Johns Hopkins University, Dept. of Biostatistics Working Papers

8-24-2016

IMPROVING PRECISION BY ADJUSTING FOR BASELINE VARIABLES IN RANDOMIZED TRIALS WITH BINARY OUTCOMES, WITHOUT REGRESSION MODEL ASSUMPTIONS

Jon Arni Steingrimsson
Johns Hopkins Bloomberg School of Public Health

Daniel F. Hanley
Department of Neurology, Johns Hopkins University

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, mrosenbl@jhsph.edu

Improving Precision by Adjusting for Prognostic Baseline Variables in Randomized Trials with Binary Outcomes, without Regression Model Assumptions

Jon Arni Steingrimsson

Department of Biostatistics

Johns Hopkins University, Baltimore MD 21205

jsteing5@jhu.edu

Daniel F. Hanley

Department of Neurology

Johns Hopkins University, Baltimore MD 21205

dhanley@jhmi.edu

Michael Rosenblum *

Department of Biostatistics

Johns Hopkins University, Baltimore MD 21205

mrosen@jhu.edu

*To whom correspondence should be addressed.

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.

Keywords: Covariate Adjustment, Post-Stratification

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. We then make practical recommendations.

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 important and 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. 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, disease severity, etc.). 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. The impact is that the logistic coefficient estimator is uninterpretable without making the strong assumption that the treatment has the same effect (on the log odds scale) for every stratum of covariates. 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 should) be used to construct covariate adjusted estimators of the marginal risk difference or relative risk, by using the standardized estimator of Moore and van der Laan[9] as described below.

Our main practical recommendation is that in trials with binary outcomes and prognostic baseline variables, the standardized estimator of the marginal risk difference or relative risk is a good method to use for covariate adjustment. Unlike the logistic coefficient estimator, the standardized estimator is consistent without requiring any parametric model assumptions, so is not vulnerable to being uninterpretable as described above. The advantage of this covariate adjusted estimator compared to the unadjusted estimator (which ignores baseline variables) is that it can have greater precision.

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 6.

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.

Figure 1 illustrates such a case of treatment effect heterogeneity, where W is a single, ordinal variable, representing a disease severity score at baseline. Let $\text{logit}(x) = \log(x/(1-x))$. The first plot in Figure 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 shown in the first two plots in Figure 1 are constant. (The curves in Figure 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.4 of the online supplement.)

Even if the conditional effect could be represented by a single number, there can be several additional reasons for model mispecification, such as missing interaction terms between 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.

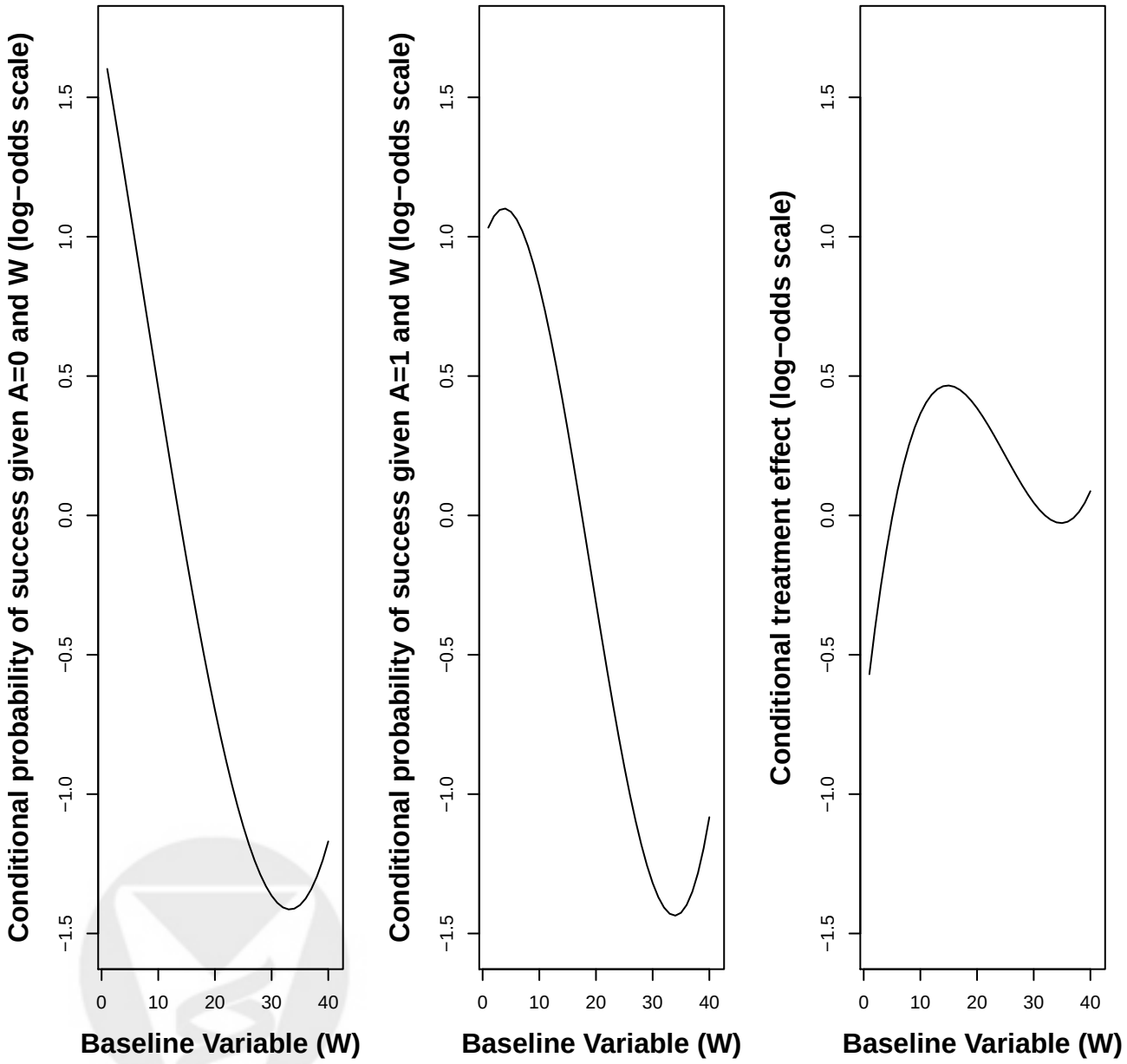

Figure 1: An example of a conditional effect that depends on the value of the baseline variables and cannot be represented using a single number. The formulas for the quantities on the vertical axes are given in the main text.

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). 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 marginal risk difference. Precision gains from covariate adjustment

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

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

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 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) (≤ 3 vs > 3) at 180 days. The study protocol has been described in detail elsewhere.[13] 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[14] 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 that 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 esti-

	Outcome Type:	
	Modified Rankin Score ≤ 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)

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 day modified Rankin score ≤ 3; the right column corresponds to the primary outcome being the indicator of 180 day survival.

mators, 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.

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:[14]

- 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.,[13] 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.

Comparisons between estimators are often in terms of signal to noise ratios. An estimator's signal to noise ratio is the quantity it converges to divided by the standard error. For example, since the unadjusted and standardized estimators both converge to the average treatment effect, this is the numerator in the signal to noise ratio for each; their denominators will generally differ. For the logistic coefficient estimator, the numerator is defined to be the probability limit of the estimated coefficient on the study arm variable; when the logistic model is correctly specified, the numerator is the conditional log odds ratio. Following the literature,[5,7,15] we define the relative

	Estimator	Average Value of Estimator	Empirical Standard Error	RE	RSS
Setting 1	Unadjusted	0.13	4.5×10^{-2}	1	0
	Standardized	0.13	3.7×10^{-2}	1.40	29%
	Logistic Coefficient	0.74	0.22	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	0.54	0.19	0.94	−7%

Table 3: Average value of estimator over the 10000 simulations, empirical standard error, relative efficiency (RE) compared to unadjusted estimator, and reduction in sample size (RSS) compared to the unadjusted estimator. For both the unadjusted and standardized estimator the true marginal treatment effect is 0.13 in both settings. In setting two the true log odds effect is 0.54. 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.

efficiency (RE) of one estimator compared to 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 practical importance of relative efficiency is its direct relationship to sample size savings. The Wald statistic corresponding to an estimator is the estimator value divided its standard error. The relative reduction in the required sample size for the Wald statistic based on one estimator to achieve the same power 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.

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 10000 simulated trials is given in Table 3. The evaluation measures used are: value of the estimators averaged over the 10000 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 1.40. 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 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 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. (The efficiency gain of the logistic coefficient estimator compared to the unadjusted estimator is similar to that seen in other stroke trials.[7,16]) In setting 2, both adjusted estimators lose efficiency compared to the unadjusted, but the loss is 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.[17] 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 untestable assumptions and can result in less precise estimators).

In stroke trials, prognostic baseline variables are often available as baseline severity is commonly a strong predictor for the outcome. 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[14] 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.

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]

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] This causes no problems for the standardized estimator, which 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.[18]

There are several covariate adjusted estimators for binary outcomes that share the desirable properties of the standardized estimator.[14] 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.[14] The standardized estimator can also adjust for baseline variables when responder analysis is used to define the outcomes.[19]

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.[20,21] For longitudinal studies, such as if mRS was measured at 30, 90, and 180 days, the targeted maximum likelihood estimator[22] 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.

7 Acknowledgements

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 Funded by the National Institute of Neurological Disorders and Stroke; ClinicalTrials.gov NCT00784134. This work is solely the responsibility of the authors and does not represent the views of the above people and agencies.

References

[1] Guideline on adjustment for baseline covariates in clinical trials. www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf. Accessed: 2016-05-18.

[2] Stuart J Pocock, Susan E Assmann, Laura E Enos, and Linda E Kasten. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in medicine*, 21(19):2917–2930, 2002.

[3] Peter C Austin, Andrea Manca, Merrick Zwarenstein, David N Juurlink, and Matthew 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. *Journal of Clinical Epidemiology*, 63(2):142–153, 2010.

[4] Sung C Choi. Sample size in clinical trials with dichotomous endpoints: use of covariables. *Journal of Biopharmaceutical Statistics*, 8(3):367–375, 1998.

[5] Adrián V Hernández, Ewout W Steyerberg, and J Dik F Habbema. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *Journal of clinical epidemiology*, 57(5):454–460, 2004.

[6] Adrían V Hernández, Ewout W Steyerberg, Gillian S Taylor, Anthony Marmarou, J Dik F Habbema, and Andrew IR Maas. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review. *Neurosurgery*, 57(6):1244–1253, 2005.

[7] Adrián V Hernández, Ewout W Steyerberg, Isabella Butcher, Nino Mushkudiani, Gillian S Taylor, Gordon D Murray, Anthony Marmarou, Sung C Choi, Juan Lu, J Dik 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. *Journal of Neurotrauma*, 23(9):1295–1303, 2006.

[8] Ewout W Steyerberg, Patrick MM Bossuyt, and Kerry L Lee. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *American heart journal*, 139(5):745–751, 2000.

[9] Kelly L Moore and Mark J van der Laan. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine*, 28(1):39, 2009.

[10] David A Freedman. Randomization does not justify logistic regression. *Statistical Science*, 23(2):237–249, 2008.

[11] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of Longitudinal Data*. OUP Oxford, 2013.

[12] Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Rejoinder to "adjusting for nonignorable drop-out using semiparametric nonresponse models". *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.

[13] Wendy C Ziai, Stanley Tuhrim, Karen Lane, Nichol McBee, Kennedy Lees, Jesse Dawson, Kenneth Butcher, Paul Vespa, David W Wright, Penelope 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). *International Journal of Stroke*, 9(4):536–542, 2014.

[14] Elizabeth Colantuoni and Michael Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in medicine*, 34(18):2602–2617, 2015.

[15] Elizabeth L Turner, Pablo Perel, Tim Clayton, Phil Edwards, Adrian V Hernández, Ian Roberts, Haleema Shakur, Ewout W Steyerberg, CRASH Trial Collaborators, et al. Covariate adjustment increased power in randomized controlled trials: an example in traumatic brain injury. *Journal of Clinical Epidemiology*, 65(5):474–481, 2012.

[16] LJ Gray, P Bath, and T Collier. Should stroke trials adjust functional outcome for baseline prognostic factors? *Stroke*, 40(3):888–894, 2009.

[17] Laurence D Robinson and Nicholas P Jewell. Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review/Revue Internationale de Statistique*, pages 227–240, 1991.

[18] Christopher Jennison and Bruce W Turnbull. *Group sequential methods with applications to clinical trials*. CRC Press, 1999.

[19] Kyra M Garofolo, Sharon D Yeatts, Viswanathan Ramakrishnan, Edward C Jauch, Karen C Johnston, and Valerie L Durkalski. The effect of covariate adjustment for baseline severity in acute stroke clinical trials with responder analysis outcomes. *Trials*, 14(1):98, 2013.

[20] Michael Rosenblum and Mark J van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1), 2010.

[21] Iván Díaz, Elizabeth Colantuoni, and Michael Rosenblum. Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 2015.

[22] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[23] Karen C Johnston, Alfred F Connors, Douglas P Wagner, and E Clarke Haley. Risk adjustment effect on stroke clinical trials. *Stroke*, 35(2):e43–e45, 2004.

[24] James Robins. A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512, 1986.

[25] Chris S Weaver, Jo Leonardi-Bee, Fiona J Bath-Hextall, and Philip MW Bath. Sample size calculations in acute stroke trials: a systematic review of their reporting, characteristics, and relationship with outcome. *Stroke*, 35(5):1216–1224, 2004.

[26] Li Yang and Anastasios A Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

[27] Nazmus Saquib, Juliann Saquib, and John PA Ioannidis. Practices and impact of primary outcome adjustment in randomized controlled trials: meta-epidemiologic study. *BMJ*, 347:f4313, 2013.

A Online Supplement.

In this online supplement we:

1. Give additional simulation results where the unadjusted and standardized estimator are transformed to the log odds scale.
2. Describe how to calculate the standardized estimator, the logistic coefficient estimator, the unadjusted estimator, and the modified R^2 estimator.
3. Give the curves based on raw data used to create Figure 1.
4. Give further details on how the simulation studies were conducted, and give R and Stata code for calculating the standardized estimator and variance estimator for the standardized estimator.

We start by defining some common notation. Let A be the study arm variable, W be a vector of the baseline variables, and Y be the outcome. In the context of the CLEAR III trial discussed in the main article, A is an indicator if the participant recieved IVH removal or not, Y is the dichotomized mRS at 180 days, and W is the vector of baseline variables (age, intracerebral hemorrhage volume, IVH volume, Glascow coma scale, NIHSS).

A.1 Simulation Results on the Log Odds Scale

Table 4 shows simulation results comparing the unadjusted, standardized, and logistic coefficient estimator where the unadjusted and standardized estimator are transformed to a log odds scale. The simulation setting used is described in Section A.3. In Setting two, where the baseline variables are independent of the outcome, the logistic regression model is correctly specified. This is because in that case $P(Y|A,W) = P(Y|A)$, and therefore the logistic regression model with all coefficients equal to 0 for baseline variable terms is correct.

A.2 Calculating the Estimators

Define $p_1 = P(Y = 1|A = 1)$ as the true but unknown proportion of successful outcomes in the treatment group and $p_0 = P(Y = 1|A = 0)$ as the true but unknown proportion of successful

	Estimator	Average Value of Estimator	Standard Error	RE	RSS
Setting 1	Unadjusted	0.51	0.18	1	0
	Standardized	0.52	0.16	1.41	29%
	Logistic Coefficient	0.74	0.22	1.36	26%
Setting 2	Unadjusted	0.53	0.18	1	0
	Standardized	0.53	0.19	0.99	−1%
	Logistic Coefficient	0.54	0.19	0.97	−3%

Table 4: Average treatment effect, average standard error , relative efficiency (RE), and reduction in sample size (RSS) of the unadjusted estimator, the standardized, and logistic coefficient estimators for the two settings described in Section 4. All estimators are given on the log odds scale. The logistic coefficient estimator estimates a conditional effect, while the other estimators estimate a marginal effect.

outcomes in the control group. The true risk difference between the treatment and control group is defined as $p_1 - p_0$. The unadjusted estimators for p_1 and p_0 are given by

$$\hat{p}_1 = \frac{1}{\sum_{i=1}^n A_i} \sum_{i=1}^n A_i Y_i, \quad \hat{p}_0 = \frac{1}{\sum_{i=1}^n (1 - A_i)} \sum_{i=1}^n (1 - A_i) Y_i.$$

That is, $\hat{p}_1$ and $\hat{p}_0$ are the proportions of successful outcomes observed in the treatment and control group, respectively. The unadjusted estimator for the risk difference between the treatment and control group is defined as $\hat{p}_1 - \hat{p}_0$.

Now we describe how the standardized estimator for the difference between the proportions of successful outcomes in the treatment and control group is calculated. Define the expit function as $\text{expit}(x) = 1/(1 + e^{-x})$. The first step is to fit the following main terms logistic regression model:

$$P(Y = 1|A, W) = \text{expit}(\beta_0 + \beta_1 A + \beta_2 W), \tag{1}$$

where β_2 and W are vectors of the same length. Denote the estimated coefficents by $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. The estimator for the probability $P(Y = 1|W, A)$ from the above model is given by

$$\hat{L}(A, W) = \text{expit}(\hat{\beta}_0 + \hat{\beta}_1 A + \hat{\beta}_2 W).$$

Here, $\hat{L}(1, W_i)$ is the predicted probability for a participant who receives treatment and has baseline variables W_i. Similarly, $\hat{L}(0, W_i)$ is the predicted probability for a participant in the control group

that has baseline variables W_i. Define the adjusted estimator for p_1 as

$$\hat{p}_1^{(ad)} = \frac{1}{n}\sum_{i=1}^{n} \hat{L}(1, W_i),$$

and the adjusted estimator for p_0 as

$$\hat{p}_0^{(ad)} = \frac{1}{n}\sum_{i=1}^{n} \hat{L}(0, W_i).$$

The adjusted estimator is defined as $\hat{p}_1^{(ad)} - \hat{p}_0^{(ad)}$. As written, the logistic regression model given by equation (1) only includes main effects, but it can easily be extended to incorporate interaction terms; in that case, the standardized estimator is still a consistent estimator for the average treatment effect.[9]

The logistic coefficient estimator is defined as the coefficient $\hat{\beta}_1$ in model (1).

The estimator for the modified R^2 is given by

$$1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{L}(A_i, W_i))^2}{\sum_{i=1}^{n}(Y_i - \hat{p}_{A_i})^2},$$

where $\hat{p}_{A_i}$ is the unadjusted estimator from the treatment group for participants in the treatment group and the unadjusted estimator for the control group for participants in the control group. As described in[14] leave one-out cross-validation can be used to avoid overfitting. To implement leave one out cross-validation, the models $\hat{L}(A_i, W_i)$ and $\hat{p}_{A_i}$ are fit on the dataset that does not include participant i.

A.3 Simulations Setup

In this section we will describe in more detail how the simulations in the results section are conducted.

For setting 1, the pair (Y, W) is sampled with replacement from the CLEAR III data. Initially, the treatment indicator A is generated from a Bernoulli distribution with probability 0.5, independent of (Y, W). To create a positive treatment effect of 0.13, the value of Y for the participants with $A = 1$ is set (overwritten) to 1 with probability $0.13/P(Y = 0)$.

For setting 2, W is sampled with replacement from the CLEAR III trial. The study arm variable A is simulated from a Bernoulli distribution with parameter 0.5, independenet of W. The outcome Y is simulated from a Bernoulli distribution with probability $P(Y = 1) + 0.065$ when $A = 1$ and $P(Y = 1) - 0.065$ when $A = 0$.

One participant had missing Glasgow Coma Scale and 37 participants had missing NIHSS score (7.5%). In order to keep the analysis simple, the results presented in sections 3 and 4.2 use median imputation for the missing baseline variables.

A.4 Treatment Effect as a Function of NIHSS in CLEAR III Data

Figure 2 shows the conditional probability of success within study arm (on the log-odds scale) as a function of the baseline variable NIHSS in the CLEAR III trial. The first two plots in Figure 2 are created using local polynomial regression with span equal to 0.75. The first plot shows the log odds of the probability of obtaining a successful outcome when assigned to control as a function of NIHSS, $\text{logit}(P(Y = 1|A = 0, W))$. The second plot is the analogous plot for treatment, $\text{logit}(P(Y = 1|A = 1, W))$. The third plot shows the treatment effect on the log-odds scale as a function of NIHSS, $\text{logit}(P(Y = 1|A = 1, W)) - \text{logit}(P(Y = 1|A = 0, W))$. The logistic coefficient estimator assumes that the effect seen in the third plot is a constant, that is, it is the same for all values of NIHSS. The plots in Figure 1 in the main paper are created by fitting a third degree polynomial to the data shown in plots one and two in Figure 2.

A.5 R and Stata code to Calculate the Standardized Estimator

The following R function calculates the standardized estimator.

```
# The first function calculates the standardized
# estimator and the second function calculates
# the bootstrapped variance estimator.
# The inputs are: outcome Y
# the treatment indicator A
# the baseline covariates W
```

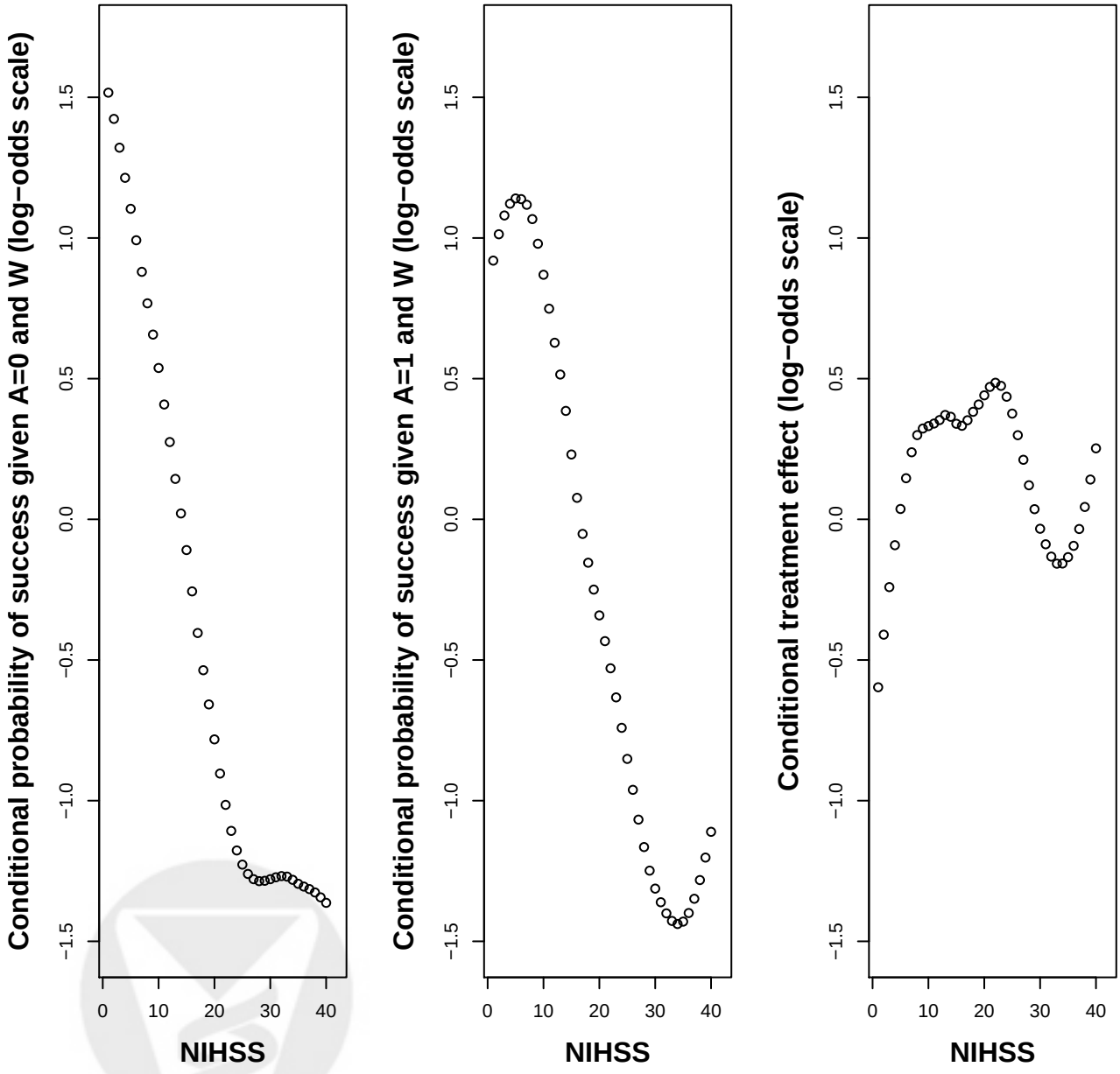

Figure 2: The conditional effect (on log-odds scale) of treatment as function of the NIHSS score in the CLEAR III data. The first two plots show the conditional probability of success (on the log odds scale) as a function of NIHSS, when assigned to treatment and control group, respectively. The third plot shows the conditional treatment effect (on the log odds scale) as a function of NIHSS.

```
#  For the second function n.boot is the number of bootstrap
# samples used to calculate the variance estimator.
# The output of the first function is the standardized estimator
# and the output of the second function is the  variance of the
# standardized estimator.

stand.est = function(Y, A, W){

# Creating the data frame
data.used = data.frame(W, A, Y)

# Fitting the logistic regression model
log.reg = glm(Y ~., data = data.used, family = "binomial")

# Creating dataset for calculating the predictions corresponding to
# A = 1 and A = 0
data.a.1 = data.used
data.a.1$A = 1
data.a.0 = data.used
data.a.0$A = 0

# Calculating the predictions
pred.1 =  predict.glm(log.reg, newdata =
data.a.1[, colnames(data.used) != "Y"]
, type = "response")
pred.0 =  predict.glm(log.reg, newdata =
data.a.0[, colnames(data.used) != "Y"]
, type = "response")
res.gcomp = mean(pred.1) - mean(pred.0)
```

```
return(res.gcomp)
}

# Calculating the variance estimator
var.stand.est = function(Y, A, W, n.boot){

# Creating the data frame
data.used = data.frame(W, A, Y)

# Calculating the variance estimator
boot.gcomp = rep(NA, n.boot)

for(i in 1:n.boot){

# Finding the bootstrap sample
bs = sample(1:nrow(data.used), size = nrow(data.used), replace = TRUE)

# Fitting the bootstrapped logistic regression model
log.reg.bs = glm(Y ~., data = data.used[bs, ], family = "binomial")
data.a.1 = data.used[bs, ]
data.a.1$A = 1
data.a.0 = data.used[bs, ]
data.a.0$A = 0
# Calculating the predictions
p.1.bs =  predict.glm(log.reg.bs, newdata =
data.a.1[, colnames(data.used) != "Y"]
, type = "response")
p.0.bs =  predict.glm(log.reg.bs, newdata =
data.a.0[, colnames(data.used) != "Y"]
, type = "response")
```

```
# Calculating the bootstrap estimator
boot.gcomp[i] = mean(p.1.bs) - mean(p.0.bs)
}

# Returning the estimator and variance estimator
return(var(boot.gcomp))
}
```

The following Stata code (written on Stata version 12) calculates the standardized estimator and the corresponding bootstrapped variance estimator.

```
* Stata code that calculates the standardized
* estimator as well as bootstrap variance
* estimator. The variables used are a binary
* outcome Y a binary study arm variable A
* and 5 covariates W1-W5. In the following
* all the covariates are continuous expect for
* W3 which is categorical.



* Fit the logistic regression model
logit Y i.A W1 W2 i.W3 W4 W5
* Getting the adjusted estimators for the risk difference
margins, dydx(A)

* Creating a program the bootstraps the
* standardized estimator for the risk difference
program bootmargins, rclass
logit Y i.A W1 W2 i.W3 W4 W5
margins, dydx(A)
matrix b = r(b)
```

```
return scalar riskdiff = b[1,2]
end

* Performing the bootstrap procedure
. bootstrap r(riskdiff), reps(1000) seed(1): bootmargins
```