Contemporary Clinical Trials 54 (2017) 18–24

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Improving precision by adjusting for prognostic baseline variables in randomized trials with binary outcomes, without regression model assumptions

Jon Arni Steingrimsson[a], Daniel F. Hanley[b], Michael Rosenblum[a, c, *]

[a] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, United States
[b] Department of Neurology, Brain Injury Outcomes Coordinating Center, Johns Hopkins University, Baltimore, MD 21231, United States
[c] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, United States

ARTICLE INFO

Article history:
Received 9 August 2016
Received in revised form 21 December 2016
Accepted 31 December 2016
Available online 4 January 2017

Keywords:
Covariate adjustment
Post-stratification

ABSTRACT

In randomized clinical trials with baseline variables that are prognostic for the primary outcome, there is potential to improve precision and reduce sample size by appropriately adjusting for these variables. A major challenge is that there are multiple statistical methods to adjust for baseline variables, but little guidance on which is best to use in a given context. The choice of method can have important consequences. For example, one commonly used method leads to uninterpretable estimates if there is any treatment effect heterogeneity, which would jeopardize the validity of trial conclusions. We give practical guidance on how to avoid this problem, while retaining the advantages of covariate adjustment. This can be achieved by using simple (but less well-known) standardization methods from the recent statistics literature. We discuss these methods and give software in R and Stata implementing them. A data example from a recent stroke trial is used to illustrate these methods.

1. Introduction

In a recent regulatory guideline on the analysis of clinical trials, the European Medicines Agency states "in case of a strong or moderate association between a baseline covariate(s) and the primary outcome measure, adjustment for such covariate(s) generally improves the efficiency of the analysis and avoids conditional bias from chance covariate imbalance" [1]. Such covariate adjustment is not uncommon; Pocock et al. [2], who surveyed 50 clinical trial reports from major medical journals, found that 36 used some form of adjustment for baseline variables. However, they concluded that "the statistical properties of covariate-adjustment are quite complex and often poorly understood, and there remains confusion as to what is an appropriate statistical strategy." A more recent survey reached a similar conclusion [3]. We attempt to resolve some of this confusion by addressing two important misconceptions about covariate adjustment for binary outcomes.

The first misconception involves the logistic coefficient estimator, defined as the estimated coefficient on the treatment term in a main effects logistic regression of the outcome on treatment and baseline variables. This is a commonly used method for covariate adjustment when outcomes are binary [4–8]. An underappreciated vulnerability of the logistic coefficient estimator is that it is uninterpretable unless one assumes the conditional treatment effect (on the log odds scale) is precisely the same value for every possible stratum of the covariates adjusted for. In contrast, there are covariate adjusted estimators that don't have this drawback, described below.

A second misconception about binary outcomes is that covariate adjustment involving logistic regression models cannot be used to estimate the marginal risk difference or relative risk. Austin et al. [3] state "For binary outcomes, risk differences and relative risks (assuming a uniform relative risk) are collapsible estimators. However, their use precludes the use of regression adjustment". This claim is correct in so far as "regression adjustment" refers only to the logistic coefficient estimator, which is the setting of their paper. However, logistic regression models can (and we argue often should) be used to construct covariate adjusted estimators of the marginal risk difference or relative risk, by using the standardized estimator developed by Moore and van der Laan [9] as described below.

Moore and van der Laan proved theoretical properties of covariate adjusted estimators for marginal effects based on the general theory of targeted maximum likelihood estimation [9]. Unlike the logistic

[*] Corresponding author.
E-mail addresses: jsteing5@jhu.edu (J. Steingrimsson), dhanley@jhmi.edu (D. Hanley), mrosen@jhu.edu (M. Rosenblum).

http://dx.doi.org/10.1016/j.cct.2016.12.026
1551-7144/© 2016 Published by Elsevier Inc.

coefficient estimator, these estimators are interpretable without requiring any parametric model assumptions. They can also have greater precision compared to the unadjusted estimator (which ignores baseline variables). Despite their appealing properties, these estimators are rarely used in clinical trials. In order to make these estimation techniques accessible to a wider audience, we describe in a non-technical manner the simplest of these estimators (called the standardized estimator). We then compare the properties of the standardized estimator to the logistic coefficient and the unadjusted estimator. We compare the performance of these estimators in a real data example from a completed stroke trial, as well as in simulations. We then make practical recommendations and provide software to calculate the standardized estimator.

2. Description of estimators

We consider trials where the primary outcome Y is binary, representing success ($Y = 1$) or failure ($Y = 0$). For simplicity, the trial is assumed to have two study arms: treatment and control. Study arm assignment uses simple or block randomization independent of the baseline variables. The study arm A is an indicator of being assigned to the treatment arm ($A = 1$) or control arm ($A = 0$). We adhere to the intention-to-treat principle, that is, we consider the effect of assignment to the treatment or control arm. The vector of baseline variables, denoted by W, can be any mix of continuous, binary, ordinal, and categorical variables. The number of baseline variables should be small relative to the sample size, as discussed in Section 5.

The average treatment effect involves two proportions: the proportion of the target population who would have a successful outcome if all were assigned to treatment and the same quantity if all were assigned to control. The average treatment effect is a contrast between these two population proportions, such as their difference (called the risk difference), their ratio (called the relative risk) or their log odds ratio. The unadjusted estimator of the average treatment effect involves replacing the population proportions by sample proportions using data from the completed trial.

When discussing estimators that involve a logistic regression model, we focus on the commonly used case where this model includes an intercept and main terms for study arm and baseline variables (and no interaction terms). We discuss the implications of including interactions in Section 6.

No assumptions are made on the relationships among Y, A, W, except that study arm A is assigned independent of the baseline variables W (which holds by randomization). In particular, we do not assume that a logistic regression model correctly captures the relationships among these variables. We assume each participant i in the trial contributes data vector (W_i, A_i, Y_i), which is an independent, identically distributed draw from the unknown, joint distribution on (W, A, Y).

2.1. Logistic coefficient estimator

The logistic coefficient estimator is defined as the fitted coefficient on A from a logistic regression model with intercept and main terms for A and each component of W. It estimates the conditional treatment effect within strata of baseline variables, on the log odds scale (under the assumption that the logistic regression model is correct). The logistic coefficient estimator and the unadjusted estimator do not estimate the same quantity. The former estimates a conditional effect, while the latter estimates an unconditional (also called marginal, or average) effect. Conditional and unconditional effects can substantially differ, both in value and in interpretation, as emphasized by Freedman in the article “Randomization Does Not Justify Logistic Regression” [10].

Which type of effect is preferred depends on the study objective. Diggle et al. [11] recommend marginal effects when the aim of the study is to make population based inference (which is our focus) and the conditional effect when interest lies in modeling participant-specific effects. Knowing the true conditional effect would give a more fine-grained understanding of treatment effects and heterogeneity, compared to the marginal effect. But, estimating the conditional effect typically requires strong model assumptions. For example, the logistic coefficient estimator requires the logistic regression model to be correctly specified. A logistic regression model is misspecified when it does not correctly capture the relationship between the outcome and the treatment and baseline variables. If there is any treatment effect heterogeneity, i.e., if the conditional treatment effect varies depending on the baseline variables, then the conditional treatment effect cannot be represented by a single number. In such a case, the logistic regression model with main terms is guaranteed to be misspecified, and therefore the logistic coefficient estimator is uninterpretable.

The above discussion highlights an important vulnerability of the logistic coefficient estimator, i.e., it is only interpretable under the assumption that the treatment has the same effect (on the log odds scale) for every stratum of baseline covariates. This is a strong assumption since there is no a priori reason to believe that a treatment would lead to exactly the same benefit for every type of patient (regardless of, e.g., age and disease severity). The state of knowledge about an experimental treatment is typically quite limited before starting a trial (hence the reason for running the trial), making such an assumption hard to justify. Also, it is difficult to verify this assumption holds when the covariates include continuous or categorical variables with many levels, since then there are few or no participants in some strata.

Fig. 1 illustrates a case where the treatment effect differs within strata of a single, ordinal variable W, representing a disease severity score at baseline. Let $\text{logit}(x) = log(x/(1 - x))$. The first plot in Fig. 1 depicts $\text{logit}(P(Y = 1|A = 0, W))$, i.e., the log odds of the probability of obtaining a successful outcome when assigned to control, within strata of the baseline variable W. The second plot shows the analogous function under assignment to treatment, $\text{logit}(P(Y = 1|A = 1, W))$. The third plot shows the conditional treatment effect, i.e., the difference between the curves in the second and first plots: $\text{logit}(P(Y = 1|A = 1, W))-$ $\text{logit}(P(Y = 1|A = 0, W))$. The logistic coefficient estimator is only interpretable if the curve in the third plot is a horizontal line. The main terms logistic regression model not only assumes the conditional effect is constant, but also assumes that the conditional probabilities (on the log-odds scale) shown in the first two plots in Fig. 1 are straight lines with the same slope. (The curves in Fig. 1 are based on smoothing the data from our trial example in Section 3, using only National Institutes of Health Stroke Scale as the baseline variable W, as described in Section A.5 of the online supplement.)

Even if the conditional effect could be represented by a single number, there can be several additional reasons for model misspecification, such as missing interaction terms among the baseline variables or using the wrong form for the baseline variables, e.g. not log transforming when a log transformation is appropriate. In the case of continuous baseline variables there are infinitely many possible transformations to choose from. This makes it hard to determine the appropriate transformation in order to establish a linear relationship to the log odds of the probability of success. In confirmatory randomized trials, the model and variables used for covariate adjustment need to be prespecified in the study protocol, and hence can only be based on prior information such as scientific knowledge and earlier phase trial data [1]. This makes evaluating the correctness of the model difficult, with goodness of fit tests often having low power. As a consequence, the final logistic regression model is expected to be at least somewhat misspecified.

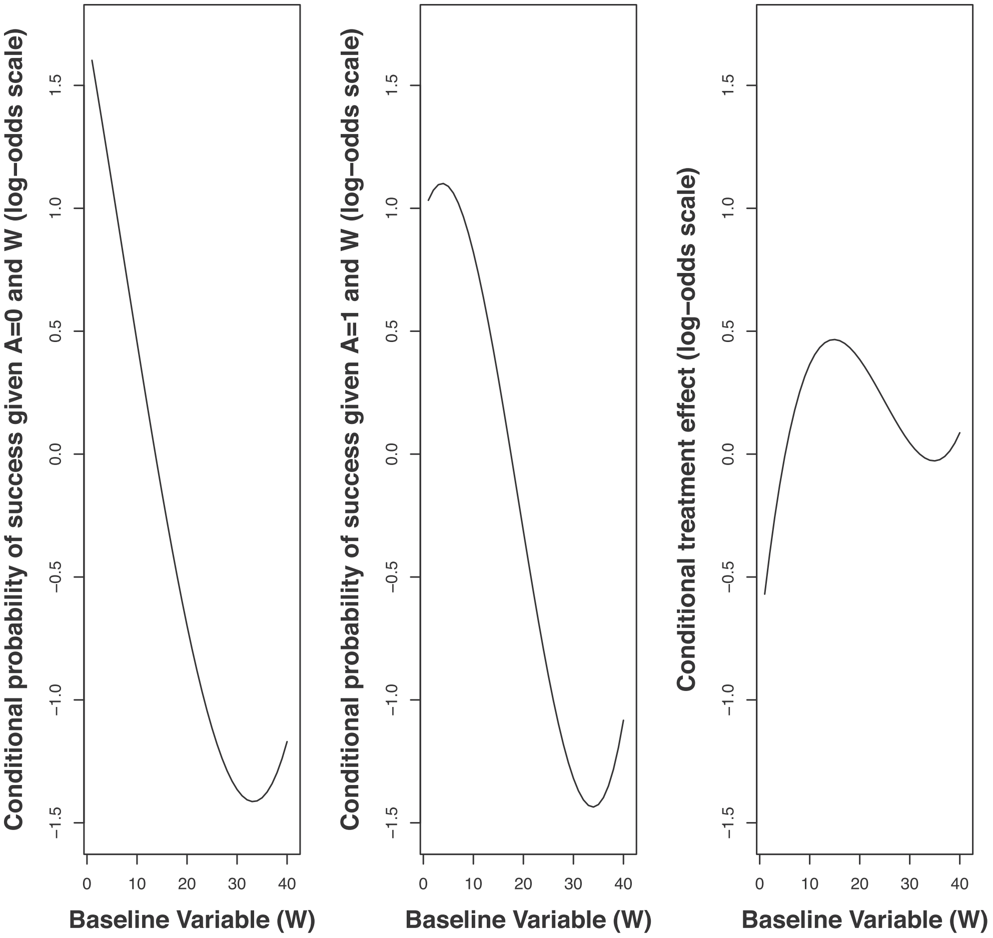

Fig. 1. An example of a conditional effect that depends on the value of the baseline variables and cannot be represented using a single number. The formulas for the quantities on the vertical axes are given in the main text.

2.2. Standardized estimator

The standardized estimator of Moore and van der Laan [9] estimates the average treatment effect (the same effect estimated by the unadjusted estimator) defined as a contrast between the proportion of the target population who would have a successful outcome under treatment versus control. The standardized estimator first estimates each of these proportions; then any contrast such as the risk difference, relative risk, or log odds ratio can be estimated by plugging in these two proportions. For the risk difference, the standardized estimator is a special case of the method of Scharfstein et al. [12] applied to randomized trials.

To estimate the above proportions using the standardized estimator, one first fits a main effects logistic regression model. Using the model fit, the predicted probability of a successful outcome under treatment is calculated for each participant i (regardless of actual study arm) by setting the study arm variable to $A = 1$ and using that participant's baseline variables W_i. Similarly, a predicted probability under control is calculated for every participant setting $A = 0$. This gives two predictions for each participant in the dataset, one corresponding to assignment to treatment and the other to control. The final estimator for the population proportion who would have a successful outcome under assignment to treatment (control) is the average of the predictions with the study arm variable set to treatment (control); each average is taken over the entire set of study participants (pooling both arms). We refer to these estimated proportions as estimated success probabilities. Intuitively, the purpose of the above steps is to correct for chance imbalances in baseline variables between treatment and control arms, e.g., more high disease severity patients assigned to one of the arms by chance.

Both the standardized and the logistic coefficient estimators require specifying a logistic regression model. The standardized estimator is a consistent estimator of the average treatment effect even if the model is arbitrarily misspecified, as proved by Moore and van der Laan [9]. Hence, the standardized estimator, unlike the logistic coefficient estimator, does not rely on any regression model assumptions in order to be consistent.

Table 1 summarizes the key properties of the unadjusted, standardized, and logistic coefficient estimators. A more technical description of the estimators is given in the online supplement accompanied by R and Stata code for computing the standardized estimator along with a corresponding confidence interval for the

Table 1
Properties of the unadjusted, standardized, and logistic coefficient estimators.

Estimator	Effect it estimates	Requires regression model assumptions?	Adjusts for baseline variables?
Unadjusted	Marginal effect	No	No
Standardized	Marginal effect	No	Yes
Logistic coefficient	Conditional effect	Yes	Yes

marginal risk difference. Precision gains from covariate adjustment translate into shorter confidence intervals compared to the unadjusted estimator (asymptotically).

3. CLEAR III trial application

The Clot Lysis Evaluation of Accelerated Resolution of Intraventricular Hemorrhage (CLEAR III) trial [13] is a completed phase III randomized trial comparing removal of intraventricular hemorrhage (IVH) using a low dose of recombinant tissue plasminogen activator versus standard of care. The primary outcome is a measure of disability evaluated using a dichotomized modified Rankin scale (mRS) ($\leq$3 vs $>$ 3) at 180 days. The study protocol has been described in detail elsewhere [14]. We use data from the 491 uncensored participants, out of 500 enrolled.

Phase II data and prior scientific knowledge indicated that baseline (pre-randomization) age, intracerebral hemorrhage volume, IVH volume, Glasgow coma scale, and the National Institutes of Health Stroke Scale (NIHSS) are prognostic baseline variables; we let W denote these variables, which are used by the adjusted estimators. Using the CLEAR III trial data, we calculated the approximate prognostic value of these variables using a modified R^2 computation [15] described in our online supplement; the result was $R^2 = 0.35$. This indicates that the baseline variables are moderately to strongly prognostic in the CLEAR III trial. This information would not be available when planning a trial and was not used to select the baseline variables. It is presented to show an example where efficiency gains may be expected using estimators that adjust for baseline variables.

The logistic coefficient estimator is only interpretable if the conditional treatment effect is the same for all combinations of baseline age, intracerebral hemorrhage volume, IVH volume, Glascow coma scale, and NIHSS. In contrast, the unadjusted estimator and the standardized estimators do not require any such assumption to be consistent for the average treatment effect.

Table 2 shows point estimates and 95% confidence intervals based on each of the three estimators, applied to the CLEAR III data set of 491 participants. The effects are estimated for both the primary mRS outcome and survival at 180 days (as a binary indicator), the latter being a secondary outcome. The unadjusted and standardized estimators target the marginal risk difference. The logistic regression estimator aims to estimate a conditional effect within strata of baseline variables on the log odds scale. Here and in our simulations, the 95% confidence interval based on each estimator is constructed using the nonparametric bootstrap with 1000 replicated data sets, as implemented by the R and Stata code in the online supplement.

For each estimator, the 95% confidence interval excludes 0 when the outcome is 180 day survival; the opposite holds when the outcome is 180 day mRS. The unadjusted estimator and the standardized estimators are similar. The confidence interval for the standardized estimator is 10% narrower for mRS and 17% narrower for 180 day survival, compared to the unadjusted estimator.

Table 2
Re-analysis of the CLEAR III trial using the unadjusted, standardized, and logistic coefficient estimators. The first two estimate the marginal risk difference, and the third aims to estimate the conditional log odds ratio. Each cell gives the corresponding point estimate followed by the 95% confidence interval. The middle column corresponds to the primary outcome being the indicator of 180 days modified Rankin score$\leq$3; the right column corresponds to the primary outcome being the indicator of 180 days survival.

	Outcome type:	
	Modified Rankin score $\leq$ 3	Survival at 180 days
Estimator:		
Unadjusted	0.03 (−0.07, 0.10)	0.11 (0.03,0.19)
Standardized	0.01 (−0.07, 0.08)	0.10 (0.03, 0.17)
Logistic coefficient	0.05 (−0.45, 0.55)	0.68 (0.19, 1.17)

The logistic coefficient estimator has wider confidence intervals than the other estimators, for each outcome; that remains true even if the unadjusted and standardized estimators are transformed to the log odds scale. However, it is difficult to make a direct comparison since the logistic coefficient estimator aims to estimate a conditional rather than unconditional treatment effect.

4. Simulations based on the CLEAR III trial data

4.1. Data generating distributions

We constructed data generating distributions for our simulations to mimic certain features of the CLEAR III data. This was done by resampling with replacement from the CLEAR III trial, and then making modifications described below. The reason we simulate from a trial rather than a parametric model is that we believe the former more accurately reflects complexities in real trial data distributions. All simulated trials have total sample size 491. The same variables are used as in the previous section. We simulate two different settings (distributions) adapted from Colantuoni and Rosenblum[15]:

- Setting 1. Baseline variables prognostic for the outcome.
- Setting 2. Baseline variables independent of the outcome.

Setting 1 is constructed to mimic the following features of the CLEAR III data: the correlation structure within the baseline variables, and the relationship between the baseline variables and the outcome. Setting 2 only mimics the former feature.

In both settings, the simulated distributions were constructed to have an average treatment effect of 13% on the risk difference scale. This choice was based on the sample size calculations for the CLEAR III trial described in Ziai et al. [14], which was powered to detect approximately this magnitude of average treatment effect. Modifications to the resampling distribution were made in order to achieve the 13% treatment effect; full details are given in the online supplement.

An important consideration when choosing the analysis technique for a randomized trial is the power of the corresponding hypothesis test. Each of the three estimators (the unadjusted, the logistic coefficient, and the standardized estimator) can be converted to a corresponding Wald statistic by dividing the estimator by its standard error. We focus on testing the null hypothesis of no average treatment effect, i.e., $H_0 : P(Y = 1|A = 1) - P(Y = 1|A = 0) = 0$. When the logistic regression model is correctly specified (which is required in order for the logistic coefficient estimator to be interpretable), this null hypothesis is equivalent to the coefficient on the treatment term in that model being equal to 0; in this case, the Wald statistic based on each estimator leads to a test of H_0 with asymptotically correct Type I error.

The one-sided test at level α is implemented by computing the Wald statistic for a given estimator, and rejecting the null hypothesis H_0 if the Wald statistic exceeds $\boldsymbol{\Phi}^{-1}(1-\alpha)$, where $\boldsymbol{\Phi}$ is the standard normal cumulative distribution function. Similarly, a two-sided test at level α involves replacing the rejection threshold $\boldsymbol{\Phi}^{-1}(1-\alpha)$ by $\boldsymbol{\Phi}^{-1}(1-\alpha/2)$, and using the absolute value of the Wald statistic; this test is equivalent to comparing the square of the Wald statistic to the $1-\alpha$ quantile of the chi-squared distribution with 1 degree of freedom. The power of the tests based on different estimators can differ, as described next.

The power of a Wald test is directly related to the mean of the Wald statistic, i.e., the estimator mean divided by its standard error; this quantity is called the signal to noise ratio of the estimator. Each of the three estimators has a different signal to noise ratio. The larger the signal to noise ratio of an estimator is, the larger the power of the corresponding Wald test is. Following the literature [5,7,16,17], we define the relative efficiency (RE) for testing H_0 using the Wald statistic based on one estimator compared to the Wald statistic based on a second estimator as the square of the following: the signal to noise ratio of the first estimator divided by the signal to noise ratio of the second. The formal justification for this definition of RE is given by van der Vaart [18].

Relative efficiency has a direct relationship to sample size savings. The relative reduction in the required sample size for the Wald statistic based on one estimator to achieve the same power for testing H_0 as another estimator is $1 - (1/\text{RE})$. We refer to this formula as the "reduction in sample size" (RSS) from using one estimator vs. another, to achieve the same power for a Wald test of H_0.

4.2. Simulation results

In simulation setting one, the baseline variables are prognostic for the outcome. Therefore, adjusted estimators have potential to leverage information in baseline variables to improve efficiency compared to the unadjusted estimator. In setting two, the outcome is independent of the baseline variables, i.e., the baseline variables are pure noise; this setting is used to get an idea of how much efficiency loss (if any) occurs when the baseline variables are not prognostic.

A summary of the results from 10,000 simulated trials is given in Table 3. The evaluation measures used are: value of the estimators averaged over the 10,000 simulations, the empirical standard error of the estimators, relative efficiency, and reduction in sample size; the last two measures are comparisons with the unadjusted estimator.

In setting one, the standardized estimator has smaller variance than the unadjusted estimator, with relative efficiency of 1.41. This corresponds to the standardized estimator requiring 29% smaller sample size than the unadjusted estimator to have the same power. In setting two, where the baseline variables are independent of the outcome, the standardized and unadjusted estimators have similar efficiency, with relative efficiency of 0.99. This corresponds to the standardized estimator requiring 1% larger sample size to achieve the same power as the unadjusted estimator. The simulations show that both the standardized and unadjusted estimators are approximately unbiased for the marginal treatment effect in both settings.

Table 3 also shows the performance of the logistic coefficient estimator. The logistic regression model is likely not correct in setting 1. (In setting 2, where the outcome is generated independent of baseline variables, the logistic model is correct, as discussed in the online supplement.) Therefore, in setting 1 the logistic coefficient estimator is uninterpretable. Even if it were interpretable, the efficiency gain from adjustment (RE) is slightly less for the logistic coefficient estimator compared to the standardized estimator; the same is true even if all estimators are converted to the log odds scale, as shown in the online supplement. (Also, the efficiency gain of the logistic coefficient estimator compared to the unadjusted estimator is similar to that seen in other stroke trials [7,19].) In setting 2, both adjusted estimators lose efficiency compared to the unadjusted, but the loss is slightly worse for the logistic coefficient estimator.

The efficiency gains from the logistic coefficient estimator compared to the unadjusted estimator are primarily a consequence of the treatment effect being further away from the null, rather than a reduction in estimator variance [16]. This is different from the standardized estimator, where the treatment effect being estimated (the average treatment effect) is the same as for the unadjusted estimator, and the efficiency gains are purely a consequence of variance reduction.

5. Recommendations for practice

For reasons described in Section 2 and illustrated using the CLEAR III trial data in Section 3, we recommend to use the standardized estimator combined with bootstrapped confidence intervals, when it is expected that baseline variables will be moderately to strongly prognostic for the outcome.

When the outcome is always observed but a substantial proportion of the baseline variables have missing data, the unadjusted estimator is preferred over the model standardization estimator (since adjusting for missing data requires making additional assumptions and can result in less precise estimators).

In stroke trials, baseline stroke severity as measured by NIHSS is a relatively strong predictor for the outcome, and can be used for covariate adjustment. Ideally, prognostic baseline variables should be selected based on clinical understanding, and then evaluated using prior data sets. E.g., for a future phase III trial being planned, phase II data can be used to evaluate the prognostic value of the baseline variables. Colantuoni and Rosenblum [15] propose a modified R^2 method (as used earlier in our paper) for doing so. A possible rule of thumb is to build the standardized estimator into the phase III study protocol as the primary analysis if the modified R^2 (which approximates the reduction in sample size due to adjustment) is at least 10%, based on a completed phase II randomized trial with at least 100 participants. We caution that since the population enrolled in phase III may differ from phase II, there is no guarantee that relative efficiency from the latter will be similar to the former.

We emphasize that the statistical analysis plan in a phase III trial must be prespecified. If this includes using a covariate adjusted estimator, the precise details of the estimator need to be prespecified (including the type of estimator and the corresponding model and variables to be used). A conservative approach is to select just a few variables that are thought to be prognostic for the outcome based on medical knowledge and on prior data as described above.

The relative efficiency gains resulting from the use of the standardized estimator are expected to be similar in large and moderately sized trials; this holds in general for covariate adjusted estimators, as noted by Pocock et al. [2]. This makes the potential absolute reduction in sample size from covariate adjustment greater in larger trials.

Table 3
Average value of estimator over the 10,000 simulations, empirical standard error, relative efficiency (RE) compared to unadjusted estimator, and reduction in sample size (RSS) compared to the unadjusted estimator. The unadjusted and standardized estimator estimate the marginal risk difference while the logistic coefficient aims to estimate a conditional effect on the log-odds scale. For both the unadjusted and standardized estimator, the true marginal treatment effect is 0.13 in both settings. In setting two, the true conditional treatment effect on the log odds scale is 0.52. As the logistic regression model is not necessarily correct in setting one, it is unclear if the true conditional effect is interpretable as a single number.

	Estimator	Average value of estimator	Empirical standard error	RE	RSS
Setting 1	Unadjusted	0.13	4.5×10^{-2}	1	0
	Standardized	0.13	3.8×10^{-2}	1.41	29%
	Logistic coefficient	0.76	0.23	1.31	24%
Setting 2	Unadjusted	0.13	4.5×10^{-2}	1	0
	Standardized	0.13	4.5×10^{-2}	0.99	−1%
	Logistic coefficient	0.53	0.19	0.94	−7%

The ratio of the standardized estimator to its standard error can be used as a test statistic for the null hypothesis of no average treatment effect. Precision gains of the standardized estimator compared to the unadjusted estimator lead to this test having higher asymptotic power compared to the analogous test for the unadjusted estimator. We emphasize that a robust standard error method, e.g., the nonparametric bootstrap, must be used, for which code is given in the online supplement.

6. Discussion

The main benefits of the standardized estimator over the logistic coefficient estimator are that (i) the former does not require correct model specification in order to be consistent, and (ii) the former is a consistent estimator of the average treatment effect, which always has an interpretation as a single population value (the same as being estimated by the unadjusted estimator), unlike the conditional treatment effect (which may be a complex function rather than a single value).

Our results illustrate the potential advantages of using the standardized estimator when analysing data from randomized trials. They are not meant as a complete analysis of the CLEAR III trial. One important component not considered here is that an adaptive randomization scheme was used when randomizing participants to study arms in the CLEAR III trial. For simplicity and since many trials use simple or block randomization, we focused on this case. When covariate adaptive randomization is used, it is recommended to adjust for the covariates in the analysis [1].

Consider the case when the sampling distribution for the trial (i.e., the data generating distribution for the trial participants) differs from the distribution in the target population. For example, it may be that the patients with very high disease severity are more likely to enroll in the trial than those with low disease severity. In that case, if the target population consists of those with both high and low disease severity, all three estimators can suffer from bias for estimating the treatment effect in the target population. The standardized and logistic coefficient estimators require the sampling distribution of the covariates and outcome to match that of the target population, in order to be consistent. The unadjusted estimator only requires the outcome distribution to match that of the target population in order to be consistent.

Throughout, we considered logistic regression models with main terms only. This accords with the European Medicines Agency guideline on covariate adjustment, which recommends that the primary analysis should not include treatment by baseline variable interactions [1]. The standardized estimator is guaranteed to be consistent for the marginal treatment effect whether the true population distribution involves interactions or not.

Stratified block randomization can be used to balance, by design, the levels of baseline variables across study arms. However, such a design can only be used to balance a small number of strata. The standardized estimator can incorporate multiple variables (continuous, ordinal, and/or categorical), and can therefore potentially leverage more of the prognostic information compared to using stratified block randomization on a small number of strata. Alternatively, the standardized estimator can be used in conjunction with this randomization scheme to leverage additional prognostic information in variables that were not stratified on by design.

Sample size calculations for the standardized estimator require specifying how prognostic the baseline variables are for the outcome. A conservative approach is to calculate the sample size as if there would be no precision gain compared to an unadjusted estimator, but plan to use the standardized estimator in the primary analysis. When the baseline variables are prognostic, this would result in a study with higher power than originally intended. The potential reduction in variance can result in smaller expected sample sizes associated with the standardized estimator in group sequential trials with information monitoring [20].

There are several covariate adjusted estimators for binary outcomes that share the desirable properties of the standardized estimator [15]. For simplicity, we only focused on the standardized estimator. If some participants have missing outcomes and the missingness probability can be correctly modeled, the standardized estimator can adjust for the resulting bias by adding weights to the logistic regression models [15]. The standardized estimator can also adjust for baseline variables when responder analysis is used to define the outcomes [21].

There are several other settings where estimators that share the desirable properties of the standardized estimator are available. Estimators with similar qualities have been derived for other generalized linear models such as when the outcome is continuous, ordinal, or a count measure [22,23]. For longitudinal studies, such as if mRS was measured at 30, 90, and 180 days, the targeted maximum likelihood estimator [24] can be used. For group sequential designs, precision can be improved by using the standardized estimator at every analysis, as long as covariances are computed using a robust method such as the nonparametric bootstrap.

Acknowledgments

This work was supported by the Patient-Centered Outcomes Research Institute [ME-1306-03198], the U.S. Food and Drug Administration [HHSF223201400113C]. The CLEAR III trial was supported by the grant 5U01 NS062851-05, awarded to DFH from the National Institutes of Health (NIH), National Institute of Neurological Disorders and Stroke (NINDS); this study is registered with ClinicalTrials.gov, NCT00222573. This work is solely the responsibility of the authors and does not represent the views of the above people and agencies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.cct.2016.12.026.

References

[1] Guideline on adjustment for baseline covariates in clinical trials, http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf, accessed: 2016-05-18.
[2] S.J. Pocock, S.E. Assmann, L.E. Enos, L.E. Kasten, Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems, Stat. Med. 21 (19) (2002) 2917–2930.
[3] P.C. Austin, A. Manca, M. Zwarenstein, D.N. Juurlink, M.B. Stanbrook, A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals, J. Clin. Epidemiol. 63 (2) (2010) 142–153.
[4] S.C. Choi, Sample size in clinical trials with dichotomous endpoints: use of covariables, J. Biopharm. Stat. 8 (3) (1998) 367–375.
[5] A.V. Hernández, E.W. Steyerberg, J.D.F. Habbema, Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements, J. Clin. Epidemiol. 57 (5) (2004) 454–460.
[6] A.V. Hernández, E.W. Steyerberg, G.S. Taylor, A. Marmarou, J.D.F. Habbema, A.I. Maas, Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review, Neurosurgery 57 (6) (2005) 1244–1253.
[7] A.V. Hernández, E.W. Steyerberg, I. Butcher, N. Mushkudiani, G.S. Taylor, G.D. Murray, A. Marmarou, S.C. Choi, J. Lu, J.D.F. Habbema, et al. Adjustment for strong predictors of outcome in traumatic brain injury trials: 25% reduction in sample size requirements in the IMPACT study, J. Neurotrauma 23 (9) (2006) 1295–1303.
[8] E.W. Steyerberg, P.M.M. Bossuyt, K.L. Lee, Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am. Heart J. 139 (5) (2000) 745–751.
[9] K.L. Moore, M.J. van der Laan, Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation, Stat. Med. 28 (1) (2009) 39.

[10] D.A. Freedman, Randomization does not justify logistic regression, Stat. Sci. 23 (2) (2008) 237–249.

[11] P. Diggle, P. Heagerty, K.-Y. Liang, S. Zeger, Analysis of Longitudinal Data, OUP Oxford. 2013.

[12] D.O. Scharfstein, A. Rotnitzky, J.M. Robins, Rejoinder to “Adjusting for nonignorable drop-out using semiparametric nonresponse models”, J. Am. Stat. Assoc. 94 (448) (1999) 1096–1120.

[13] D.F. Hanley, K. Lane, N. McBee, W. Ziai, S. Tuhrim, K.R. Lees, J. Dawson, D. Gandhi, N. Ullman, W.A. Mould, et al. Thrombolytic removal of intraventricular haemorrhage in treating severe stroke: results of the CLEAR III trial, a randomised, controlled trial, Lancet (2016)

[14] W.C. Ziai, S. Tuhrim, K. Lane, N. McBee, K. Lees, J. Dawson, K. Butcher, P. Vespa, D.W. Wright, P.M. Keyl, et al. A multicenter, randomized, double-blinded, placebo-controlled phase III study of Clot Lysis Evaluation of Accelerated Resolution of Intraventricular Hemorrhage (CLEAR III), Int. J. Stroke 9 (4) (2014) 536–542.

[15] E. Colantuoni, M. Rosenblum, Leveraging prognostic baseline variables to gain precision in randomized trials, Stat. Med. 34 (18) (2015) 2602–2617.

[16] L.D. Robinson, N.P. Jewell, Some surprising results about covariate adjustment in logistic regression models, Int. Stat. Rev./Revue Internationale de Statistique (1991) 227–240.

[17] E.L. Turner, P. Perel, T. Clayton, P. Edwards, A.V. Hernández, I. Roberts, H. Shakur, E.W. Steyerberg, C.T. Collaborators, et al. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury, J. Clin. Epidemiol. 65 (5) (2012) 474–481.

[18] A. van der Vaart, Asymptotic Statistics, Cambridge University Press, Cambridge, 1998.

[19] L.J. Gray, P. Bath, T. Collier, Should stroke trials adjust functional outcome for baseline prognostic factors? Stroke 40 (3) (2009) 888–894.

[20] C. Jennison, B.W. Turnbull, Group Sequential Methods with Applications to Clinical Trials, CRC Press. 1999.

[21] K.M. Garofolo, S.D. Yeatts, V. Ramakrishnan, E.C. Jauch, K.C. Johnston, V.L. Durkalski, The effect of covariate adjustment for baseline severity in acute stroke clinical trials with responder analysis outcomes, Trials 14 (1) (2013) 98.

[22] M. Rosenblum, M.J. van der Laan, Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables, Int. J. Biostat. 6 (1). (2010)

[23] I. Díaz, E. Colantuoni, M. Rosenblum, Enhanced precision in the analysis of randomized trials with ordinal outcomes, Biometrics (2015)

[24] M.J. Van der Laan, S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Springer Science & Business Media. 2011.