Johns Hopkins University, Dept. of Biostatistics Working Papers

8-1-2016

SENSITIVITY OF TRIAL PERFORMANCE TO DELAY OUTCOMES, ACCRUAL RATES, AND PROGNOSTIC VARIABLES BASED ON A SIMULATED RANDOMIZED TRIAL WITH ADAPTIVE ENRICHMENT

Tiachen Qian
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics, tqian2@jhu.edu

Elizabeth Colantuoni
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Aaron Fisher
Johns Hopkins Bloomberg School of PublicHealth, Department of Biostatistics

Michael Rosenblum
Johns Hopkins Bloomberg School of Public Health, Department of Biostatistics

Sensitivity of Trial Performance to Delayed Outcomes, Accrual Rates, and Prognostic Variables based on a Simulated Randomized Trial with Adaptive Enrichment

Tianchen Qian*†, Elizabeth Colantuoni†, Aaron Fisher†, Michael Rosenblum†

for the Alzheimer's Disease Neuroimaging Initiative‡

July 27, 2016

Abstract

Adaptive enrichment designs involve rules for restricting enrollment to a subset of the population during the course of an ongoing trial. This can be used to target those who benefit from the experimental treatment. To leverage prognostic information in baseline variables and short-term outcomes, we use a semiparametric, locally efficient estimator, and investigate its strengths and limitations compared to standard estimators. Through simulation studies, we assess how sensitive the trial performance

*Corresponding author email address: tqian2@jhu.edu

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

‡Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

(Type I error, power, expected sample size, trial duration) is to different design characteristics. Our simulation distributions mimic features of data from the Alzheimer's Disease Neuroimaging Initiative, and involve two subpopulations of interest based on a generic marker. We investigate the impact of the following design characteristics: the accrual rate, the delay time between enrollment and observation of the primary outcome, and the prognostic value of baseline variables and short-term outcomes. We apply information-based monitoring, and evaluate how accurately information can be estimated in an ongoing trial.

Keywords: multiple testing procedure; treatment effect heterogeneity

1 Introduction

Adaptive enrichment designs involve pre-planned rules for restricting enrollment based on accrued data in an ongoing trial (Wang et al., 2007). If, for example, a subpopulation shows evidence of no benefit of treatment, its enrollment could be stopped while the complementary subpopulation continues to be enrolled. Stallard et al. (2014) give an overview of statistical methods for adaptive enrichment designs, including the p-value combination approach (Bretz et al., 2006; Schmidli et al., 2006; Jennison and Turnbull, 2007; Brannath et al., 2009); the conditional error function approach (Friede et al., 2012); and approaches using group sequential computations (Stallard, 2011; Magnusson and Turnbull, 2013). We use an adaptive enrichment design from the general class of Rosenblum et al. (2016), which is based on the group sequential computation approach.

We consider trials where the primary outcome is observed a fixed amount of time from enrollment (called the delay). To illustrate, we use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. We set the primary outcome to be a measure of change in severity of dementia symptoms from baseline to 1 year of follow-up described below; this is similar to the primary outcome in an ongoing, phase 3 clinical trial of a drug to slow cogni-

tive and functional decline from early Alzheimer's Disease (Biogen, 2016). Also recorded are baseline variables and the short-term outcome of change in severity of dementia symptoms measured at 6 months of follow-up.

To leverage prognostic information in baseline variables and the short-term outcome, we use a semiparametric, locally efficient estimator (called the adjusted estimator, for conciseness) from van der Laan and Gruber (2012). The adjusted estimator in a randomized trial is consistent under mild regularity conditions without requiring any parametric model assumptions. It has potential to improve precision, power, expected sample size, and trial duration when variables are sufficiently prognostic for the outcome. In trials with delayed outcomes, the adjusted estimator uses information from pipeline participants, i.e., enrollees whose primary outcome has not yet been observed.

An open question is how useful the above estimators are in adaptive enrichment designs with delayed outcomes, under different configurations of delay, accrual rates and prognostic value. We use simulation studies that mimic features of data from the ADNI study, and examine the impact of delay, accrual rates, prognostic baseline variables, and prognostic short-term outcomes.

The simulated trials involve multiple stages, and information-based monitoring is used to determine the time of interim analyses. We evaluate the accuracy of information estimates when using the adjusted estimator versus the unadjusted estimator, which is critical in order that the familywise Type I error be controlled.

In Section 2 we describe the ADNI study. In Section 3 we present notation. The simulation setup is given in Section 4. Section 5 presents simulation results, including the impact of prognostic baseline variables and a short-term outcome (Section 5.1), the impact of varying delay time (Section 5.2), and the impact of varying the accrual rates (Section 5.3) on the performance of the adaptive design. In Section 6 we discuss information accrual rates and

how accurately these can be estimated in an ongoing trial. Section 7 discusses limitations and future research directions.

2 Data Example

Our simulations are based on distributions that mimic features of the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), an observational longitudinal study of cognitive impairment and progression to Alzheimer's disease. The ADNI was initiated in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of the study has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early Alzheimer's disease.[1] The Clinical Dementia Rating (CDR) scale is used to assess the severity of dementia symptoms and provides both a numeric global score ranging from 0 to 3, and a sum of boxes (SOB) score ranging from 0 to 18.

Our data come from 286 patients who entered the ADNI study with mild cognitive impairment (CDR 0.5 with a SOB score 2.5 or less) and who remained in the study for the full 12 months of follow-up. For conciseness, we refer to the sum of the CDR global score and the SOB score as the CDR score. We define the primary outcome Y as the difference between the CDR score at baseline and at 12 months. We define the short-term outcome L as the difference between the CDR score at baseline and at 6 months. Let W denote the following five prognostic baseline variables: CDR score at baseline; age; $A\beta_{42}$ (a type of amyloid plaque involved in Alzheimer's disease progression); Alzheimer's Disease Association (ADA, 13 items) scale; and the Mini Mental State Examination (MMSE) score. We consider two distinct subpopulations defined by apolipoprotein E (APOE) $\epsilon 4$ carrier

[1] For up-to-date information, see www.adni-info.org.

status. Subpopulation 1 consists of those with no $\epsilon 4$ alleles, and subpopulation 2 consists of those with at least one $\epsilon 4$ allele. (no alleles being subpopulation 1 vs. at least one allele being subpopulation 2). Among the 286 patients, 47% carry no APOE $\epsilon 4$ alleles. We consider a hypothetical treatment whose goal is to delay the progression of disease. Since we had more measurements at 12 and 24 months in the dataset, but we wanted to use the timescale of 6 and 12 months in our simulated trial, we mapped each 12- and 24-month outcome to 6-and 12-month, respectively, throughout our paper.

3 Notation

When followed up completely, each participant i in the trial has full data vector $\mathbf{D}_i = (S_i, W_i, A_i, L_i, Y_i)$. We use the vector $\mathbf{D} = (S, W, A, L, Y)$ when referring to a generic participant. The variable $S_i \in \{1, 2\}$ denotes the subpopulation that participant i belongs to; W_i denotes a vector of baseline variables; A_i denotes the treatment assignment indicator; L_i denotes the short-term outcome; and Y_i denotes the primary outcome. We assume that (S_i, W_i, A_i) are observed when participant i is enrolled, and that L_i and Y_i are observed at duration d_L and d_Y, respectively, from the time of enrollment, with $d_L \leq d_Y$. Each vector $\mathbf{D}_i$ is assumed to be an independent, identically distributed draw from an unknown distribution Q, with the only restriction being that A is randomized by design with equal probability of being 0 or 1, independent of S, W. The short-term outcome L can be any predefined measurement made after randomization. No assumptions on its relationship to Y are needed in order that our estimators (adjusted and unadjusted) are consistent and asymptotically normal.

For a given population, the average treatment effect is defined to be the difference between the population mean of the primary outcome under treatment ($A = 1$) versus control ($A = 0$). Denote the average treatment effect in subpopulation 1, subpopulation 2, and the combined

population by Δ_1, Δ_2, and Δ_0, respectively, where $\Delta_0 = E(Y|A=1) - E(Y|A=0)$ and for each subpopulation $s \in \{1,2\}$, $\Delta_s = E(Y|A=1,S=s) - E(Y|A=0,S=s)$. Let p_s denote the proportion of subpopulation s in the combined population. Then $\Delta_0 = p_1\Delta_1 + p_2\Delta_2$. Define the null hypotheses

$$H_{01} : \Delta_1 \leq 0; \qquad H_{02} : \Delta_2 \leq 0; \qquad H_{00} : \Delta_0 \leq 0,$$

which represent no average treatment benefit in subpopulation 1, subpopulation 2, and the combined population, respectively.

We quantify the prognostic value of W and L for explaining variance in the primary outcome Y for the combined population. Define the R-squared of W and R-squared of L as

$$R^2_W = \frac{\text{var}\{E(Y \mid W)\}}{\text{var}(Y)}, \quad R^2_L = \frac{\text{var}\{E(Y \mid L)\}}{\text{var}(Y)}. \tag{1}$$

R^2_W represents the fraction of variance in Y explained by W. Similarly, R^2_L represents the fraction of variance in Y explained by L.

Using the ADNI study data, we approximated (1) to roughly determine how much of the variance of the outcome Y is explained by W or L. The empirical R^2_W is computed as in (1), with $E(Y \mid W)$ estimated by a linear model with intercept and main terms W_3, W_4, and the variances are estimated by the empirical variance. (We use only W_3, W_4 in the working model for constructing the adjusted estimator; see Section 4.2.) A similar computation was done to obtain the empirical R^2_L replacing W by L. The resulting values are 0.20 and 0.48 for R^2_W and R^2_L, respectively. Roughly speaking, this indicates both variables are moderately to strongly prognostic for Y.

We estimated R^2_W and R^2_L within each subpopulation, and found the prognostic values differ by subpopulation. The corresponding empirical R^2_W is 0.30 for subpopulation 1 and

0.14 for subpopulation 2; the empirical R^2_L is 0.44 for subpopulation 1 and 0.50 for subpopulation 2. This differential prognostic value by subpopulation impacts information accrual and power for the adjusted estimator as described in Section 5. In what follows, R^2_W and R^2_L refer to (1) for the combined population.

4 Simulation Setup

4.1 Overview

Our goal is to evaluate the performance of an adaptive enrichment design with a delayed outcome when we vary the prognostic values in baseline variables and short-term outcome, accrual rates, delay time, and estimator used. The performance is evaluated based on Type I error, power, expected sample size and average duration of the trial, and is based on two estimators: the unadjusted estimator (the difference between the sample means of the primary outcome between the two study arms), and an adjusted estimator that leverages baseline variables and the short-term outcome. The latter is a targeted maximum likelihood estimator (TMLE) of van der Laan and Gruber (2012) implemented in the `R` package `ltmle` (Schwab et al., 2015). The R code we used for the adjusted estimator is provided in the Supplementary Materials. We also could have used adjusted estimators such as those of Lu and Tsiatis (2011); Rotnitzky et al. (2012); Gruber and van der Laan (2012). Both the unadjusted and adjusted estimators are consistent and asymptotically normal under mild regularity conditions (van der Laan and Gruber, 2012).

We vary the following in our simulation studies: the prognostic value of baseline variables W and short-term outcome L represented by the R-squared formulas in Section 3; the delay time d_L to observe the short-term outcome; the delay time d_Y to observe the final outcome; and the accrual rate.

4.2 Data Generating Distributions Based on ADNI Data

Hypothetical trials are populated with participants, each of whose data vector $\mathbf{D}$ is drawn independently from a data generating distribution Q, which may differ by simulation study. We constructe each Q to mimic certain observed relationships between W, L and Y within each subpopulation $s \in \{1, 2\}$ in the ADNI study. For simplicity, we center W within each subpopulation S.

Since there is no treatment in the ADNI study, we assign the treatment variable A independent of S, W, and having a relationship with Y as described next. The minimum, clinically meaningful, average treatment effect for our hypothetical trials is $\delta_{\min} = 0.42$, which corresponds to a 30% relative improvement in mean CDR score change, i.e., a 30% reduction in disease progression. Within each of our five simulation studies (described below), we generate data under four treatment effect settings (abbreviated as "effect setting" hereafter): (a) treatment benefits neither subpopulation ($\Delta_1 = \Delta_2 = 0$); (b) treatment benefits subpopulation 1 only ($\Delta_1 = \delta_{\min}, \Delta_2 = 0$); (c) treatment benefits subpopulation 2 only ($\Delta_1 = 0, \Delta_2 = \delta_{\min}$); and (d) treatment benefits both subpopulations ($\Delta_1 = \Delta_2 = \delta_{\min}$). Effect settings (b) and (c) involve treatment effect heterogeneity.

The data generating distribution in each set of simulations is denoted by

$$Q = Q\left(\Delta_1, \Delta_2, R^2_W, d_Y, \text{accrual rate}, L \text{ measured}, R^2_L, d_L\right),$$

and is determined by the following: the pair of treatment effects for each subpopulation (Δ_1, Δ_2); the prognostic value of the baseline covariates R^2_W; the delay between enrollment and the primary outcome d_Y; the accrual rate; whether the short-term outcome L is measured and if so, its prognostic value R^2_L and delay time d_L from enrollment. We set the enrollment process to be random, where the enrollment time of the patients follows a ho-

mogeneous Poisson process with intensity equal to the accrual rate. We assume that each subpopulation's accrual rate is proportional to its prevalence in the combined population. In each simulation study, we vary one or several of the above at a time to assess the impact on trial performance.

First, consider the case where the short-term outcome L is not measured. Within each subpopulation $S = s$, Y is drawn from the linear model:

$$Y = \beta_0^s + \beta_W^s W + \beta_A^s A + \epsilon_Y, \qquad \epsilon_Y \sim N\left(0, (\sigma_Y^s)^2\right) \tag{2}$$

with ϵ_Y independent of (W, A). The values β_0^s, β_W^s and σ_Y^s are based on the above model fit to the ADNI study data separately within each stratum $S = s$ and leaving out A. We set $\beta_A^s = \Delta_s$ to be the desired treatment effect, which depends on the effect settings (a)-(d).

For the case where L is measured, within each subpopulation $S = s$, Y and L are generated from the linear models:

$$L = \alpha_0^s + \alpha_W^s W + \alpha_A^s A + \epsilon_L, \qquad \epsilon_L \sim N\left(0, (\sigma_L^s)^2\right) \tag{3}$$

$$Y = \beta_0^s + \beta_W^s W + \beta_A^s A + \beta_L^s L + \epsilon_Y, \qquad \epsilon_Y \sim N\left(0, (\sigma_Y^s)^2\right) \tag{4}$$

with ϵ_Y and ϵ_L independent of (W, A) and of each other. The values of β_0^s, β_W^s, β_L^s, σ_Y^s, α_0^s, α_W^s and σ_L^s are based on the above models fit to the ADNI study data separately within each stratum $S = s$ and leaving out A. For simplicity we set $\alpha_A^s = 0$, and set $\beta_A^s = \Delta_s$ to be the desired treatment effect. The β values in (4) are not the same as those in (2); however, we do not distinguish between these in our notation, when there is no ambiguity as to whether L is measured in the simulation scenario.

We construct simulation distributions with a range of R_W^2 and R_L^2 values by varying α and β. We do so in such a way that the average treatment effect within each subpopulation

is unchanged, and the variance of Y within each subpopulation and each treatment arm is unchanged; our method is summarized below with details given in the Supplementary Material. The result is that the (asymptotic) performance of the unadjusted estimator is unchanged, providing a benchmark to compare against. In simulation scenarios where L is not measured, to change the prognostic value of W we multiply the original fits of β_W^1, β_W^2 from the ADNI study data by a tuning parameter p_W in (2), and change σ_Y^s accordingly so that the variance of Y given A, S and the average treatment effect given S are unchanged. In simulation scenarios where L is measured, to change the prognostic value of L, we multiply β_L^1, β_L^2 by p_L in (4) and change σ_Y^s accordingly to ensure that the variance of Y given A, S and the average treatment effect given S are unchanged. These modifications do not affect the unadjusted estimator, although they do impact the adjusted estimator (as we will show in Section 5).

Let the default simulation scenario be the one with design characteristics corresponding to the empirical distribution of the ADNI study data: $R_W^2 = 0.20$, $R_L^2 = 0.48$; the default scenario sets $d_L = 0.5$ years, $d_Y = 1$ year, and the accrual rate for the combined population to be 334 patients/year. We conduct 5 sets of simulations with various design characteristics that are summarized in Table 1. Each combination of $(R_W^2, d_Y, \text{accrual rate}, L \text{ measured}, R_L^2, d_L)$ is referred to as a simulation scenario. For example, in simulation study 1 (row 1 in Table 1), R_W^2 is varied from 0 to 0.6, the short-term outcome is not measured, and all other characteristics are set to the default value.

In all simulations, we use the full set of baseline covariates $(W_1, W_2, W_3, W_4, W_5)$ in the data generating distributions (2)-(4) for L and Y, but we only include baseline variables W_3, W_4 ($A\beta_{42}$ and ADA) in the working models used by the adjusted estimator. We intentionally induced such model misspecification, since in practice the working models used by the adjusted estimator will generally be misspecified. In addition, the TMLE estimator

Table 1: Summary of setups for 5 simulation studies. Default value of parameter: $R^2_W = 0.20$, $R^2_L = 0.48$, $d_L = 0.5$ years, $d_Y = 1$ year, accrual rate 334 patients/year. Ranges of values $x - y$ indicate the design characteristic(s) varied in the corresponding simulation study.

Simulation study	R^2_W	d_Y (yrs)	accrual rate (patients/year)	L measured	R^2_L	d_L (yrs)
1	$0 - 0.6$	default	default	No	NA	NA
2	0	default	default	Yes	$0 - 0.6$	default
3	default	$0 - 2$	default	No	NA	NA
4	default	$0.05, 0.5, 1, 1.5, 2$	default	Yes	default	$0 - d_Y$
5	default	default	$100 - 1000$	Yes	default	default

uses logistic regression working models (by first scaling the outcome to the interval $[0, 1]$) rather than linear models, which can lead to additional misspecification. Though the adjusted estimator is robust to the above model misspecification in that it is still consistent and asymptotically normal, the misspecification may reduce its precision.

4.3 Adaptive Enrichment Design

We define a new adaptive enrichment design using the general framework developed by Rosenblum et al. (2016). We consider two subpopulations: $S = 1$ if the patient has no APOE $\epsilon 4$ allele, and $S = 2$ if the patient has one or more APOE $\epsilon 4$ allele. Define $S = 0$ to be the combined population. We consider an adaptive enrichment design with maximum number of stages $K = 5$. At each analysis $k \leq K$, let $Z_{s,k}$ denote the Wald statistic (estimator divided by its standard error) for null hypothesis H_{0s}, $s \in \{0, 1, 2\}$. For each population s and stage $k \leq K$, let $u_{s,k}$ denote the efficacy boundary for the null hypothesis H_{0s} ($s \in \{0, 1, 2\}$), and let $l_{s,k}$ denote the futility stopping boundary ($s \in \{1, 2\}$). The multiple testing procedure at each analysis $k \leq K$ consists of the following steps:

1. For each $s \in \{1, 2\}$, if subpopulation s has not had enrollment stopped at a previous analysis, and if $Z_{s,k} > u_{s,k}$, reject H_{0s}.

2. For each $s \in \{1, 2\}$, if H_{0s} is rejected or $Z_{s,k} < l_{s,k}$, stop subpopulation s enrollment.

3. If both H_{01} and H_{02} are rejected, or (if both subpopulations have not had enrollment stopped at a previous analysis and $Z_{0,k} > u_{0,k}$), reject H_{00}.

The trial continues until both subpopulations terminate enrollment or the final analysis K is reached.

Define the power of H_{01} to be the probability to reject at least H_{01} under effect setting (b), power of H_{02} to be the probability to reject at least H_{02} under effect setting (c), and power of H_{00} to be the probability to reject at least H_{00} under effect setting (d). The design's goals are to achieve at least 80% power to reject the corresponding null hypothesis under each effect setting (b), (c), and (d), and to strongly control the familywise Type I error rate at level 0.025, asymptotically. For example, the requirement under effect setting (b) is 80% power for H_{01}.

The Type I error spent at each stage, futility boundaries $l_{s,k}, s \in \{1, 2\}, 1 \leq k \leq K$ and the information level (inverse of the estimator's variance) used for analysis timing are in Table 2. They were constructed by approximately solving the following optimization problem: for the unadjusted estimator under the default simulation scenario, minimize the expected sample size averaged over effect settings (a)-(d), subject to the Type I error and power constraints in the previous paragraph. The optimization was solved using an approach from Fisher and Rosenblum (2016), and does not necessarily equal the true optimum solution (which is currently an open research question). The asymmetry in the solution is because the proportion $p_1 = 0.47$ and the variances differ by subpopulation. In determining the values of efficacy boundaries $u_{s,k}, s \in \{0, 1, 2\}, 1 \leq k \leq K$, we use the error spending approach as described in Rosenblum et al. (2016, Section 3.2), which extends the approach of Slud and Wei (1982); Lan and DeMets (1983) to multiple populations; see the Supplementary Material for details. These efficacy boundaries depend on the covariance matrix of the

estimator being used. The design is guaranteed to strongly control the familywise Type I error rate at level 0.025, asymptotically, for Wald statistics based on either the unadjusted or adjusted estimators.

Table 2: Adaptive enrichment design, and efficacy boundaries under default simulation scenario.

Analysis (k)	1	2	3	4	5
Type I error spent for Subpop. 1	0.0007	0.0007	0.0028	0.0015	0.0038
Type I error spent for Subpop. 2	0.0001	0.0023	0.0012	0.0026	0.0027
Type I error spent for Comb. Pop.	0.0028	0.0006	0.0009	0.0013	0.0012
Futility boundary ($l_{1,k}$)	-4.12	0.40	-1.48	0.94	-
Futility boundary ($l_{2,k}$)	-0.10	0.29	0.42	0.93	-
Information threshold for Subpop. 1	13.0	20.2	24.9	40.1	69.1
Information threshold for Subpop. 2	13.4	20.2	25.7	41.1	69.6
Information threshold for Comb. Pop.	27.1	40.8	50.1	80.3	138.5
Efficacy boundaries for the unadjusted estimator under default simulation scenario					
Effcacy boundary ($u_{1,k}$)	3.12	3.06	2.64	2.77	2.53
Effcacy boundary ($u_{2,k}$)	3.52	2.76	2.78	2.63	2.62
Effcacy boundary ($u_{0,k}$)	2.78	3.08	2.92	2.86	2.89

4.4 Analysis Timing and Information Accrual

We present our method to determine the time of each analysis based on information monitoring. Consider either the adjusted or unadjusted estimator. There are 3 populations of interest (the two subpopulations and the combined population) in our design. For each population there is a treatment effect estimator whose variance changes over time as patients are continuously enrolled. We define the information accrued for each population as the reciprocal of the corresponding estimator's variance. The kth analysis occurs at the earliest time when the information accrued for every population is above its corresponding, preset threshold (which is a preset function of the Type I error allocated at that stage, i.e., part of

the trial design). Information thresholds in the design, shown in Table 2, were set such that for the unadjusted estimator in the default simulation scenario, the information accrual for each population crosses its threshold at the same calendar time. Information can accrue at different rates depending on whether the unadjusted or adjusted estimator is used, as shown in our simulations. Faster information accrual can lead to earlier analyses in calendar time and usually smaller sample size at each analysis.

Since in practice the variance of each estimator is unknown, one could use a variance estimator that is updated whenever new data accrues. (See Section 6 where we investigate the accuracy of information estimation at given time points.) However, it is not computationally feasible to implement this in our simulations where each data generating distribution is used to simulate 10,000 trials. Instead, we set analysis timing once for each simulation scenario and estimator type, using an approximation described in the Supplementary Material.

Table 3 shows the calendar times of each analysis for the unadjusted and the adjusted estimators under the default simulation scenario. The cumulative sample size at each analysis time is random due to the random accrual process; Table 3 is an example realization. Time of analysis and sample sizes are substantially smaller for the adjusted estimator compared to the unadjusted due to the former having a faster information accrual rate.

5 Results

We simulated 10,000 trials for each simulation scenario and effect setting combination. Table 4 shows the empirical probability of rejecting each hypothesis under the four effect settings in the default simulation scenario. The numbers with * indicate Type I error, i.e., rejecting at least one true null hypothesis. Under effect setting (a), all null hypotheses are true; under effect setting (b) (or (c)), only H_{01} (or H_{02}) is true; under effect setting (d), none of the null hypotheses are true.

Table 3: Calendar time to conduct interim analysis for unadjusted and adjusted estimators under default simulation scenario. For one realization of the trial we show the cumulative sample size (CSS) with the format: number of participants with Y observed (+ number of pipeline participants). If no early stop occurs, "stop enroll" column shows the time of last participant enrolled, and we wait until all participants have Y observed then conduct the final analysis (analysis 5).

Analysis (k)	1	2	3	4	stop enroll	5 (final)
			Unadjusted estimator			
Time (years)	2.2	2.8	3.2	4.5	6.1	7.1
CSS (Subpop. 1)	108 (+245)	202 (+252)	275 (+248)	479 (+237)	730 (+238)	968 (+0)
CSS (Subpop. 2)	104 (+290)	241 (+257)	309 (+256)	528 (+248)	788 (+275)	1063 (+0)
CSS (Comb. Pop.)	212 (+535)	443 (+509)	584 (+504)	1007 (+485)	1518 (+513)	2031 (+0)
			Adjusted estimator			
Time (years)	1.8	2.3	2.7	3.8	5.1	6.1
CSS (Subpop. 1)	54 (+251)	137 (+238)	192 (+245)	375 (+236)	585 (+222)	807 (+0)
CSS (Subpop. 2)	53 (+276)	133 (+283)	223 (+257)	415 (+258)	626 (+242)	868 (+0)
CSS (Comb. Pop.)	107 (+527)	270 (+521)	415 (+502)	790 (+494)	1211 (+464)	1675 (+0)

Across all the simulation scenarios we considered, the familywise Type I error rate was always controlled at 0.025 for both adjusted and unadjusted estimators. All the power goals in Section 4.3 are met. For the unadjusted estimator, the powers of H_{00}, H_{01} and H_{02} (defined in Section 4.3) are all between $80\% - 83\%$ under different simulation scenarios. This is as expected due to our method of determining the analysis timing (as described in Section 4.4). For the adjusted estimator, the power of H_{02} also stays near 80% under different simulation scenarios, whereas the power of H_{01} and H_{02} can be much higher than 80% under certain simulation scenarios, e.g. when the prognostic value in W is considerably high ($R^2_W > 0.3$). This is because R^2_W is always higher in subpopulation 1 than in subpopulation 2 due to the way we vary p_W in Section 4.2. (See Section A.3 for a more detailed discussion.) If one intended to have exactly 80% power for all three hypotheses for the adjusted estimator, we could have optimized a separate adaptive design for the adjusted estimator to incorporate the different R^2_W in two subpopulations. However, this would make it harder to do a head-to-head comparison of the unadjusted and the adjusted estimators.

Table 4: Type I error / power for two estimators under default simulation scenario. Type I errors (numbers with *) are computed assuming nonbinding futility boundaries; powers are computed assuming binding futility boundaries. In "Percent probability to reject", to reject an individual hypothesis means to reject at least that hypothesis; All/Any means to reject all/any of the three hypotheses. The empirical values corresponding to the power requirements are in bold for each scenario (b)-(d).

	Effect setting	Percent probability to reject				
		H_{00}	H_{01}	H_{02}	All	Any
Adjusted estimator	(a) $\Delta_1 = \Delta_2 = 0$	0.7*	1.0*	1.1*	0.0*	2.5*
	(b) $\Delta_1 = \delta_{\min}, \Delta_2 = 0$	12	**87**	1.1*	1.0*	88
	(c) $\Delta_1 = 0, \Delta_2 = \delta_{\min}$	16	1.1*	**80**	0.9*	80
	(d) $\Delta_1 = \Delta_2 = \delta_{\min}$	**83**	88	80	70	98
Unadjusted estimator	(a) $\Delta_1 = \Delta_2 = 0$	0.6*	1.0*	1.1*	0.0*	2.5*
	(b) $\Delta_1 = \delta_{\min}, \Delta_2 = 0$	12	**82**	1.0*	0.9*	82
	(c) $\Delta_1 = 0, \Delta_2 = \delta_{\min}$	15	1.1*	**81**	1.0*	81
	(d) $\Delta_1 = \Delta_2 = \delta_{\min}$	**82**	81	81	66	97

In what follows, we focus on comparing the expected sample size (ESS) and the expected duration (ED) as summaries of trial performance under different simulation scenarios and between the two estimators.

5.1 Simulation Studies 1-2: Effect of Prognostic Value of Baseline Variables and Short-term Outcome

Figure 1 illustrates how ESS and ED are affected when one of R^2_W or R^2_L varies. The performance of the unadjusted estimator remains the same when the prognostic value in W and L changes, providing a benchmark to compare with. The adjusted estimator performs similar to the unadjusted when there is no prognostic value in W or L, i.e. $R^2_W = R^2_L = 0$. As R^2_W or R^2_L increases, the adjusted estimator leverages this to achieve faster information accrual and fewer participants per stage, which leads to smaller ESS and ED. In simulation study 1, R^2_W is varied from 0 to 0.6; in simulation study 2, R^2_L is varied from 0 to 0.6 (Table 1).

Our results indicate that for the adjusted estimator, a prognostic baseline variable is more valuable than an equally prognostic short-term outcome in terms of reducing ESS and ED. For instance, under effect setting (d), increasing R^2_W from 0 to 0.25 results in a 19% drop in ESS, whereas increasing R^2_L from 0 to 0.25 only renders a 2% drop. This is because all enrolled patients' baseline variables contribute to the precision of the adjusted estimator; however, although the short-term outcome of every participant is used, the efficiency gain from adjusting for L is proportional to the number of participants in the pipeline (i.e., those who have L but not Y observed). Moreover, a participant's baseline variables potentially improve precision for estimation of both $E(Y|A=1)$ and $E(Y|A=0)$, while a participant's short-term outcome is only used toward improving precision for one of these, corresponding to the treatment that participant received.

5.2 Simulation Studies 3-4: Effect of Delay Times d_Y and d_L

We assess the impact of delay times d_Y and d_L on the performance of the design. In simulation study 3, we vary d_Y from 0 years (immediate Y) to 2 years with L not measured. In simulation study 4, with L measured we set d_Y to several levels, and in each case vary d_L from 0 (immediate L) to d_Y.

Figure 2 shows the comparison under simulation study 3. ESS and ED increase with longer d_Y for both estimators. This is intuitive: the longer it takes to observe the primary outcome, the more time is needed to accumulate the necessary information. The adjusted estimator leads to smaller ESS and ED than the unadjusted estimator uniformly over all values of d_Y because of gains from adjusting for baseline variables W. In addition, ESS and ED for both estimators are approximately linear in d_Y.

Figure 3 shows the comparison under simulation study 4. When d_Y is fixed, the performance of the unadjusted estimator remains the same regardless of the length of d_L, because

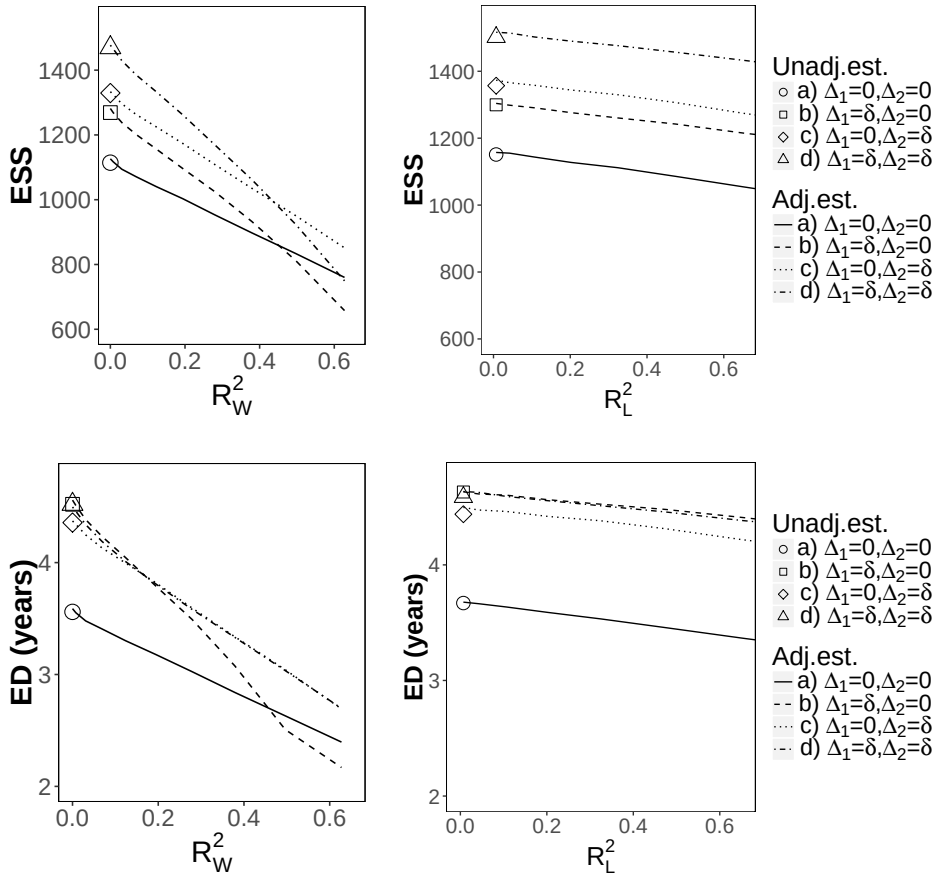


Figure 1: Left: impact of R^2_W on ESS and duration in simulation study 1. Right: impact of R^2_L on ESS and duration in simulation study 2. Since the results corresponding to unadjusted estimator do not change as R^2_W and R^2_L are varied, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols. δ refers to $\delta_{\min}$.

L is not used in the unadjusted estimator. For the adjusted estimator, a longer d_L results in a smaller proportion of pipeline participants who have L observed—hence, slower information accrual and larger ESS and ED. Even when $d_L = d_Y$, which implies no asymptotic precision gain from adjusting for L, the adjusted estimator still gains from adjusting for prognostic W.

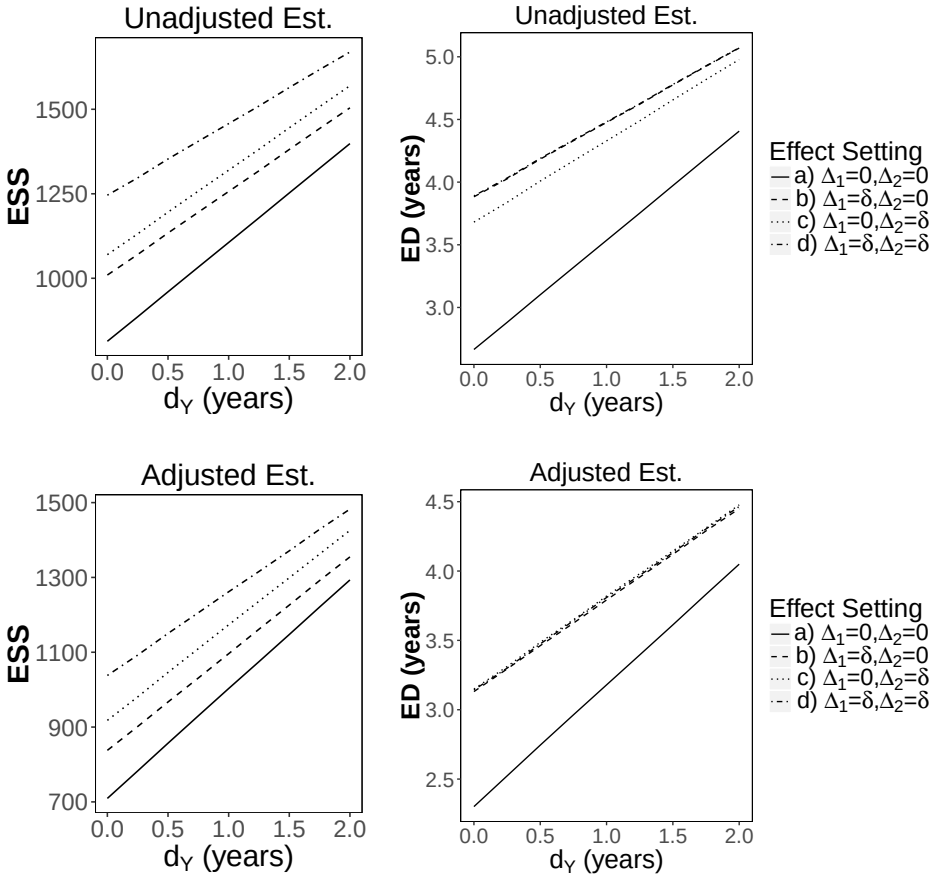


Figure 2: Impact of d_Y on ESS and ED in simulation study 3. Different line types indicate the ESS and ED under four effect settings. For the adjusted estimator, the lines for ED under effect settings (b)-(d) are clustered together. For the unadjusted estimator, the lines for ED under effect settings (b) and (c) are clustered together. δ refers to $\delta_{\min}$.

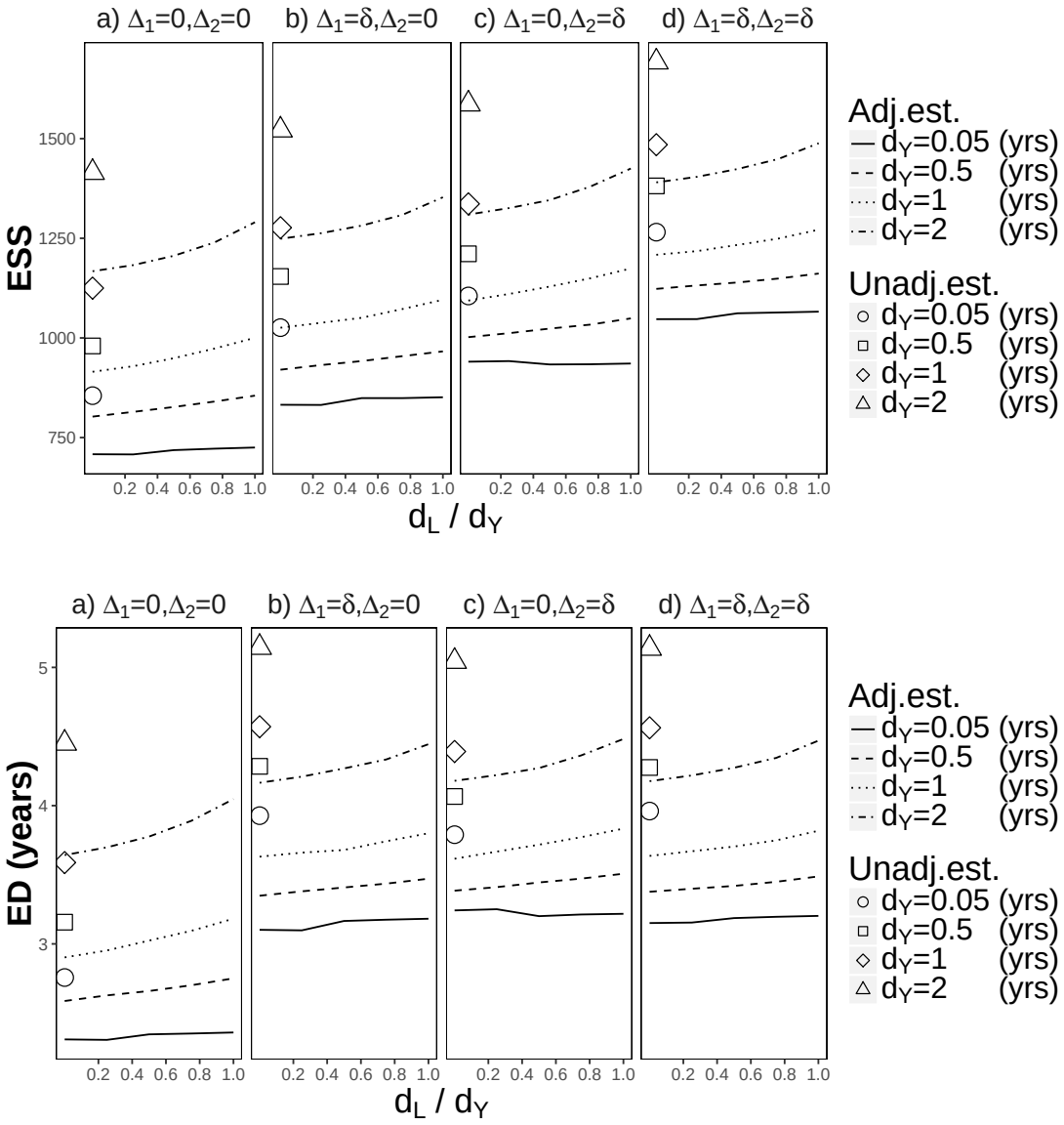


Figure 3: Effect of d_Y and d_L on ESS and ED in simulation study 4. Since the results corresponding to unadjusted estimator do not change when d_L varies as long as d_Y is fixed, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols. δ refers to $\delta_{\min}$.

5.3 Simulation Study 5: Effect of Accrual Rate

Figure 4 illustrates how the ESS and ED are affected by different accrual rates. Because the information depends either entirely (for the unadjusted estimator) or largely (for the adjusted estimator) on the number of participants who have the delayed outcome Y observed, with faster accrual there will generally be more pipeline participants at interim analyses. These additional pipeline participants make ESS larger. On the other hand, ED gets shorter with faster accrual. The impact of accrual rate on ESS and ED is similar across the two estimators.

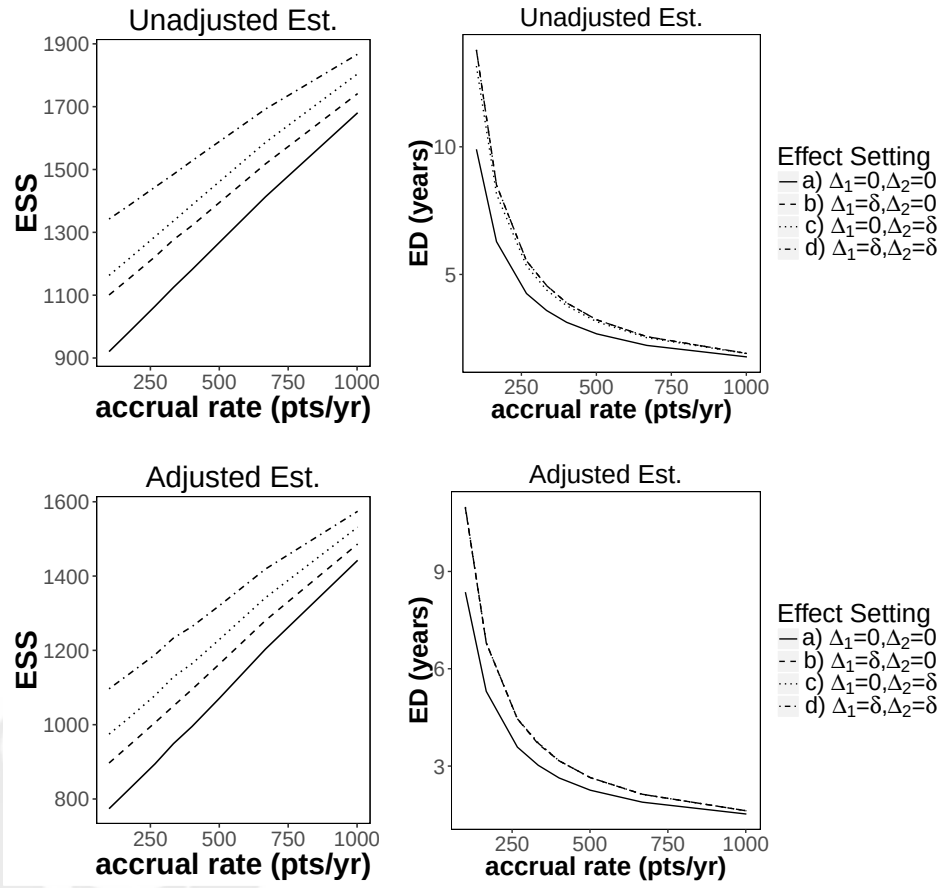


Figure 4: Impact of accrual rate on ESS and ED. Different line types indicate the ESS and ED under four effect settings. For each estimator, the curves for ED under effect settings (b)-(d) are clustered together. δ refers to $\delta_{\min}$.

6 Information Accrual Rates and Estimating Information Levels

In Section 4.4 we presented our approach for determining the time for analyses based on information monitoring. Here we explore information accrual more thoroughly and discuss how accurately information can be estimated in an ongoing trial. At each time, we are interested in two types of information level: the *current* information, i.e., the inverse of variance of the estimator computed using available data at the time, and the *wait-for-pipeline* information, i.e., the inverse of variance of the estimator using available data at the time plus the not yet observed L and Y of the pipeline participants. The current information is used for determining time for interim analyses, and the wait-for-pipeline information is used for determining time for the final analysis when we wait until all pipeline participants finish the trial and then test hypotheses.

Figure 5(a) shows how the two types of information accrue over time for the two estimators under the default simulation scenario when enrollment is not stopped.

For the unadjusted estimator, the information at a given time is proportional to the number of patients with Y observed; for the adjusted estimator, such proportionality is only approximate because the pipeline participants also contribute information. There is an approximately constant gap between the current information and the wait-for-pipeline information for each estimator, because the extra information in the not yet observed outcomes from the pipeline participants stays roughly constant over time. The adjusted estimator results in a faster information accrual compared to the unadjusted estimator, which is consistent with better trial performance (as shown in Section 5). The information accrual rates do not depend on Δ_1, Δ_2 since in our setup these do not impact the estimator's variance.

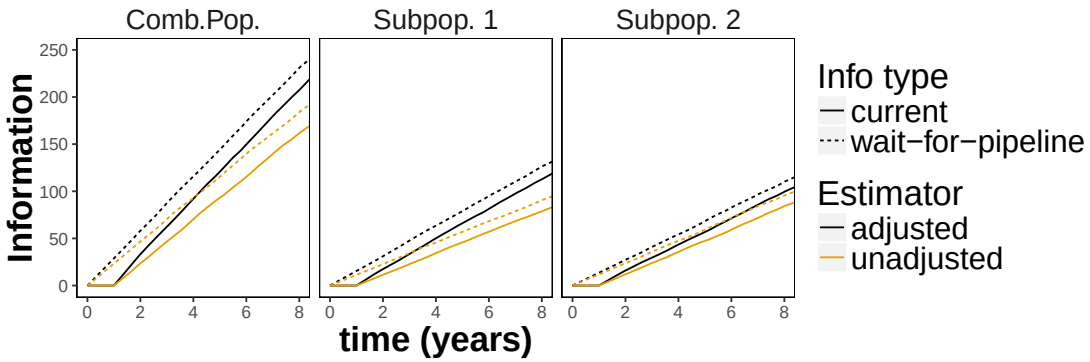
In practice, one needs a reliable method for estimating the information level using data from the ongoing trial in order to determine information-based timing for interim and final analyses. The sample variance is used to estimate the true variance of the unadjusted

estimator. For the adjusted estimator, its variance can be estimated using the nonparametric bootstrap or by the influence curve. The `ltmle` package computes an influence-curve-based variance estimate (ICVE) for the TMLE estimator. ICVE can be conservative in the sense that it may overestimate the variance (van der Laan and Gruber, 2012); in our simulation, however, it approximates the variance quite well.

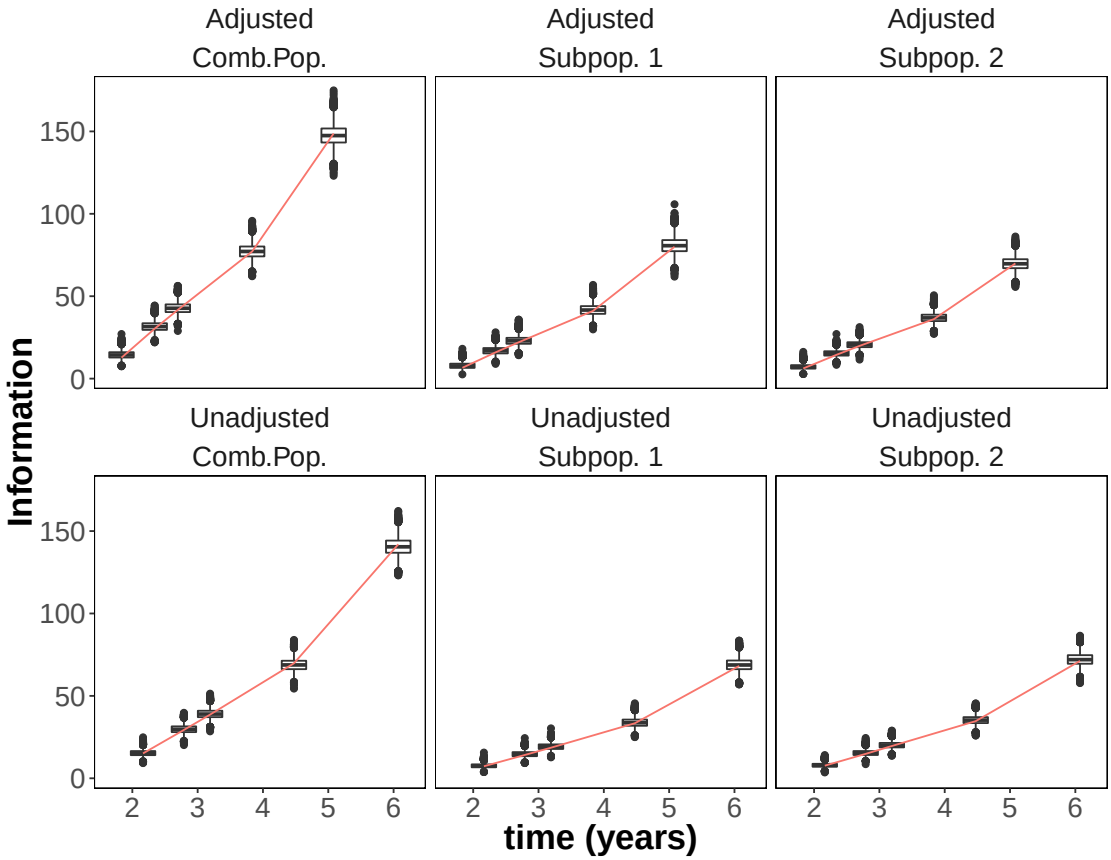
Figure 5(b) summarizes the performance of the variance estimators under the default simulation scenario. The solid red line connects the true information levels over time, and the box-plots represent the distribution of ICVE at 5 analyses assuming no early stopping. The mean and the spread of the distribution of ICVE increase with time (and hence with sample size n), because the information level is approximately n times the reciprocal of the variance of the estimator's influence curve, and the latter is estimated with standard error proportional to $n^{-1/2}$ asymptotically. Therefore, the spread in the box plots representing the approximate interquartile range grows at rate $n^{1/2}$. A similar observation applies to the sample variance estimate for the unadjusted estimator. Estimation accuracy for information accrual is similar for the two estimators.

7 Discussion

In simulation studies 3 and 4 in Section 5.2, we set constant prognostic values R^2_W and R^2_L, while varying d_L and d_Y. It may also be of interest to consider a range of simulation scenarios where the prognostic value changes with delay. For example, it is possible that with longer d_Y, the baseline variables W become less correlated with the final outcome Y, e.g., if these variables measure the same quantity at different time points. In addition, if d_L is closer to d_Y then the correlation between L and Y may be stronger. It is an area of future research to explore such simulation scenarios, in which there is a trade-off such that shorter d_L means more participants will have L but not Y observed, but such L is less prognostic for Y.



(a) Information accrual under the default simulation scenario. Yellow corresponds to unadjusted estimator and black to adjusted estimator.



(b) Box-plots of estimated information level for adjusted estimator (using influence-curve-based method) and unadjusted estimator (using sample variance) at each of the five analyses assuming no early stopping of enrollment (so that enrollment stops $d_Y = 1$ year before the final analysis; see Table 3). The red solid line connects the true information levels, and each box-plot shows the spread of the estimated information level.

Figure 5: Information accrual rates and box-plots of estimated variance for the adjusted and unadjusted estimators under the default simulation scenario.

We used the full set of baseline variables $(W_1, \ldots, W_5)$ in generating data, and only used (W_3, W_4) in the adjusted estimator. Since model misspecification is likely to occur in practice, we think it is important to have incorporated this in our simulation study. We could also include subpopulation information S as a baseline variable in estimating the treatment effect for the combined population with the adjusted estimator. Another potential modification would be to separately optimize the trial design for the adjusted estimator (rather than use the same information-based design that was optimized for the unadjusted estimator). These modifications could further increase the gains due to adjustment.

In (3) we set $\alpha_A = 0$, i.e., the treatment doesn't affect the short-term outcome. Setting this to be nonzero could impact efficiency gains from prognostic L.

Open research problems include investigating the impact of subpopulation proportion, and generalizing the findings to other designs and data generating mechanisms. Another problem is to evaluate the impact of dropout in the simulation. The adjusted estimator can provide advantages over the unadjusted estimator for handling dropout under the missing at random assumption, in which case the unadjusted estimator will typically be inconsistent (van der Laan and Gruber, 2012).

Acknowledgments

This research was supported by the Patient-Centered Outcomes Research Institute (ME-1306-03198). This paper's contents are solely the responsibility of the author and do not represent the views of the above agency. We thank Mary Joy Argo for helpful comments.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bio-

engineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Biogen (2016). 221AD302 phase 3 study of Aducanumab (BIIB037) in early Alzheimer's disease (EMERGE). In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2016 May 11]. Available at `https://clinicaltrials.gov/ct2/show/nct02484547?term=mci+biogen&rank=1`.

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. *Stat. Med.* **28,** 1445–1463.

Bretz, F., Schmidli, H., König, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical J.* **48,** 623–634.

Fisher, A. and Rosenblum, M. (2016). Stochastic optimization of adaptive enrichment designs for two subpopulations. *Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 279.* `http://biostats.bepress.com/jhubiostat/paper279` .

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

Gruber, S. and van der Laan, M. J. (2012). Targeted minimum loss based estimator that outperforms a given estimator. *Int. J. Biostat.* **8,**.

Jennison, C. and Turnbull, B. W. (2007). Adaptive seamless designs: selection and prospective testing of hypotheses. *J. Biopharm. Stat.* **17,** 1135–1161.

Lan, 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 Anal.* **17,** 566–593.

Magnusson, B. P. and Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Stat. Med.* **32,** 2695–2714.

Rosenblum, M., Qian, T., Du, Y., Qiu, H., and Fisher, A. (2016). Multiple testing procedures for adaptive enrichment designs: Combining group sequential and reallocation approaches. *Biostatistics. doi: 10.1093/biostatistics/kxw014* .

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

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 J.* **48,** 635–643.

Schwab, J., Lendle, S., Petersen, M., and van der Laan, M. (2015). *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*. R package version 0.9-5.

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

Stallard, N. (2011). Group-sequential methods for adaptive seamless phase II/III clinical trials. *J. Biopharm. Stat.* **21,** 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,** 168–187.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *Int. J. Biostat.* **8,**.

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

A Supplementary Material

A.1 Detail for varying prognostic values in the data generating mechanism

The fitted α, β and σ's from the ADNI study data are shown in Table 5.

Table 5: Fitted α, β and σ's from the ADNI study data

L is not measured

	$s=1$	$s=2$
β_0^s	-1.131	-1.664
$\beta_{W_1}^s$	0.007	0.172
$\beta_{W_2}^s$	-0.027	0.013
$\beta_{W_3}^s$	0.001	-0.007
$\beta_{W_4}^s$	-0.148	-0.111
$\beta_{W_5}^s$	0.027	0.150
σ_Y^s	1.552	1.773

L is measured

	$s=1$	$s=2$		$s=1$	$s=2$
β_0^s	-0.699	-0.808	α_0^s	-0.485	-0.734
$\beta_{W_1}^s$	-0.134	0.140	$\alpha_{W_1}^s$	0.158	0.028
$\beta_{W_2}^s$	-0.011	0.009	$\alpha_{W_2}^s$	-0.018	0.003
$\beta_{W_3}^s$	-0.002	-0.006	$\alpha_{W_3}^s$	0.004	-0.001
$\beta_{W_4}^s$	-0.098	-0.031	$\alpha_{W_4}^s$	-0.057	-0.068
$\beta_{W_5}^s$	-0.017	0.099	$\alpha_{W_5}^s$	0.049	0.043
β_L^s	0.890	1.167			
σ_Y^s	1.247	1.342	σ_L^s	1.044	0.996

In varying R_W^2 and R_L^2 as described in Section 4.2, it is desired that the average treatment effect within each subpopulation remains unchanged, and the variance of Y given A, S remains unchanged. This implies that $E(Y \mid A=1, S=s) - E(Y \mid A=0, S=s), s \in \{1,2\}$ and $\text{var}(Y \mid A=a, S=s)$ for $a \in \{1,2\}$ and $s \in \{1,2\}$ need to be unchanged. Throughout rest of the subsection we omit the superscript s, because the following procedures will be conducted separately within each of the subpopulations.

In simulation scenarios where L is not measured, we multiply β_W by a tuning parameter p_W in (2). The mean and variance of Y given $A = a$ become:

$$E(Y \mid A=a) = \beta_0 + \beta_A a, \tag{5}$$

$$\begin{aligned} \text{var}(Y \mid A=a) &= \text{var}(p_W \beta_W W) + \text{var}(\epsilon_Y) \\ &= p_W^2 \beta_W^T \text{var}(W) \beta_W + \sigma_Y^2. \end{aligned} \tag{6}$$

$E(Y \mid A=a)$ does not depend on p_W, and neither does the average treatment effect. For a

given p_W, we solve for σ_Y^2 so that the value of (6) is constant under different specifications of p_W. R_W^2 with $A = 0$ defined in (1) becomes:

$$R_W^2 = \frac{p_W^2 \beta_W^T \text{var}(W) \beta_W}{\text{var}(Y \mid A = 0)}. \tag{7}$$

In simulation scenarios where L is measured, we multiply β_L by a tuning parameter p_L in (4). The mean and variance of Y given $A = a$ become:

$$E(Y \mid A = a) = \beta_0 + p_L \alpha_0 \beta_L + \beta_A a, \tag{8}$$

$$\begin{aligned} \text{var}(Y \mid A = a) &= \text{var}\{(\beta_W + p_L \alpha_W \beta_L) W\} + \text{var}(\epsilon_Y + p_L \beta_L \epsilon_L) \\ &= (\beta_W + p_L \alpha_W \beta_L)^T \text{var}(W)(\beta_W + p_L \alpha_W \beta_L) + \sigma_Y^2 + p_L^2 \beta_L^2 \sigma_L^2, \end{aligned} \tag{9}$$

The average treatment effect does not depend on p_L. For a given p_L, we solve for σ_Y^2 so that the value of (9) is constant under different specifications of p_L. R_L^2 with $A = a$ defined in (1) becomes:

$$R_L^2 = \frac{p_L^2 \beta_L^2 \text{var}(L)}{\text{var}(Y \mid A = a)}. \tag{10}$$

A.2 Algorithm to Compute Efficacy Boundaries $u_{s,k}$

At stage k for the null hypothesis H_{0s}, $1 \leq k \leq K, s \in \{0, 1, 2\}$, denote $\alpha_{s,k}$ as the Type I error to be spent, and $u_{s,k}$ the efficacy boundary for the Wald statistic $Z_{s,k}$ (estimator divided by its standard deviation). Define the ordering $(s', k') \prec (s, k)$ if and only if $k' < k$ or ($k' = k$ and $s' < s$). Define $u_{s,k}$ to be the solution to

$$P\{Z_{s',k'} \leq u_{s',k'} \text{ for all } (s', k') \prec (s, k), \text{ and } Z_{s,k} > u_{s,k} \mid \Delta_1 = \Delta_2 = 0\} = \alpha_{s,k},$$

where the joint distribution of $Z_{s,k}$ is approximated by a normal distribution, with variance-covariance matrix estimated from 10,000 simulated trials.

A.3 Detail for Analysis Timing

For each fixed data generating distribution

$$Q = Q\left(\Delta_1, \Delta_2, R_W^2, d_Y, \text{accrual rate}, L \text{ measured}, R_L^2, d_L\right)$$

and each estimator, our method to determine time of analyses consists of four steps:

Step 1: Generate 10,000 pilot simulated trials where interim analyses are conducted at 25 pre-selected calendar time points $t_1, \ldots, t_{25}$, such that approximately 50 patients from subpopulation 1 are enrolled between t_j and t_{j+1}. For each t_j, we record the estimated treatment effect τ_j at that time and the "wait-for-pipeline" treatment effect $\tilde{\tau}_j$ that is obtained by assuming enrollment is stopped at t_j and estimating the treatment effect after Y is measured for all pipeline participants.

Step 2: Compute the variances of τ_j and $\tilde{\tau}_j$ from the 10,000 pilot simulated trials, the inverse of which are the *current* information and *wait-for-pipeline* information at t_j, respectively.

Step 3: For interim analyses 1 - 4, linearly interpolate to find calendar time T_k of which the current information equals that listed in Table 2 for the corresponding $k \in \{1, 2, 3, 4\}$. For the final analysis, linearly interpolate to find the calendar time T_5 of which the wait-for-pipeline information equals that listed in Table 2 for $k = 5$.

Step 4: In the simulated trials, interim analyses 1 - 4 are conducted at calendar times $T_1, \ldots, T_4$, enrollment stops at T_5 (if no early stopping occurs), and final analysis is conducted at calendar time $T_5 + d_Y$.

In step 3, for each k we identify the calendar time such that the information accrued for

subpopulation 1, subpopulation 2 and the combined population all exceed the corresponding threshold in Table 2. The design is optimized for the unadjusted estimator, so that when using the unadjusted estimator, the thresholds for the three populations are crossed at almost the same time. However, for the adjusted estimator, since W is more prognostic in subpopulation 1 than in subpopulation 2 in the ADNI study data, by the time the information accrued for subpopulation 2 reaches the threshold, information accrued for subpopulation 1 and the combined population already exceed their corresponding thresholds. Thus, for the adjusted estimator at each interim analysis the information for subpopulation 2 is exactly as in Table 2, whereas the information for subpopulation 1 and the combined exceeds the thresholds in Table 2. This makes the power for H_{01} and H_{00} higher than 80% for the adjusted estimator (as presented in Section A.4.1, Figure 6).

A.4 Additional Simulation Results

A.4.1 Impact of Prognostic Value, Delay Time and Accrual Rate on power

Figure 6 shows that for the adjusted estimator power of H_{01} and H_{00} increases with larger R^2_W, whereas the power of H_{02} remains roughly constant. Change in R^2_L, d_Y, d_L or accrual rate does not substantially affect power. For the unadjusted estimator power is always constant.

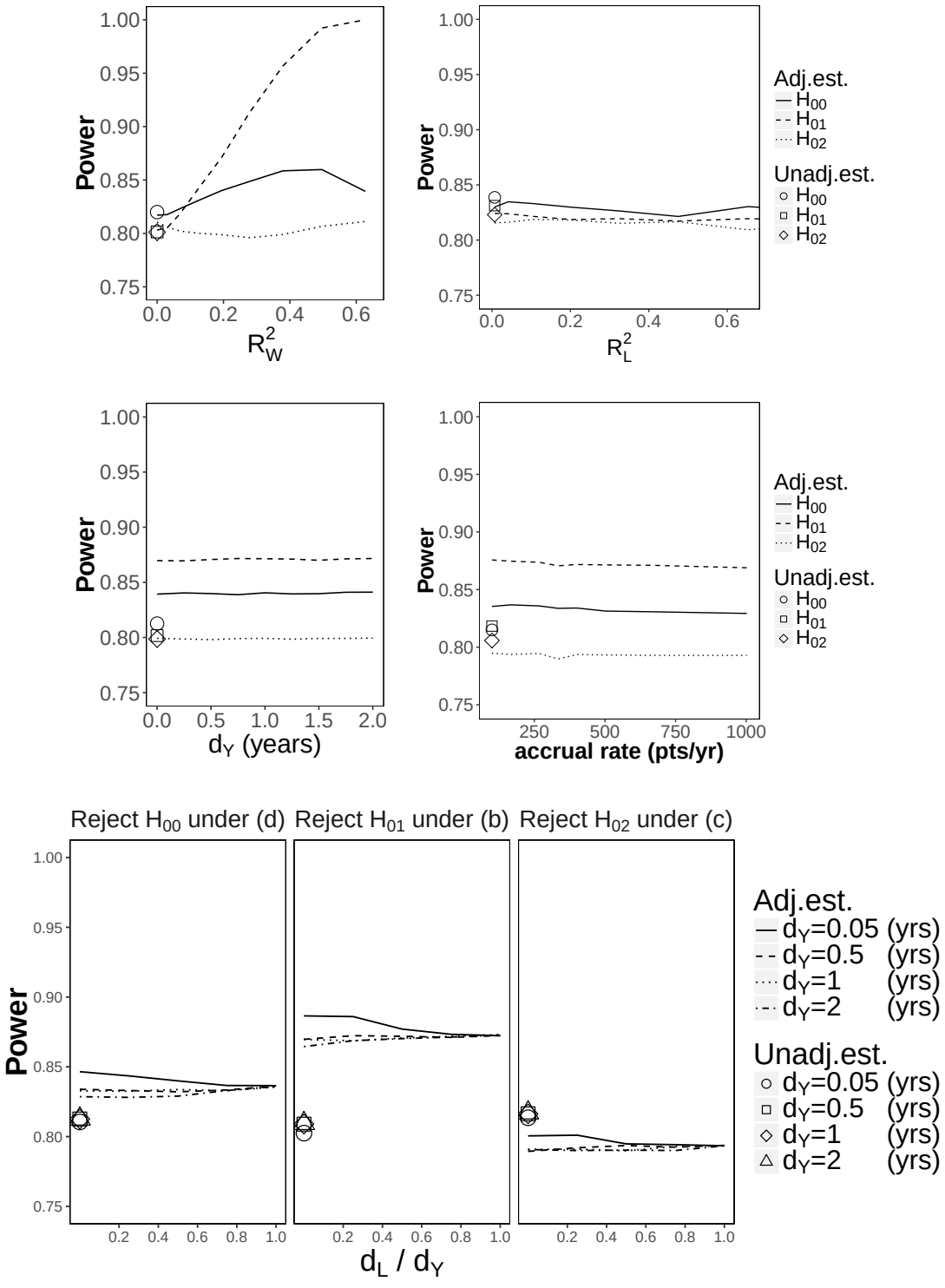


Figure 6: Change in power under simulation studies 1-5. The power of $H_{00}/H_{01}/H_{02}$ is the probability to reject at least $H_{00}/H_{01}/H_{02}$ under effect setting (d)/(b)/(c). Since the results corresponding to unadjusted estimator do not change as the design characteristics are varied, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols.