4-24-2015

ADAPTIVE ENRICHMENT DESIGNS FOR RANDOMIZED TRIALS WITH DELAYED ENDPOINTS, USING LOCALLY EFFICIENT ESTIMATORS TO IMPROVE PRECISION

Michael Rosenblum
Johns Hopkins University, Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, mrosen@jhu.edu

Tianchen Qian
Johns Hopkins University, Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Yu Du
Johns Hopkins University, Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Huitong Qiu
Johns Hopkins University, Johns Hopkins Bloomberg School of Public Health, Depaetment of Biostatistics

Adaptive Enrichment Designs for Randomized Trials with Delayed Endpoints, using Locally Efficient Estimators to Improve Precision

MICHAEL ROSENBLUM*, TIANCHEN QIAN, YU DU, and HUITONG QIU

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, 21205

mrosen@jhu.edu

Summary

Adaptive enrichment designs involve preplanned rules for modifying enrollment criteria based on accrued data in an ongoing trial. For example, enrollment of a subpopulation where there is sufficient evidence of treatment efficacy, futility, or harm could be stopped, while enrollment for the remaining subpopulations is continued. Most existing methods for constructing adaptive enrichment designs are limited to situations where patient outcomes are observed soon after enrollment. This is a major barrier to the use of such designs in practice, since for many diseases the outcome of most clinical importance does not occur shortly after enrollment. We propose a new class of adaptive enrichment designs for delayed endpoints. At each analysis, semiparametric, locally efficient estimators leverage information in baseline variables and short-term outcomes to improve precision. This can reduce the sample size required to achieve a desired power. We propose new multiple testing procedures tailored to this problem, which we prove to strongly

*To whom correspondence should be addressed.

control the familywise Type I error rate, asymptotically. These methods are illustrated through simulations of a trial for a new surgical intervention for stroke.

Key words: multiple testing procedure; treatment effect heterogeneity

1. Introduction

We address the problem of designing a confirmatory randomized trial of an experimental treatment versus control when the primary outcome is measured with delay and there are multiple subpopulations of interest. Our methods were developed to solve a problem in designing a trial of a new surgical treatment for stroke, with outcomes measured a fixed time (180 days) from enrollment. However, our general method can also be applied to time-to-event outcomes.

To illustrate our approach, consider an analysis that occurs just after 50% of a trial's total enrollment. Due to delayed outcomes, less than 50% of final (i.e., primary) outcomes are observed. However, all enrolled participants have baseline variables observed, some have short-term outcomes observed, and a further subset have the final outcome observed. If the short-term outcomes and baseline variables are correlated with the final outcome, they can provide valuable information that we harness through the semiparametric, locally efficient estimators of van der Laan and Gruber (2012). In a randomized trial, these estimators converge to the true average treatment effect, without having to make any parametric model assumptions. To the best of our knowledge, we give the first application of such an estimator in adaptive enrichment designs with delayed outcomes. In simulations that mimic key features of a completed stroke trial, this leads to tangible improvements in precision and a 19-20% reduction in both expected sample size and maximum sample size, compared to the standard, unadjusted estimator that ignores short-term outcomes and baseline variables.

Our designs strongly control the familywise Type I error rate as required, e.g., by the U.S.

Food and Drug Administration in their draft guidance on adaptive designs for drugs and biologics (FDA, 2010). This means that the probability of rejecting one or more true null hypotheses is at most the desired level, regardless of the sign and magnitude of each subpopulation treatment effect. Two general techniques for ensuring strong control of the familywise Type I error rate in adaptive enrichment designs are the the p-value combination approach (Bretz *and others*, 2006; Schmidli *and others*, 2006; Jennison and Turnbull, 2007; Brannath *and others*, 2009), and the approach based on modified, group sequential computations (Stallard, 2011; Magnusson and Turnbull, 2013). These approaches require assumptions that are not guaranteed to hold when using semiparametric, locally efficient estimators in our context, as described in Section 4. To take advantage of precision gains from these estimators, we propose a class of multiple testing procedures that do not require these assumptions; these procedures build on ideas from the modified, group sequential computation approach. An alternative to our approach is to use conditional error functions as in the adaptive enrichment designs of (Friede *and others*, 2012); however, this involves computational challenges described in Section 7.

We present our motivating application in Section 2. The general problem is defined in Section 3. In Section 4, we describe the semiparametric, locally efficient estimators used in our designs. The proposed class of adaptive enrichment designs for delayed outcomes is given in Section 5. In Section 6, we apply our designs in simulations that mimic features of the data from a completed stroke trial. Section 7 describes extensions, limitations, and areas for future research.

2. Motivating Application: MISTIE stroke trial

We consider the problem of planning a Phase III trial to evaluate a new surgical treatment for stroke, called Minimally-Invasive Surgery Plus rt-PA for Intracerebral Hemorrhage Evacuation, abbreviated as MISTIE (Morgan *and others*, 2008). The aim is to assess whether the MISTIE surgical treatment is superior to the standard of care. The primary outcome is a participant's

degree of disability as measured by the modified Rankin Scale (mRS) at 180 days from enrollment. A mRS score of 3 or less is considered a successful outcome. Define the average effect of the MISTIE treatment to be the difference between the probability of a successful outcome under assignment to MISTIE (treatment) versus standard of care (control).

Prior data indicated greater uncertainty of the treatment effect for the subpopulation of participants with large (at least 10ml) intraventricular hemorrhage (IVH) at baseline, called large IVH participants. All others are called small IVH participants. The clinical investigators thought two scenarios were most likely to occur if the treatment was effective at all: either the new treatment would benefit both subpopulations, or it would benefit only small IVH participants. We explore adaptive designs for testing the corresponding null hypotheses of no mean treatment benefit for the overall population and for the small IVH subpopulation.

3. Problem Definition

3.1 *Subpopulations, Data Structure for Each Participant, and Analysis Timing*

We assume the overall population is partitioned into m disjoint subpopulations, which are pre-planned functions of variables measured before randomization. For each $s \in \{1, \ldots, m\}$, let p_s denote the proportion of the overall population in subpopulation s, which we assume is known.

Each participant i has full data vector $\mathbf{D}_i = (E_i, S_i, W_i, A_i, L_i^{(1)}, \ldots, L_i^{(T)}, Y_i)$ when followed up completely, where E_i is the enrollment time, S_i denotes subpopulation, W_i is a vector of baseline (pre-randomization) variables, A_i is the treatment indicator ($A_i = 1$ indicates treatment and $A_i = 0$ indicates control), $L_i^{(1)}, \ldots, L_i^{(T)}$ are variables observed after randomization, and Y_i is the final (i.e., primary) outcome. We assume that $L_i^{(1)}, \ldots, L_i^{(T)}, Y_i$ are observed at preplanned durations (in days) $\mathbf{d} = (d_1, \ldots, d_T, d_Y)$, respectively, from the time of enrollment, such that $0 < d_1 < \cdots < d_T < d_Y$. The subscript i is omitted when referring to a generic participant.

A special case of interest is where $L^{(1)}, \ldots, L^{(T)}$ represent the same quantity as in the primary

outcome, but measured at the earlier times $d_1, \ldots, d_T$; we refer to $L^{(1)}, \ldots, L^{(T)}$ as short-term outcomes, though in general they can be any variables measured after randomization. For example, in the MISTIE trial we have $T = 1$ and the following data are measured for each participant: enrollment time E, subpopulation $S \in \{1, 2\}$ (small or large IVH, respectively); baseline variables $W =$ (NIH Stroke Scale, clot volume, and Glasgow Coma Scale); treatment indicator A; indicator $L^{(1)}$ of functional disability score (mRS) $\leqslant 3$ at 30 days from enrollment; the primary outcome Y, which is the indicator of mRS $\leqslant 3$ at 180 days from enrollment.

Let K denote the maximum number of stages and $N_{s,\max}$ denote the maximum, cumulative sample size for each subpopulation s, both of which are preplanned. The maximum total sample size is $N = \sum_{s \leqslant m} N_{s,\max}$. Define $q_s = N_{s,\max}/N$. At the start of the trial, all subpopulations are continuously enrolled. Let e denote the combined population enrollment rate in participants per day, which we assume to be a constant. We assume enrollment rates are proportional to subpopulation sizes, i.e., if enrollment has not stopped for subpopulations s and s', the ratio of the cumulative number enrolled from subpopulation s and s' is $p_s/p_{s'}$. The enrollment time for the lth participant from subpopulation s is $l/(ep_s)$, for each $s \leqslant m, l \leqslant N_{s,\max}$. We order the set of all participants by increasing enrollment time (with ties broken arbitrarily), and denote the full data vector for the ith participant in this ordering by $\mathbf{D}_i$ as defined above. Each stage's duration can be any preplanned function of calendar time and/or information accrued, as defined in Section 5. The maximum trial duration $\mathcal{D}$, which is when analysis K occurs, is the time to enroll all participants plus d_Y, i.e., $\mathcal{D} = \max_{s \leqslant m}\{N_{s,\max}/(ep_s)\} + d_Y$.

3.2 *Assumptions, Hypotheses, Censoring, and Accrual Modification Rules*

For each participant i, we assume that conditioned on E_i and $S_i = s$, his/her baseline data W_i is a random draw from an unknown distribution $Q_s(W)$, independent of the data from all previously enrolled participants. By design, each participant is randomized with probability 1/2

to either study arm independent of (E, S, W), i.e., $P(A = 1 \mid E, S, W) = 1/2$; we call this the randomization assumption. We assume that for each participant i, conditioned on $E_i, S_i = s, W_i, A_i$, the vector $(L_i^{(1)}, \ldots, L_i^{(T)}, Y_i)$ is a random draw from distribution $Q'_s(L^{(1)}, \ldots, L^{(T)}, Y \mid A = A_i, W = W_i)$ independent of the data from all previously enrolled participants. Denote the unknown distributions by $Q = \{(Q_s, Q'_s) : s \leqslant m\}$. We make no parametric model assumptions, nor do we assume any relationships between distributions for different subpopulations. We use a nonparametric model $\mathcal{Q}$ where the only assumptions are the randomization assumption and that Q satisfies regularity conditions given in Appendix A of the Supplementary Material.

Let $\delta_s = E(Y \mid A = 1, S = s) - E(Y \mid A = 0, S = s)$ denote the average treatment effect for subpopulation $s \in \{1, \ldots, m\}$. For each $j \in \{0, \ldots, J\}$, let $\tilde{\mathcal{S}}_j \subseteq \{1, \ldots, m\}$ denote the jth composite population of interest, consisting of the union of subpopulations in $\tilde{\mathcal{S}}_j$. The overall population will generally be of interest, and we denote it by $\tilde{\mathcal{S}}_0 = \{1, \ldots, m\}$. The average treatment effect in population $\tilde{\mathcal{S}}_j$ is $\Delta_j = E(Y \mid A = 1, S \in \tilde{\mathcal{S}}_j) - E(Y \mid A = 0, S \in \tilde{\mathcal{S}}_j) = \sum_{s \in \tilde{\mathcal{S}}_j} p_s \delta_s \Big/ \sum_{s \in \tilde{\mathcal{S}}_j} p_s$. For each $j \in \{0, \ldots, J\}$, define the null hypothesis $H_{0j} : \Delta_j \leqslant 0$ and alternative hypothesis $\Delta_j > 0$. Our general problem is to construct adaptive enrichment designs to test the set of null hypotheses $\{H_{0j} : j = 0, \ldots, J\}$. For any distribution $Q \in \mathcal{Q}$, let $\mathcal{T}(Q)$ denote the set of true null hypotheses under Q. We require at least $n_{\min} > 1$ participants from each population $\tilde{\mathcal{S}}_j$ to have Y observed before analysis 1 takes place.

We assume $L^{(0)} = (E, S, W, A)$ are observed at enrollment. Let $L^{(T+1)} = Y$. For each participant i, stage k, and $t \leqslant T + 1$, let $C_{i,k}^{(t)}$ denote the indicator that $L_i^{(t)}$ is observed by the end of stage k. In the special case of the MISTIE trial, for any participant i and stage k, the vector $(C_{i,k}^{(0)}, C_{i,k}^{(1)}, C_{i,k}^{(2)})$ has one of the following forms:

$(0, 0, 0)$: no data observed, i.e., not yet enrolled by end of stage k;

$(1, 0, 0)$: enrollment time E, subpopulation S, baseline variables W and study arm A observed;

$(1, 1, 0)$: E, S, W, A, and short-term outcome $L^{(1)}$ observed;

$(1,1,1)$: complete data vector $(E, S, W, A, L^{(1)}, Y)$ observed.

In general, we assume a monotone missingness structure, i.e., that $C_{i,k}^{(t)} \geqslant C_{i,k}^{(t+1)}$ for each $t \in \{0,\ldots,T\}, k \leqslant K, i \leqslant N$, and that $C_{i,k}^{(t)} \leqslant C_{i,k+1}^{(t)}$ for each $t \in \{0,\ldots,T+1\}, k \leqslant K-1, i \leqslant N$. Define the pipeline participants at analysis k to be those enrolled with Y not yet observed.

We assume the only cause of missing data is administrative censoring due to some participants not yet having experienced short-term and/or final outcomes at an interim analysis. In Section 7 we describe how to incorporate additional right censoring due to loss to follow-up. For clarity of exposition, we assume throughout that variances and covariances are known; in practice, they will be estimated as the trial progresses, e.g., using the nonparametric bootstrap as described in Appendix D of the Supplementary Material.

An early stop of accrual (of information) for subpopulation s means that both enrollment and continuation of follow-up are halted. Unless subpopulation s accrual is stopped early, it has ongoing enrollment (until $N_{s,\max}$ is reached) and follow-up (until all $N_{s,\max}$ participants have Y observed). Let r_k denote the preplanned rule for accrual modification at analysis $k < K$. It can be any measurable function from the data available at analysis k to the set of subpopulations for which accrual will continue during stage $k+1$, under the restrictions that accrual can only be stopped at interim analysis times and that once a subpopulation's accrual has been stopped it cannot be restarted. An example is given in Section 6.3. Let $\mathbf{r} = (r_1, \ldots, r_{K-1})$.

3.3 *Asymptotic Framework*

Our asymptotic results, such as consistency and asymptotic normality of estimators, involve a sequence of hypothetical trials with sample sizes in all stages converging to infinity. We fix the proportions p_s, the delay times $\mathbf{d}$, the maximum duration $\mathcal{D}$, the analysis times, and the distribution Q. We set $N_{s,\max} = q_s N$ for fixed constants $q_s > 0$ that sum to 1, and consider a sequence of trials in which N goes to infinity. This implies that the enrollment rate e goes to

infinity, the number enrolled in each stage goes to infinity, and if enrollment has not stopped for subpopulation s by the end of stage k then the proportion of its participants in the pipeline is a positive constant that depends on k; these results are proved in Appendix A.1 of the Supplementary Material. This framework is similar to (Scharfstein *and others*, 1997, Section 2), except ours is more restricted because the enrollment process and delay times are fixed. Though our asymptotic results only require $N_{s,\max}/N$ to converge to q_s, for clarity of exposition we consider the case where $N_{s,\max}/N = q_s$. Consider any $Q \in \mathcal{Q}$, rule $\boldsymbol{r}$, and multiple testing procedure M. We say M controls the familywise Type I error rate at level α^*, asymptotically, if $\limsup_{N\to\infty} P_{Q,\boldsymbol{r},N}\{M \text{ rejects at least one null hypothesis in } \mathcal{T}(Q)\} \leqslant \alpha^*$, where $P_{Q,\boldsymbol{r},N}$ denotes probability under distribution Q, rule $\boldsymbol{r}$, and maximum sample size N. If this holds for any $Q \in \mathcal{Q}$, then M *strongly* controls the familywise Type I error rate, asymptotically. Familywise Type I error control for fixed N is defined similarly, except dropping $\limsup_{N\to\infty}$.

3.4 *Unadjusted Estimator*

We define all estimators, statistics, and corresponding covariance matrices as if a rule $\mathbf{r}$ is used such that no subpopulation's accrual is ever stopped early. This poses no problem since our testing procedures never use an estimator or statistic if any of its component subpopulations had accrual stopped early at a prior analysis. For a given population, the unadjusted estimator of the average treatment effect is the difference between sample means of Y comparing those assigned to $A = 1$ versus $A = 0$. At any analysis time, only the data from participants who have Y observed are used. For each subpopulation $s \in \{1, \ldots, m\}$, the unadjusted estimator of δ_s at analysis k is

$$\hat{\delta}_{s,k}^{unadj} = \frac{\sum_i Y_i C_{i,k}^{(T+1)} 1\left[A_i = 1, S_i = s\right]}{\sum_i C_{i,k}^{(T+1)} 1\left[A_i = 1, S_i = s\right]} - \frac{\sum_i Y_i C_{i,k}^{(T+1)} 1\left[A_i = 0, S_i = s\right]}{\sum_i C_{i,k}^{(T+1)} 1\left[A_i = 0, S_i = s\right]},$$

where $1[X]$ is the indicator variable taking value 1 if X is true and 0 otherwise. The unadjusted estimator of Δ_j for composite population $\tilde{\mathcal{S}}_j$ at stage k is $\hat{\Delta}_{j,k}^{unadj} = \sum_{s\in\tilde{\mathcal{S}}_j} p_s \hat{\delta}_{s,k}^{unadj} \Big/ \sum_{s\in\tilde{\mathcal{S}}_j} p_s$.

4. Semiparametric, Locally Efficient Estimators

To take advantage of prognostic information in baseline variables and short-term outcomes, we use estimators that build on the general theory of semiparametric, locally efficiency of Robins and Rotnitzky (1992). When baseline variables and short-term outcomes are strongly correlated with the final outcome, as in the MISTIE trial, these estimators can have greater precision than the unadjusted estimator. We use a targeted maximum likelihood estimator (TMLE) for longitudinal data developed by van der Laan and Gruber (2012) and implemented in the R package **ltmle** (Schwab *and others*, 2014). This estimator combines features of the general targeted maximum likelihood template of van der Laan and Rubin (2006) with the sequential regression approach of Robins (2000); Bang and Robins (2005). We call this the adjusted estimator. Let $\hat{\Delta}_{j,k}^{adj}$ denote the adjusted estimator of Δ_j based on all data from participants in population $\tilde{\mathcal{S}}_j$ collected up to and including stage k. The precise definition of this estimator is given in Appendix B of the Supplementary Material. It is also possible to use the semiparametric, locally efficient estimators of, e.g., Lu and Tsiatis (2011); Rotnitzky *and others* (2012); Gruber and van der Laan (2012); Parast *and others* (2014); Zhang (2015), as we discuss in Section 7.

The adjusted estimator involves working models that are fit using data accrued at a given analysis. The term "working model" means we do not assume that the true, unknown data generating distribution Q satisfies any of the assumptions of these models. For example, our TMLE implementation uses a logistic regression working model for $P(Y = 1 \mid L^{(1)}, A, W, S \in \tilde{\mathcal{S}}_j)$ for each $j \leqslant J$, but we do not assume the conditional distribution of Y given $(L^{(1)}, A, W, S)$ has the functional form of a logistic regression model. Under regularity conditions given in Appendix A of the Supplementary Material, the adjusted estimator is consistent regardless of whether the working models are correctly specified. If they are correctly specified, then the adjusted estimator achieves the semiparameteric efficiency bound; this is the local efficiency property of the estimator.

When our discussion applies to a generic estimator, we suppress *adj* and *unadj* in the subscript.

For a given estimator $\hat{\Delta}_{j,k}$ of Δ_j, define the corresponding Wald statistic $Z_{j,k} = \hat{\Delta}_{j,k}/\text{Var}(\hat{\Delta}_{j,k})^{1/2}$, where Var denotes variance. When the adjusted estimator involves at least a few baseline variables or short-term outcomes that are continuous-valued (or discrete-valued with many levels and treated as continuous in working models), e.g., as in the MISTIE example, then we expect these models to be at least somewhat misspecified. We consider such a situation throughout the paper.

For a given $\tilde{\mathcal{S}}_j$, the statistics $Z_{j,1}, \ldots, Z_{j,K}$ for the adjusted estimator are not guaranteed to have the canonical covariance that arises when estimators (rescaled by the information) have the independent increments property described by Scharfstein *and others* (1997); Jennison and Turnbull (1999). This was shown for estimators based on generalized estimating equations by Shoben and Emerson (2014). We show this occurs for the TMLE and some other locally efficient estimators when working models are misspecified, in Appendix A of the Supplementary Material. We furthermore show this occurs even when baseline variables but no short-term outcomes are used in working models. The upshot is that the general techniques for ensuring familywise Type I error control listed in Section 1 cannot be directly applied, since they assume the canonical covariance or more generally the so-called p-clud property, neither of which is guaranteed to hold for the adjusted estimator. An exception involving only linear models is described in Appendix A.

5. Multiple Testing Procedure using Interleaved Error Spending Functions

Error-spending functions were introduced by Slud and Wei (1982); Lan and DeMets (1983) for a single population, but have not been applied in the manner we describe below for multiple populations in adaptive enrichment designs with delayed outcomes. Error-spending functions set boundaries for early stopping based on the information accrued. We define a separate error spending function for each composite population $\tilde{\mathcal{S}}_j$. Tests are interleaved in a way that takes advantage of correlations among related statistics, including statistics for the same population at different analysis times, and statistics for different but overlapping populations. We focus

on efficacy boundaries, whose corresponding error spending functions are called alpha spending functions. A special case of our general class of designs was given by Rosenblum *and others* (2015), who only considered immediately observed outcomes, unadjusted estimators, and more restricted designs than the class below.

For each null hypothesis H_{0j}, at each analysis k, let $\mathcal{I}_{j,k} = 1/\text{Var}(\hat{\Delta}_{j,k})$ denote the accrued information corresponding to the estimator $\hat{\Delta}_{j,k}$. Define the information fraction $\tau_{j,k} = \mathcal{I}_{j,k}/\mathcal{I}_{j,\max}$, where $\mathcal{I}_{j,\max}$ is a predefined maximum information level for population $\tilde{\mathcal{S}}_j$, and $\mathcal{I}_{j,0} = 0$ for all j. Let α^* denote the desired upper bound on the familywise Type I error rate, e.g., $\alpha^* = 0.025$ (since we use one-sided tests). Define error spending functions $\alpha_j : [0,\infty) \to [0,\alpha^*]$ for each $j \in \{0,\ldots,J\}$ that are nondecreasing, take the value 0 at $\tau = 0$, and satisfy $\sum_{j=0}^{J} \alpha_j(\tau) \leqslant \alpha^*$ for all $\tau > 0$. Define $\pi_{j,k} = \max\{0, \alpha_j(\tau_{j,k}) - \alpha_j(\tau_{j,k-1})\}$ and let $\boldsymbol{\pi} = \{\pi_{j,k}\}_{j\leqslant J, k\leqslant K}$.

Consider any accrual modification rule $\mathbf{r}$, nonnegative increments $\boldsymbol{\pi}$, and $Q \in \mathcal{Q}$. We first give results for an arbitrary vector of statistics $\tilde{\mathbf{Z}} = \{\tilde{Z}_{j,k}\}_{j\leqslant J, k\leqslant K}$ with covariance matrix $\tilde{\boldsymbol{\Sigma}}$ having 1's on the main diagonal, and then apply these results for $\tilde{\mathbf{Z}}$ equal to the Wald statistics corresponding to the unadjusted or adjusted estimator. For each k define the following set of null hypotheses: $B_k(\mathbf{r}) = \{H_{0j} : \text{for each } s \in \tilde{\mathcal{S}}_j, \text{ subpopulation } s \text{ accrual not stopped early prior to analysis } k\}$. Define the ordering $(k',j') \prec (k,j)$ to mean that $k' < k$ or ($k' = k$ and $j' < j$). Denote the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ by $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$. The following procedure takes as input the statistics $\tilde{\mathbf{Z}}$, their covariance $\tilde{\boldsymbol{\Sigma}}$, the rule $\mathbf{r}$, and increments $\boldsymbol{\pi}$:

Multiple Testing Procedure $M(\tilde{\mathbf{Z}},\tilde{\boldsymbol{\Sigma}},\mathbf{r},\boldsymbol{\pi})$**:** Define $\mathbf{Z}' = \{Z'_{j,k}\}_{j\leqslant J, k\leqslant K}$ to be a random vector with distribution $N(\mathbf{0},\tilde{\boldsymbol{\Sigma}})$. At each analysis $k \leqslant K$, for each population $\tilde{\mathcal{S}}_j$, $j = 0,\ldots,J$ in turn:

1. Define $u_{j,k}$ to be the solution to:

$$P\left\{Z'_{j',k'} \leqslant u_{j',k'} \text{ for all } (k',j') \prec (k,j); \text{ and } Z'_{j,k} > u_{j,k}\right\} = \pi_{j,k}. \tag{5.1}$$

2. Reject H_{0j} if all of the following hold: it hasn't already been rejected, $j \in B_k(\mathbf{r})$, and $\tilde{Z}_{j,k} > u_{j,k}$.

The left side of (5.1) can be computed using the multivariate normal distribution function, e.g., implemented in the mvtnorm R package of Genz *and others* (2014), which takes as input $\tilde{\mathbf{\Sigma}}$. Given the previously computed values $\{u_{j',k'} : (k', j') \prec (k, j)\}$, the solution $u_{j,k}$ to (5.1) can be computed to high precision by the bisection (binary search) method. In the special case that $\pi_{j,k} = 0$, we set $u_{j,k} = \infty$. The null hypotheses rejected at the end of the trial are those that were rejected at any stage.

Let $\mathcal{T} = \mathcal{T}(Q)$ as defined on page 6. For any vector $\tilde{\mathbf{Z}} = \{\tilde{Z}_{j,k}\}_{j \leqslant J, k \leqslant K}$, define the subvector $\tilde{\mathbf{Z}}_{\mathcal{T}} = \{\tilde{Z}_{j,k} : k \leqslant K, j : H_{0j} \in \mathcal{T}\}$ and let $\tilde{\mathbf{\Sigma}}_{\mathcal{T}}$ denote its covariance matrix. For any random vectors $\mathbf{U}, \mathbf{U}'$ taking values in $\mathbb{R}^v$, we say $\mathbf{U}'$ is stochastically smaller than $\mathbf{U}$ if for any $\mathbf{u} \in \mathbb{R}^v$, $P(\text{for some } v' \leqslant v, U'_{v'} > u_{v'}) \leqslant P(\text{for some } v' \leqslant v, U_{v'} > u_{v'})$. A sequence $\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots$ taking values in $\mathbb{R}^v$ is asymptotically, stochastically smaller than $\mathbf{U}$ if for any $\mathbf{u} \in \mathbb{R}^v$, we have $\limsup_{l>0} P(\text{for some } v' \leqslant v, U^{(l)}_{v'} > u_{v'}) \leqslant P(\text{for some } v' \leqslant v, U_{v'} > u_{v'})$.

Consider any $Q \in \mathcal{Q}$, accrual modification rule $\mathbf{r}$, and nonnegative increments $\boldsymbol{\pi}$ that sum to at most α^*. In Appendix C of the Supplementary Material we prove:

THEOREM 5.1 Consider any vector of statistics $\tilde{\mathbf{Z}} = \{\tilde{Z}_{j,k}\}_{j \leqslant J, k \leqslant K}$ with covariance matrix $\tilde{\mathbf{\Sigma}}$ having 1's on the main diagonal. If $\tilde{\mathbf{Z}}_{\mathcal{T}}$ is stochastically smaller than $N(\mathbf{0}, \tilde{\mathbf{\Sigma}}_{\mathcal{T}})$, then the procedure $M(\tilde{\mathbf{Z}}, \tilde{\mathbf{\Sigma}}, \mathbf{r}, \boldsymbol{\pi})$ controls the familywise Type I error rate at level α^*.

THEOREM 5.2 Consider any sequence $\tilde{\mathbf{Z}}^{(N)}$ with corresponding covariance matrices $\tilde{\mathbf{\Sigma}}^{(N)}$ such that $\tilde{\mathbf{\Sigma}}^{(N)}_{\mathcal{T}}$ converges to a limit $\tilde{\mathbf{\Sigma}}^*_{\mathcal{T}}$ with 1's on the main diagonal, and $\tilde{\mathbf{Z}}^{(N)}_{\mathcal{T}}$ is asymptotically, stochastically smaller than $N(\mathbf{0}, \tilde{\mathbf{\Sigma}}^*_{\mathcal{T}})$. Then $M(\tilde{\mathbf{Z}}^{(N)}, \tilde{\mathbf{\Sigma}}^{(N)}, \mathbf{r}, \boldsymbol{\pi})$ controls the familywise Type I error rate, asymptotically, at level α^*.

Consider a sequence of trials with maximum sample size N going to infinity, as described in Section 3.3. For each N, let $\hat{\Delta}^{(N)}_{j,k}$ denote the unadjusted or adjusted estimator for population

$\tilde{\mathcal{S}}_j$ at stage k, with corresponding information $\mathcal{I}_{j,k}^{(N)}$ and Wald statistic $Z_{j,k}^{(N)}$. Similar to Scharfstein *and others* (1997, Section 3), we assume that for each $j \leqslant J, k \leqslant K$, $\lim_{N\to\infty} \mathcal{I}_{j,k}^{(N)}/N = \lim_{N\to\infty}\{N\text{Var}(\hat{\Delta}_{j,k}^{(N)})\}^{-1} = \mathcal{I}_{j,k}^{*}$ where $\mathcal{I}_{j,k}^{*}$ is defined in Appendix A of the Supplementary Material. This implies the covariance matrix $\mathbf{\Sigma}^{(N)}$ of $\mathbf{Z}^{(N)}$ converges to $\mathbf{\Sigma}^{*}$ (defined in that Appendix) with 1's on the main diagonal. Under the regularity conditions in that Appendix, the centered Wald statistics $\{(\hat{\Delta}_{j,k}^{(N)} - \Delta_j)/\text{Var}(\hat{\Delta}_{j,k}^{(N)})^{1/2}\}_{j\leqslant J, k\leqslant K}$ converge in distribution to $N(\mathbf{0}, \mathbf{\Sigma}^{*})$. For each true H_{0j}, we have $Z_{j,k}^{(N)} = \hat{\Delta}_{j,k}^{(N)}/\text{Var}(\hat{\Delta}_{j,k}^{(N)})^{1/2} \leqslant (\hat{\Delta}_{j,k}^{(N)} - \Delta_j)/\text{Var}(\hat{\Delta}_{j,k}^{(N)})^{1/2}$, and so $\mathbf{Z}_{\mathcal{T}}^{(N)}$ is asymptotically, stochastically smaller than $N(\mathbf{0}, \mathbf{\Sigma}_{\mathcal{T}}^{*})$. Theorem 5.2 implies $M(\mathbf{Z}^{(N)}, \mathbf{\Sigma}^{(N)}, \mathbf{r}, \boldsymbol{\pi})$ controls the familywise Type I error rate, asymptotically, at level α^{*}. Since this holds for any $Q \in \mathcal{Q}$, the procedure strongly controls the familywise Type I error rate, asymptotically.

Magnusson and Turnbull (2013) present multiple testing procedures that use error spending functions in adaptive enrichment designs. Their method is not applicable in our context since it assumes the canonical covariance described above. It also assumes that the treatment effect cannot be negative in any subpopulation. Negative treatment effects cannot be ruled out in the MISTIE trial context, since it is possible that the new surgical procedure may cause damage.

6. Simulations

6.1 *Overview and Design Goals*

Consider the problem of planning the Phase III MISTIE trial, as introduced in Section 2. The variables $(S, W, A, L^{(1)}, Y)$ defined in the third paragraph of Section 3 are measured for each participant. We refer to those with small IVH as subpopulation 1, and those with large IVH as subpopulation 2. The composite populations of interest are the combined population denoted by $\tilde{\mathcal{S}}_0 = \{1, 2\}$, and subpopulation 1 denoted by $\tilde{\mathcal{S}}_1 = \{1\}$. We test the corresponding null hypotheses H_{00} and H_{01} using Wald statistics $Z_{0,k}$ and $Z_{1,k}$ for each $k \leqslant K$ in multiple testing procedure M. We use statistics $Z_{2,k}$, which involve only subpopulation 2, in the accrual modification rule

r defined in Section 6.3. In the adaptive enrichment design literature, it is not uncommon to consider the null hypotheses for a single subpopulation and the combined population, e.g., Wang *and others* (2007); Brannath *and others* (2009); Jenkins *and others* (2011); Freidlin *and others* (2013); Stallard *and others* (2014, Section 5). We assume $p_1 = 1/3$ based on prior studies (Hanley, 2012). We assume the enrollment rate is 50 patients per year for subpopulation 1, and 100 per year for subpopulation 2, based on the projected enrollment rates for the MISTIE Phase III trial.

The clinical investigators in the MISTIE trial were interested in the following three scenarios: (a) $\delta_1 = 12.2\%$, $\delta_2 = 12.2\%$; (b) $\delta_1 = 12.2\%$, $\delta_2 = 0\%$; (c) $\delta_1 = \delta_2 = 0$. The values of δ_1, δ_2 in scenario (a) are based on the point estimate of the average treatment effect from the MISTIE II trial. We had the following goals: (i) 80% power to reject H_{00} in scenario (a); (ii) 80% power to reject H_{01} in scenario (b); (iii) strong control of familywise Type I error rate at level $\alpha^* = 0.025$. Similar goals were also considered by Rosenblum *and others* (2015) in the context of immediately observed outcomes and no baseline variables W.

6.2 *Data Generating Distributions used in Simulation Study*

To make our simulations realistic, we mimic features in the data from the completed MISTIE II trial introduced in Section 2. A simple approach would be to construct simulated trials by resampling with replacement from the MISTIE II data, so that the data generating distribution is the corresponding empirical distribution. Unfortunately, the resampling distribution does not satisfy the randomization assumption from Section 3.2, since there are slight correlations between baseline variables and treatment assignment in the actual MISTIE II data set (as would generally be expected in any given dataset). Furthermore, since no two participants in this data have identical values of the baseline variables W, the treatment A is a deterministic function of W.

We construct data generating distributions that mimic key features of the MISTIE II data, while satisfying the randomization assumption. Specifically, we construct distributions with sim-

ilar correlations among $W, L^{(1)}, Y$ as the Phase II trial data. This is achieved by adding, for each participant from the MISTIE II data, a "twin" participant with identical baseline variables but opposite treatment assignment, and whose $L^{(1)}$ and Y are generated using regression models fit to the original data, with perturbations to the outcomes Y depending on the desired treatment effect in each subpopulation. For each scenario (a)-(c), a distribution was constructed that satisfies the assumptions from Section 3.2, as described in Appendix E of the Supplementary Material.

6.3 *Specific Adaptive Enrichment Design Used*

We define our adaptive enrichment design by first giving the multiple testing procedure and accrual modification rule $\mathbf{r}$, and then presenting the analysis timing and error spending functions. We use the multiple testing procedure M from Section 5, which generates efficacy boundaries $u_{j,k}$ at each analysis k. The accrual modification rule $\mathbf{r}$ involves futility boundaries $l_{j,k}$ defined below. The following encodes our accrual modification rule (and indicates when null hypotheses are rejected, based on multiple testing procedure M) at the analysis at the end of stage $k \leqslant K$:

1. if $Z_{0,k} > u_{0,k}$ or $Z_{1,k} > u_{1,k}$, reject the corresponding null hypotheses and stop all accrual;
2. else, if $Z_{1,k} \leqslant l_{1,k}$ or $k = K$, then stop all accrual and fail to reject both null hypotheses;
3. else, if accrual continued for both subpopulations in stage k and $Z_{2,k} \leqslant l_{2,k}$, stop subpopulation 2 accrual (and fail to reject H_{00}) but continue subpopulation 1 in stage $k+1$;
4. else, if $k < K$ accrual continues for the same subpopulations in stage $k+1$ as in stage k.

Since the above design uses multiple testing procedure M, under the conditions in Theorem 5.2 it strongly controls the familywise Type I error rate at level α^*, asymptotically. In step 1, if both $Z_{0,k} > u_{0,k}$ and $Z_{1,k} > u_{1,k}$, then H_{00} and H_{01} are rejected; if only one of these conditions holds, then only the corresponding null hypothesis is rejected. Step 2 is motivated by the clinical investigator's judgment that if the treatment benefits any subpopulation, it will very likely benefit

subpopulation 1, so the entire trial should stop for futility if $Z_{1,k} \leqslant l_{1,k}$. The above design (steps 1-4) is just one possible choice; the general multiple testing procedure M can be applied for any accrual modification rule $\mathbf{r}$.

We set the maximum number of stages $K = 5$. Alpha spending functions are from the ρ-family of Kim and DeMets (1987) at $\rho = 2$, i.e., $\alpha_j(\tau) = c_j \min\{\tau^2, 1\}$ for each population $\tilde{\mathcal{S}}_j : j \in \{0, 1\}$, for nonnegative coefficients c_0, c_1 that sum to $\alpha^* = 0.025$. Each analysis $k \leqslant K$ occurs when the accrued information for subpopulation 1 approximately reaches $\tau_{1,k}\mathcal{I}_{1,\max}$; our approximation to this information-based monitoring plan is described below.

First, consider the unadjusted estimator. The values of c_j, $\mathcal{I}_{j,\max}$, $\tau_{1,k}$, $l_{j,k}$ for each $j \in \{0, 1\}, k \leqslant K$ were chosen by searching over a set of candidate values to find those that minimize the average of the expected sample size over scenarios (a)-(c), under the constraint that goals (i)-(iii) are satisfied. We used the optimization procedure from Rosenblum *and others* (2015) to conduct this search, which resulted in $c_0 = 0.003$, $c_1 = 0.022$, $\mathcal{I}_{0,\max} = 1115$, $\mathcal{I}_{1,\max} = 795$, $(\tau_{1,1}, \tau_{1,2}, \tau_{1,3}, \tau_{1,4}, \tau_{1,5}) = (0.16, 0.32, 0.47, 0.74, 1)$ and $l_{j,k}$ as given in Table 2. The futility boundaries $l_{j,k}$ equal 0 in most cases, with the notable exception $l_{2,3} = \infty$; this causes subpopulation 2 accrual to stop at or before analysis 3; intuitively, this is because at analysis 3, sufficient information has accrued to achieve goal (i), so further enrollment of subpopulation 2 would be counterproductive. The first three analysis times, when expressed in terms of information accrued for the combined population, approximately equal $(1/3, 2/3, 1)\mathcal{I}_{0,\max}$.

Next, consider the adjusted estimator. We slightly increased $\mathcal{I}_{0,\max}$ and decreased $\mathcal{I}_{1,\max}$ so that goals (i)-(iii) are achieved using this estimator, which has a different covariance matrix than the unadjusted estimator. Consider scenario (a). Table 1 shows the per-stage information levels $\mathcal{I}_{j,k}$ for each estimator and Table 2 shows the corresponding sample sizes. Because information accrues more quickly when using the adjusted estimator, its corresponding sample sizes at each stage are smaller than those for the unadjusted estimator. The maximum sample sizes for sub-

populations 1 and 2 when using the adjusted estimator are $N_{1,\max} = 512, N_{2,\max} = 504$, while those for the unadjusted estimator are $N_{1,\max} = 648, N_{2,\max} = 624$.

For a given estimator, the per-stage sample sizes corresponding to the information levels in Table 1 were almost identical across scenarios (a)-(c). For simplicity, in our simulation study we set interim analyses to occur at the sample sizes in Table 2 (rather than at preset information levels), for each scenario. This approximation to information-based monitoring led to negligible differences in information accrued at each analysis, among the different scenarios. We emphasize that the per-stage sample sizes differ by estimator, but not by scenario.

In order to assess whether the adjusted estimator performs worse than the unadjusted estimator when covariates are pure noise, we consider a modified data generating distribution with $(W, L^{(1)})$ exogenous, i.e., independent of all other variables. Denote the TMLE under this type of data generating distribution by TMLE $\text{prog}_{\emptyset}$, and denote the TMLE using the data generating distributions in Section 6.2 (where W and $L^{(1)}$ are prognostic) by TMLE $\text{prog}_{W,L}$. We set the analysis timing for TMLE $\text{prog}_{\emptyset}$ to be the same as for the unadjusted estimator, in calendar time.

The efficacy boundaries $u_{j,k}$, which are determined by (5.1), depend on the covariance matrix $\mathbf{\Sigma}$ of the statistics under consideration. To ease the computational burden in our simulations, we precomputed an approximation to $\mathbf{\Sigma}$ using Monte Carlo simulation as described in Appendix D of the Supplementary Material, and treated $\mathbf{\Sigma}$ as known. This was done separately for each estimator and scenario (a)-(c). The resulting boundaries $u_{j,k}$ for scenario (a) and the unadjusted estimator are given in Table 3. These boundaries were quite similar for each estimator and scenario (a)-(c); the maximum absolute difference was 0.02.

Wang *and others* (2009) define adaptive enrichment to be a preplanned rule for restricting enrollment based on accrued data. The above adaptive design has such a rule for early stopping of only subpopulation 2 for futility if $Z_{2,k} \leqslant l_{2,k}$ at analysis $k = 1$ or $k = 2$. We call this the adaptive enrichment feature. Though the above design always stops subpopulation 2 enrollment at the end

of stage 3, we do not consider this to be *adaptive* enrichment since this occurs regardless of the accrued data. To show the value added by the adaptive enrichment feature, consider the same design as described above except setting $l_{2,1} = l_{2,2} = -\infty$, which disables this feature. We call this the non-adaptive design, which we compare to the adaptive design.

6.4 *Results: Power, Expected Sample Size, and Maximum Sample Size*

Based on 50,000 simulated trials for each estimator and scenario (a)-(c), we computed the empirical Type I error, power, and expected sample size (ESS, defined as the expected number enrolled, which includes those in the pipeline). These were computed under the accrual modification rule **r** from Section 6.3. An exception is that when computing the familywise Type I error rate we assumed no early stopping of accrual, in order to show that Type I error is controlled even in this case; early stopping would only leave unchanged or decrease the Type I error.

The top half of Table 4 summarizes results for the adaptive enrichment design from Section 6.3. In each scenario, the power of the different estimators is very similar due to the information-based design using similar $\mathcal{I}_{j,\max}$ values for each estimator. Essentially all the gains from adjusting for prognostic variables get channeled into reducing the expected sample size. Using the adjusted estimator (TMLE $\text{prog}_{W,L}$) instead of the unadjusted estimator leads to a reduction in expected sample size of 20% in scenario (a), 19% in scenario (b), and 19% in scenario (c). Also, the maximum sample size is 1016 for the design using the adjusted estimator (TMLE $\text{prog}_{W,L}$), which is 20% less than the maximum sample size of 1272 for the unadjusted estimator. In scenario (c), the familywise Type I error rate (assuming no early stopping of accrual) is 0.025 for each estimator, as desired. Comparing the unadjusted estimator versus TMLE $\text{prog}_{\emptyset}$ shows that when W and L provide no prognostic information, the adjusted estimator is almost identical to the unadjusted estimator in power and expected sample size.

Figure 1 gives the power of each estimator at each stage. Plots (i) and (iv) of Figure 1

display power to reject at least H_{00} under scenario (a), and power to reject at least H_{01} under scenario (b), respectively. These plots demonstrate that goals (i) and (ii) from Section 6.1 are approximately achieved by all estimators. Plots (ii) and (iii) show that for each estimator, there is low power to reject at least H_{00} when only subpopulation 1 benefits, and to reject at least H_{01} when both subpopulations benefit. This behavior may be regarded as advantageous since it is ideal to reject only H_{00} in scenario (a) and only H_{01} in scenario (b), these corresponding precisely to the populations who benefit in each scenario, respectively.

The performance of the non-adaptive design defined in the last paragraph of Section 6.3 is shown in the bottom half of Table 4. The main difference between this design and the adaptive design from Section 6.3 is that the former has substantially larger expected sample size in scenarios (b) and (c). This is not surprising, since it is in these scenarios when futility stopping of subpopulation 2 is especially useful. The two designs have similar power and Type I error rate in all three scenarios, and similar expected sample sizes in scenario (a).

In Appendix F of the Supplementary Material, we compare the bias, variance, and mean squared error for the unadjusted versus adjusted estimators, and the adaptive versus non-adaptive designs. In each scenario (a)-(c), differences in the bias, variance, and mean squared error were negligible when comparing estimators or when comparing designs.

7. Discussion

Alternative methods exist for covariate adjustment in our longitudinal setting, e.g., the estimators of Lu and Tsiatis (2011); Rotnitzky *and others* (2012); Gruber and van der Laan (2012). These estimators have enhanced efficiency properties, but to the best of our knowledge there is not currently an R package implementing any of these methods that incorporates both baseline variables and short-term outcomes. The multiple testing procedure in Section 5 can also be applied for survival times, e.g., by using Wald statistics based on a modified TMLE or the

estimators of Lu and Tsiatis (2011); Parast *and others* (2014); Zhang (2015).

We assumed that the only cause of missing data was administrative censoring due to some participants not yet having their final outcomes observed. In Appendix B of the Supplementary Material, we describe how to incorporate additional right censoring due to loss to follow-up, under the missing at random assumption (van der Laan and Gruber, 2012).

Since the unadjusted estimator does not use information from pipeline participants, it has the independent increments covariance structure. In our simulation study, the covariance matrices and resulting boundaries $u_{j,k}$ from procedure M were quite similar for the unadjusted and adjusted estimators. We conjecture that the similarity was due to a relatively low proportion of pipeline participants at each interim analysis, and that there can be greater deviations from the independent increments covariance structure when there are larger proportions of pipeline participants, stronger correlations between Y and $(W, L^{(1)}, \ldots, L^{(T)})$, and more severe model misspecification. Before a trial starts, it may be difficult to predict how much the covariance matrix will deviate from the independent increments structure; therefore, it may be useful to have a general approach as here that strongly controls the familywise Type I error rate regardless of how large this deviation is.

The specific design in Section 6.3 stops all accrual at the first rejected null hypothesis, since this suffices to achieve the power goals (i)-(ii) from Section 6.1. However, it is possible that alternative designs, which involve rules for continuing accrual after a null hypothesis is rejected, may improve performance; this is an area of future research. The general multiple testing procedure M in Section 5 does not require stopping the trial at the first analysis where a null hypothesis is rejected, and can be used with any accrual modification rule $\mathbf{r}$. It may be possible to reallocate $\pi_{j,k}$ from null hypotheses that are rejected to other null hypotheses using the graphical approaches of Bretz *and others* (2011); this is an area of future work.

Our simulations involved two subpopulations. The general framework in Section 5 can be

applied to any number of subpopulations and composite populations. However, as the number of such populations increases, so will the required sample size to achieve high power for each population while controlling the familywise Type I error rate. It is an open problem to determine how many populations can be accommodated before sample size becomes prohibitively large.

The conditional error function approach has been applied to two-stage, adaptive enrichment designs by Friede *and others* (2012). They note that the required computations become more demanding for designs with more than two stages. In our context, one would need to compute conditional probabilities for a multivariate normal random vector. This is possible but more challenging than computing the (unconditional) multivariate normal distribution function, which can be done using the sophisticated algorithms of Genz *and others* (2014). However, the conditional error function approach has more flexibility in how the testing procedure can be modified, compared to our approach.

8. Supplementary Material

The Supplementary material includes the following: asymptotic results for the adjusted estimator; the TMLE implementation; the proofs of Theorems 5.1 and 5.2; details of the data generating distributions from Section 6; a bootstrap procedure to estimate $\mathbf{\Sigma}$; R code for our simulations. It is available here:

`http://people.csail.mit.edu/mrosenblum/papers/SuppMatAdaptWithDelay.pdf`

9. Acknowledgments

This research was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198) and the U.S. Food and Drug Administration (HHSF223201400113C). This paper's contents are solely the responsibility of the author and do not represent the views of these agencies.

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**(4), 962–973.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M. and Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* **28**(10), 1445–1463.

Bretz, Frank, Posch, Martin, Glimm, Ekkehard, Klinglmueller, Florian, Maurer, Willi and Rohmeyer, Kornelius. (2011). Graphical approaches for multiple comparison procedures using weighted bonferroni, simes, or parametric tests. *Biometrical Journal* **53**(6), 894–913.

Bretz, Frank, Schmidli, Heinz, König, Franz, Racine, Amy and Maurer, Willi. (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal* **48**(4), 623–634.

FDA. (2010). Draft guidance for industry. Adaptive design clinical trials for drugs and biologics. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf.

Freidlin, Boris, Sun, Zhuoxin, Gray, Robert and Korn, Edward L. (2013). Phase iii clinical trials that integrate treatment and biomarker evaluation. *Journal of Clinical Oncology* **31**(25), 3158–3161.

Friede, Tim, Parsons, N and Stallard, Nigel. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat. Med.* **31**(30), 4309–4320.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F. and Hothorn, T.

(2014). mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-0. URL http://CRAN.R-project.org/package=mvtnorm.

Gruber, S. and van der Laan, M.J. (2012). Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics* **8**(1).

Hanley, D. (2012). http://braininjuryoutcomes.com/studies/mistie/entry/mistie/international-stroke-conference-2012-mistie-phase-2-results.

Jenkins, M., Stone, A. and Jennison, C. (2011). An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm. Stat.* **10**(4), 347–356.

Jennison, Christopher and Turnbull, Bruce W. (1999). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC Press.

Jennison, Christopher and Turnbull, Bruce W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *J. of Biopharm. Stat.* **17**(6), 1135–1161.

Kim, Kyungmann and DeMets, David L. (1987). Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika* **74**(1), 149–154.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

Lu, X. and Tsiatis, A. A. (2011). Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring. *Lifetime data analysis* **17**(4), 566–593.

Magnusson, Baldur P. and Turnbull, Bruce W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine* **32**, 2695–2714.

Morgan, T., Zuccarello, M., Narayan, R., Keyl, P., Lane, K. and Hanley, D. (2008). Preliminary findings of the minimally-invasive surgery plus rtPA for intracerebral hemorrhage evacuation (MISTIE) clinical trial. *Acta Neurochir Suppl.* **105**, 147–51.

Parast, Layla, Tian, Lu and Cai, Tianxi. (2014). Landmark estimation of survival and treatment effect in a randomized clinical trial. *J. Amer. Statist. Assoc.* **109**(505), 384–394.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. In: *Proceedings of the American Statistical Association*.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology-Methodological Issues* **297–331**.

Rosenblum, M., Thompson, R. E., Luber, B. S. and Hanley, D. F. (2015). Adaptive group sequential designs that balance the benefits and risks of expanding inclusion criteria. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 250. http://biostats.bepress.com/jhubiostat/paper250*.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, JM. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**(2), 439–456.

Scharfstein, D.O., Tsiatis, A.A. and Robins, J.M. (1997). Semiparametric efficiency and its implications on the design and analysis of group-sequential studies. *J. Amer. Statist. Assoc.* **92**(440), 1342–1350.

Schmidli, H, Bretz, F, Racine, A and Maurer, W. (2006). Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* **48**(4), 635–643.

Schwab, J., Lendle, S., Petersen, M. and van der Laan, M. (2014). *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*. R package version 0.9.3-1.

Shoben, Abigail B. and Emerson, Scott S. (2014). Violations of the independent increment assumption when using generalized estimating equation in longitudinal group sequential trials. *Statistics in Medicine* **33**(29), 5041–5056.

Slud, E. V. and Wei, L-J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc.* **77**, 862–868.

Stallard, N. (2011). Group-sequential methods for adaptive seamless phase ii/iii clinical trials. *J. Biopharm. Stat.* **21**(4), 787–801.

Stallard, N., Hamborg, T., Parsons, N. and Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *J. Biopharm. Stat.* **24**(1), 168–187.

van der Laan, M.J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics* **2**(1).

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The International Journal of Biostatistics* **8**(9).

Wang, S. J., Hung, H. and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* **51**, 358–374.

Wang, S. J., O'Neill, R. T. and Hung, H. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subsets. *Pharmaceut. Statist.* **6**, 227–244.

Zhang, Min. (2015). Robust methods to improve efficiency and reduce bias in estimating survival curves in randomized clinical trials. *Lifetime Data Analysis* **21**(1), 119–137.

Table 1: Cumulative information at each analysis for subpopulation 1, subpopulation 2, and the combined population, in scenario (a).

	adjusted estimator					unadjusted estimator				
Analysis (k)	1	2	3	4	5	1	2	3	4	5
Subpop. 1 ($\mathcal{I}_{1,k}$)	124	250	370	558	749	126	251	376	590	795
Subpop. 2 ($\mathcal{I}_{2,k}$)	256	524	763			249	487	739		
Comb. Pop. ($\mathcal{I}_{0,k}$)	389	785	1140			372	740	1115		

Table 2: Cumulative sample size (Cum.S.S.) at each analysis, which has the format: number of participants with Y observed (+ number in pipeline).

Analysis (k)	1	2	3	4	5
			Unadjusted estimator		
Cum.S.S. Subpop. 1	104 (+24)	208 (+24)	312 (+24)	480 (+24)	648 (+0)
Cum.S.S. Subpop. 2	208 (+49)	416 (+49)	624 (+0)	624 (+0)	624 (+0)
Cum.S.S. Comb. Pop.	312 (+73)	624 (+73)	936 (+24)	1104 (+24)	1272 (+0)
			Adjusted estimator		
Cum.S.S. Subpop. 1	84 (+24)	168 (+24)	252 (+24)	382 (+24)	512 (+0)
Cum.S.S. Subpop. 2	168 (+49)	336 (+49)	504 (+0)	504 (+0)	504 (+0)
Cum.S.S. Comb. Pop.	252 (+73)	504 (+73)	756 (+24)	886 (+24)	1016 (+0)
Futility Boundary ($l_{1,k}$)	0	0	0	0	-
Futility Boundary ($l_{2,k}$)	0	0	∞	-	-

Table 3: Efficacy boundaries for scenario (a) and unadjusted estimator. H_{00} is no longer tested after analysis 3.

Analysis (k)	1	2	3	4	5
H_{00} Efficacy Boundary ($u_{0,k}$)	3.41	3.06	2.84	-	-
H_{01} Efficacy Boundary ($u_{1,k}$)	3.27	2.89	2.66	2.33	2.14

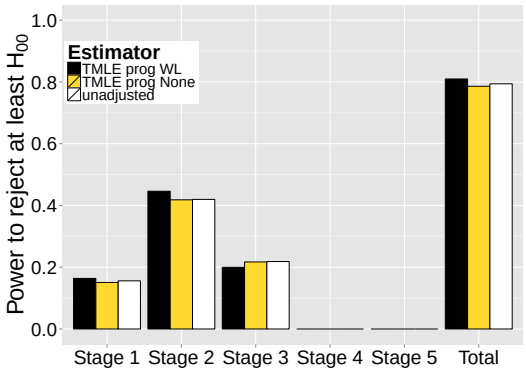
Table 4: Power and expected sample size (ESS) for adaptive and non-adaptive designs. Power under scenario (a) is the probability of rejecting at least H_{00}; power under scenario (b) is the probability of rejecting at least H_{01}.

Adaptive Design

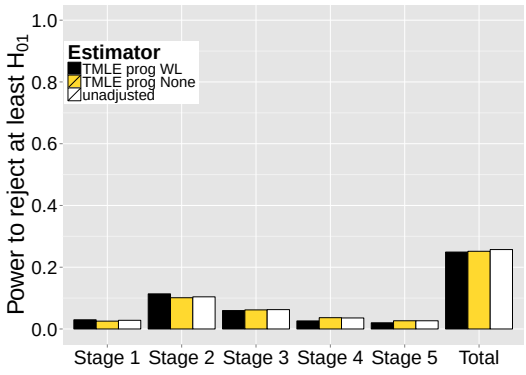
	Scenario (a)		Scenario (b)		Scenario (c)	
	power H_{00}	ESS	power H_{01}	ESS	Type I error	ESS
Estimator:						
unadjusted	0.79	712	0.82	795	0.025	640
TMLE $\text{prog}_{W,L}$	0.81	568	0.82	643	0.025	521
TMLE $\text{prog}_{\emptyset}$	0.79	711	0.82	794	0.025	638

Non-Adaptive Design

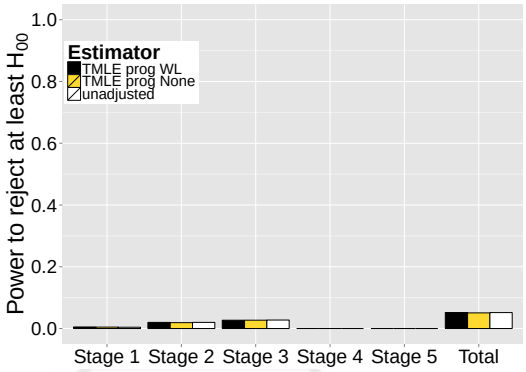
	Scenario (a)		Scenario (b)		Scenario (c)	
	power H_{00}	ESS	power H_{01}	ESS	Type I error	ESS
Estimator:						
unadjusted	0.80	718	0.82	958	0.025	729
TMLE $\text{prog}_{W,L}$	0.82	575	0.82	771	0.025	591
TMLE $\text{prog}_{\emptyset}$	0.80	718	0.81	959	0.025	727



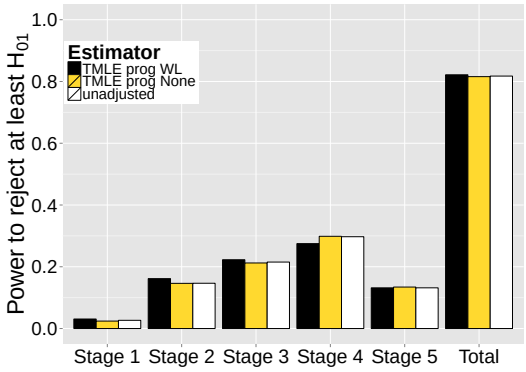
(i) Power to Reject at Least H_{00} in Scenario (a)



(ii) Power to Reject at Least H_{01} in Scenario (a)



(iii) Power to Reject at Least H_{00} in Scenario (b)



(iv) Power to Reject at Least H_{01} in Scenario (b)

Fig. 1: Stagewise and overall power comparing estimators. Top and bottom rows correspond to scenarios (a) and (b), respectively. Left and right columns represent power to reject at least H_{00} and to reject at least H_{01}, respectively. Black bar represents TMLE $\text{prog}_{W,L}$; yellow bar represents TMLE $\text{prog}_{\emptyset}$ (denoted "prog None" in legends); white bar represents unadjusted estimator.