then consistent without having to assume that patient populations are identical at different values of S (e.g., at different sites). In general, we expect the precision gains from adjustment to be reduced if certain covariates are already balanced by design. However, the number of covariates that can be balanced by design is limited, and additional precision gains can be achieved by adjusting for covariates that are not balanced by design. Exploring the interaction between stratified randomization and the aforementioned adjusted estimators is an area of future research.

Designs with covariate-adaptive randomization are another alternative; these are logistically more complicated than the aforementioned designs because each participant's randomization probability depends on the data from all previous participants. It is generally recommended to adjust for covariates used in the adaptive randomization procedure, and it is still possible to gain precision by additional adjustment for prognostic baseline variables not included in the randomization procedure [31].

An alternative to resampling pairs (W, Y) from the MISTIE II data would be to generate Y given W using, for example, a logistic regression model for Y given W. An advantage of conducting simulations based on resampling is that they more closely mimic the relationships between baseline variables and outcomes in a trial. Another advantage of simulations based on resampling is that they avoid the pitfall of generating data using a model similar in form to the working models used in estimators; this can lead to overly optimistic results because, in practice, one expects that the outcome regression model will be at least somewhat misspecified.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

References

1. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet. 2000; 355:1064–1069. [PubMed: 10744093]

2. Pocock SJ, Assmann SE, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in Medicine. 2002; 21(19):2917–2930. [PubMed: 12325108]

3. Yang L, Tsiatis AA. Effciency study of estimators for a treatment effect in a pretest–posttest trial. The American Statistician. 2001; 55:314–321.

4. Tan Z. Bounded, efficient and doubly robust estimating equations for marginal and nested structural models. Biometrika. 2010; 97:661–682.

5. Rotnitzky A, Lei Q, Sued M, Robins JM. Improved double-robust estimation in missing data and causal inference models. Biometrika. 2012; 99(2):439–456. [PubMed: 23843666]

6. Gruber S, van der Laan MJ. Targeted minimum loss based estimator that outperforms a given estimator. The International Journal of Biostatistics. 2012; 8(1) Article 11.

7. Rubin D, van der Laan MJ. Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. International Journal of Biostatistics. 2008; 4(1) Article 5.

8. Moore K, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation. Statistics in Medicine. 2009; 28(1):39–64. [PubMed: 18985634]

9. Mould WA, Carhuapoma JR, Muschelli J, Lane K, Morgan TC, McBee NA, Bistran-Hall AJ, Ullman NL, Vespa P, Martin NA, Awad I, Zuccarello M, Hanley DF, for the MISTIE investigators. Minimally invasive surgery plus recombinant tissue-type plasminogen activator for intracerebral hemorrhage evacuation decreases perihematomal edema. Stroke. 2013; 44(3):627–634. [PubMed: 23391763]

10. Campbell TB, Smeaton LM, Kumarasamy N, Flanigan T, Klingman KL, Firnhaber C, Grinsztejn B, Hosseinipour MC, Kumwenda J, Lalloo U, Riviere C, Sanchez J, Melo M, Supparatpinyo K, Tripathy S, Martinez AI, Nair A, Walawander A, Moran L, Chen Y, Snowden W, Rooney JR, Uy J, Schooley RT, De Gruttola V, Hakim JG, for the PEARLS study team of the ACTG. Efficacy and safety of three antiretroviral regimens for initial treatment of HIV-1: a randomized clinical trial in diverse multinational settings. PLoS Medicine. 2012; 9(8) Article e1001 290.

11. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. International Statistical Review / Revue Internationale de Statistique. 1991; 59(2):227–240. http://www.jstor.org/stable/1403444.

12. Freedman DA. Randomization does not justify logistic regression. Statistical Science. 2008; 23(2):237–249.

13. van der Vaart, AW. Asymptotic Statistics. Cambridge University Press; Cambridge: 1998.

14. Shen C, Li X, Li L. Inverse probability weighting for covariate adjustment in randomized studies. Statistics in Medicine. 2014; 33(4):555–568. [PubMed: 24038458]

15. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 1994; 89:846–866.

16. Rotnitzky A, Li L, Li X. A note on overadjustment in inverse probability weighted estimation. Biometrika. 2010; 97(4):997–1001. [PubMed: 22822256]

17. Austin PC. Absolute risk reductions, relative risks, relative risk reductions, and numbers needed to treat can be obtained from a logistic regression model. Journal of Clinical Epidemiology. 2010; 63:2–6. [PubMed: 19230611]

18. Rosenblum M, van der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. International Journal of Biostatistics. 2010; 6(1) Article 1.

19. Robins JM, Sued M, Lei-Gomez Q, Rotnizky A. Comment: performance of double-robust estimators when inverse probability weights are highly variable. Statistical Science. 2007; 22(4): 544–559.

20. Zhang M, Tsiatis AA, Davidian M. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. Biometrics. 2008; 64:707–715. [PubMed: 18190618]

21. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). Journal of the American Statistical Association. 1999; 94:1096–1146.

22. Tsiatis AA, Davidian M, Zhang M, Lu X. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. Statistics in Medicine. 2008; 27:4658–4677. [PubMed: 17960577]

23. Cao W, Tsiatis AA, Davidian M. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika. 2009; 96:723–734. [PubMed: 20161511]

24. Kang DYL, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). Statistical Science. 2007; 22(4):523 –580.

25. Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Double-robustness with improved efficiency in missing and causal inference models. Technical Report, Harvard School of Public Health. 2007

26. Rubin, D.; van der Laan, MJ. Working Paper Series. U.C. Berkeley Division of Biostatistics; 2008. Covariate adjustment for the intention-to-treat parameter with empirical efficiency maximization..

27. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. Biometrika. 1984; 71:1–10.

28. Tian L, Cai T, Zhao L, Wei LJ. On the covariate-adjusted estimation for an overall treatment difference with data from a randomized comparative clinical trial. Biostatistics. 2012; 13(2):256–273. [PubMed: 22294672]

29. Yuan S, Zhang HH, Davidian M. Variable selection for covariate-adjusted semiparametric inference in randomized clinical trials. Statistics in Medicine. 2012; 31:3789–3804. [PubMed: 22733628]

30. Jennison, C.; Turnbull, BW. Group Sequential Methods With Applications to Clinical Trials. Chapman and Hall/CRC Press; Boca Raton: 1999.

31. Lachin JM. Staistical properties of randomization in clinical trials. Controlled Clinical Trials. 1988; 9(4):289–311. [PubMed: 3060315]

Table I

Properties for each adjusted estimator used in our simulation studies.

Estimator	A	B	C	D	E
IPW	X	X		X	
Model standardization	X		X	X	X
DR-WLS	X		X	X	X
Tan	X	X	X	X	X
PLEASE	X	X	X	X	X
Rotnitzky *et al.*$_{K=1}$	X	X	X		X
Gruber and van der Laan	X	X	X		X

IPW, inverse probability weighted; DR-WLS, doubly-robust weighted least squares; PLEASE, precise, locally efficient, augmented, simple estimator. All estimators in the table are consistent.

Table II

Results of the 100,000 simulated randomized trials of size $n = 412$ patients based on the MISTIE II trial. For each of the seven adjusted estimators, relative efficiency is defined as the variance of the unadjusted estimator divided by the variance of the adjusted estimator.

Estimator	Bias	Variance	MSE	Rel.Efficiency
Scenario 1: Y and W dependent; zero average treatment effect				
Unadjusted	−0.000060	0.0021	0.0021	1.000
IPW	−0.000080	0.0016	0.0016	1.310
Model standardization	−0.000075	0.0015	0.0015	1.380
DR-WLS	−0.000074	0.0015	0.0015	1.390
Tan	−0.000072	0.0015	0.0015	1.440
PLEASE	−0.000058	0.0015	0.0015	1.430
Rotnitzky $et\ al._{K=1}$	−0.0000051	0.0013	0.0013	1.630
Gruber and van der Laan	−0.000073	0.0015	0.0015	1.380
Scenario 2: Y and W dependent; positive average treatment effect				
Unadjusted	0.00021	0.0023	0.0023	1.000
IPW	0.00020	0.0018	0.0018	1.230
Model standardization	0.00024	0.0018	0.0018	1.250
DR-WLS	0.000060	0.0018	0.0018	1.260
Tan	−0.00086	0.0018	0.0018	1.280
PLEASE	−0.00043	0.0018	0.0018	1.290
Rotnitzky $et\ al._{K=1}$	0.00015	0.0016	0.0016	1.400
Gruber and van der Laan	0.00039	0.0018	0.0018	1.250
Scenario 3: Y and W independent; zero average treatment effect				
Unadjusted	−0.000033	0.0021	0.0021	1.000
IPW	−0.000037	0.0021	0.0021	0.992
Model standardization	−0.000041	0.0021	0.0021	0.992
DR-WLS	−0.000037	0.0021	0.0021	0.992
Tan	−0.000042	0.0021	0.0021	0.986
PLEASE	−0.000055	0.0021	0.0021	0.979
Rotnitzky $et\ al._{K=1}$	−0.000038	0.0022	0.0022	0.972
Gruber and van der Laan	−0.0000015	0.0021	0.0021	0.991
Scenario 4: Y and W independent; positive average treatment effect				
Unadjusted	−0.00016	0.0020	0.0020	1.000
IPW	−0.00018	0.0020	0.0020	0.992
Model standardization	−0.00017	0.0020	0.0020	0.993
DR-WLS	−0.00019	0.0020	0.0020	0.992
Tan	−0.00017	0.0021	0.0021	0.987
PLEASE	−0.00016	0.0021	0.0021	0.980
Rotnitzky $et\ al._{K=1}$	−0.00020	0.0021	0.0021	0.972
Gruber and van der Laan	0.000082	0.0020	0.0020	0.991

Table III

Results of the 100,000 simulated randomized trials of size $n = 1044$ patients based on resampling from the PEARLS trial. For each of the seven adjusted estimators, relative efficiency is defined as the variance of the unadjusted estimator divided by the variance of the adjusted estimator.

Estimator	Bias	Variance	MSE	Rel.Efficiency
Scenario 1: Y and W dependent; zero average treatment effect				
Unadjusted	−0.000087	0.00055	0.00055	1.000
IPW	−0.000071	0.00054	0.00054	1.020
Model standardization	−0.000075	0.00054	0.00054	1.020
DR-WLS	−0.000075	0.00054	0.00054	1.020
Tan	−0.000079	0.00054	0.00054	1.030
PLEASE	−0.000068	0.00054	0.00054	1.020
Rotnitzky $et\ al._{K=1}$	−0.000078	0.00054	0.00054	1.030
Gruber and van der Laan	−0.000080	0.00054	0.00054	1.020
Scenario 2: Y and W dependent; positive average treatment effect				
Unadjusted	0.000052	0.00062	0.00062	1.000
IPW	0.000044	0.00061	0.00061	1.020
Model standardization	0.000040	0.00061	0.00061	1.020
DR-WLS	0.000048	0.00061	0.00061	1.020
Tan	0.000053	0.00061	0.00061	1.020
PLEASE	0.000070	0.00061	0.00061	1.020
Rotnitzky $et\ al._{K=1}$	0.000054	0.00061	0.00061	1.020
Gruber and van der Laan	0.000096	0.00061	0.00061	1.020
Scenario 3: Y and W independent; zero average treatment effect				
Unadjusted	0.00011	0.00056	0.00056	1.000
IPW	0.00010	0.00056	0.00056	0.997
Model standardization	0.00010	0.00056	0.00056	0.997
DR-WLS	0.00010	0.00056	0.00056	0.997
Tan	0.00011	0.00056	0.00056	0.995
PLEASE	0.000091	0.00056	0.00056	0.992
Rotnitzky $et\ al._{K=1}$	0.00010	0.00056	0.00056	0.991
Gruber and van der Laan	0.00013	0.00056	0.00056	0.998
Scenario 4: Y and W independent; positive average treatment effect				
Unadjusted	0.000011	0.00055	0.00055	1.000
IPW	0.0000034	0.00056	0.00056	0.997
Model standardization	0.0000023	0.00056	0.00056	0.997
DR-WLS	0.0000039	0.00056	0.00056	0.997
Tan	0.0000044	0.00056	0.00056	0.995
PLEASE	0.000018	0.00056	0.00056	0.993
Rotnitzky $et\ al._{K=1}$	0.000011	0.00056	0.00056	0.991
Gruber and van der Laan	0.000025	0.00055	0.00055	1.000