Contemporary Clinical Trials Communications 8 (2017) 39–48

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

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

Sensitivity of adaptive enrichment trial designs to accrual rates, time to outcome measurement, and prognostic variables

Tianchen Qian*, Elizabeth Colantuoni, Aaron Fisher, Michael Rosenblum, for the Alzheimer's Disease Neuroimaging Initiative[1]

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

ARTICLE INFO

Keywords:
Adaptive enrichment
Clinical trial design
Group sequential stopping
Multiple testing procedure
Treatment effect heterogeneity

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. Trial characteristics such as the accrual rate and the prognostic value of baseline variables are typically unknown when a trial is being planned; these values are typically assumed based on information available before the trial starts. Because of the added complexity in adaptive enrichment designs compared to standard designs, it may be of special concern how sensitive the trial performance is to deviations from assumptions. Through simulation studies, we evaluate the sensitivity of Type I error, power, expected sample size, and trial duration to different design characteristics. Our simulation distributions mimic features of data from the Alzheimer's Disease Neuroimaging Initiative cohort study, and involve two subpopulations based on a genetic marker. We investigate the impact of the following design characteristics: the accrual rate, the time from enrollment to measurement of a short-term outcome and the primary outcome, and the prognostic value of baseline variables and short-term outcomes. 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. We apply information-based monitoring, and evaluate how accurately information can be estimated in an ongoing trial.

1. Introduction

Adaptive enrichment designs involve pre-planned rules for restricting enrollment based on accrued data in an ongoing trial [1]. 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 [2]. give an overview of statistical methods for adaptive enrichment designs, including the p-value combination approach [3–6]; the conditional error function approach [7]; and approaches using group sequential computations [8,9]. We use an adaptive enrichment design from the general class of [10]; which is based on the group sequential computation approach.

We consider trials where the primary outcome is observed a fixed amount of time after enrollment (called the delay); we refer to such outcomes as delayed responses. To illustrate, we use data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort study. We set the primary outcome to be a measure of change in severity of dementia symptoms from baseline to 2 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 cognitive and functional decline from early Alzheimer's Disease [11]. Also recorded are baseline variables and the short-term outcome of change in severity of dementia symptoms measured at 1 year 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 Ref. [12]. 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 responses, the adjusted estimator uses information from pipeline participants, i.e., enrollees whose primary outcome has not yet been observed.

We evaluate the sensitivity of Type I error, power, expected sample

* Corresponding author.
E-mail address: tqian2@jhu.edu (T. Qian).

[1] 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.

http://dx.doi.org/10.1016/j.conctc.2017.08.003
Received 28 July 2016; Received in revised form 19 April 2017; Accepted 11 August 2017
Available online 16 August 2017

size, and trial duration to different design characteristics through simulation studies. Our simulation distributions mimic features of data from the Alzheimer's Disease Neuroimaging Initiative, and involve two subpopulations of interest based on a genetic marker. We investigate the impact of the following design characteristics: the accrual rate, the delay time of the short-term outcome and the primary outcome, and the prognostic value of baseline variables and short-term outcomes. The simulated trials involve multiple stages, and information-based monitoring is used to determine the time of interim analyses.

We focus on adaptive enrichment designs since their added complexity (compared to standard designs) may raise special concern about how sensitive their performance is to deviations from initial assumptions. Since statistics from multiple populations are used in the stopping rule and multiple testing procedure, changes to assumptions (which affect the joint distribution of these statistics) could have impacts that are not easy to predict a priori. This was observed, for example, when we varied the ratio of information accrual rates in the two subpopulations; in these cases the covariance structure of the test statistics is affected. This sometimes resulted in higher than 80% power for certain hypothesis tests, despite the fact that we used information-based monitoring (which in a single population trial design would maintain constant power at a given alternative). These results are described in Section 5.

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 concludes with discussions 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.[2] 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 CDR sum of boxes score as the CDR score. We define the primary outcome Y as the difference between the CDR score at baseline and at 2 years. We define the short-term outcome L as the difference between the CDR score at baseline and at 1 year. 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) $\varepsilon 4$ carrier status. Subpopulation 1 consists of those with no $\varepsilon 4$ alleles, and subpopulation 2 consists of those with at least one $\varepsilon 4$ allele. Among the 286 patients, 47% carry no APOE $\varepsilon 4$ alleles. We consider a hypothetical treatment whose goal is to delay the progression of disease.

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

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 time d_L and d_Y, respectively, from the time of enrollment. Assume $d_L \le d_Y$. Each vector $\mathbf{D}$ 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 [13].

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 under 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, and we have $\Delta_0 = p_1\Delta_1 + p_2\Delta_2$. Define the null hypotheses

$$H_{01}: \Delta_1 \le 0; \quad H_{02}: \Delta_2 \le 0; \quad H_{00}: \Delta_0 \le 0,$$

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

$$R_W^2 = \frac{\mathrm{Var}\{E(Y|W)\}}{\mathrm{Var}(Y)}, \quad R_L^2 = \frac{\mathrm{Var}\{E(Y|L)\}}{\mathrm{Var}(Y)}. \tag{1}$$

R_W^2 represents the fraction of variance in Y explained by W. R_L^2 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_W^2 is computed as in (1), with $E(Y|W)$ 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_L^2. The resulting values are 0.20 and 0.48 for R_W^2 and R_L^2, respectively, for the combined population.

We also estimated R_W^2 and R_L^2 within each subpopulation, and found the prognostic values differ by subpopulation. The corresponding empirical R_W^2 is 0.30 for subpopulation 1 and 0.14 for subpopulation 2; the empirical R_L^2 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_W^2 and R_L^2 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 response 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 [12] implemented in the R package ltmle [14]. Other candidate adjusted estimators include, e.g., those of Refs. [15,16,13]. Colantuoni and Rosenblum [17] showed that although the magnitude of precision gains depends on the estimator used, the impact of adjustment was qualitatively similar across estimators in their simulation studies. We conjecture that the same will hold in our simulation studies, but it is an open problem to determine this. In our simulation studies, both the unadjusted and adjusted estimators are consistent and asymptotically normal under mild regularity conditions [12].

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 (1); the delay time d_L of the short-term outcome; the delay time d_Y of 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 **D** is drawn independently from a data generating distribution Q, which differs by simulation study. We construct 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.

There is no treatment in the ADNI study. We draw each participant's study arm variable A independent of S,W as a Bernoulli random variable with success probability 0.5, 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 in Table 1), 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 is denoted by

$$Q = Q(\Delta_1, \Delta_2, R^2_W, R^2_L, d_L, d_Y, \textit{accrual rate}),$$

which 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 prognostic value of the short-term outcome R^2_L, the delay between enrollment and the short-term outcome d_L, the delay between enrollment and the primary outcome d_Y, and the accrual rate. We set the enrollment process to be random, where the enrollment time of the patients follows a homogeneous 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.

Within each subpopulation $S = s$, W is randomly sampled from the observed data, and Y and L are generated from the linear models:

$$L = \alpha_0^s + \alpha_W^s W + \alpha_A^s A + \varepsilon_L, \quad \varepsilon_L \sim N(0, (\sigma_L^s)^2) \tag{2}$$

$$Y = \beta_0^s + \beta_W^s W + \beta_A^s A + \beta_L^s L + \varepsilon_Y, \quad \varepsilon_Y \sim N(0, (\sigma_Y^s)^2) \tag{3}$$

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. We set $\alpha_A^s = 0.5\Delta_s$ and $\beta_A^s = \Delta_s - \alpha_A^s\beta_L^s$, where Δ_s is the desired treatment effect of the corresponding effect setting. This makes the treatment effect on the short-term outcome half of that on the final outcome, which we believe is plausible.

We construct simulation distributions with a range of R^2_W and R^2_L values by varying β's and σ's. 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. This is to ensure that the (asymptotic) performance of the unadjusted estimator is unchanged, providing a benchmark to compare against. To obtain different values of R^2_W and R^2_L, we multiply the original fits β_W^s and β_L^s from the ADNI data set by a tuning parameter p, and change β_0^s, β_A^s, σ_Y^s and σ_L^s accordingly to ensure that the variance of Y given A,S and the average treatment effect given S are unchanged. Details are given in the Supplementary Material.

Let the default simulation scenario be the one with design characteristics corresponding to the empirical distribution of the ADNI study data: $R^2_W = 0.20$, $R^2_L = 0.48$, $d_L = 1$ year, $d_Y = 2$ years, and the accrual rate for the combined population 167 patients/year. We conduct 5 sets of simulations with various design characteristics that are summarized in Table 1. Each combination of (R^2_W, R^2_L, d_L, d_Y,accrual rate) is referred to as a simulation scenario. For example, in simulation study 1 (row 1 in Table 1), R^2_W is varied from 0 to 0.6 and all other characteristics are 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) and (3) 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 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. (The usage of logistic regression on bounded continuous variables is justified in Ref. [18] because it preserves the bounds on the outcome.) 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 [13]; Section 4).

Table 1
Summary of setups for 5 simulation studies. Default value of parameter: $R^2_W = 0.20$, $R^2_L = 0.48$, $d_L = 1$ years, $d_Y = 2$ year, accrual rate 167 patients/year. Ranges of values $x - y$ indicate the design characteristic(s) varied in the corresponding simulation study.

Simulation study	R^2_W	R^2_L	d_L (years)	d_Y (years)	Accrual rate (patients/year)
1	0 − 0.6	0	default	default	default
2	0	0 − 0.6	default	default	default
3	default	0	default	0 − 4	default
4	default	default	0 − d_Y	0.1,1,2,3,4	default
5	default	default	default	default	50 − 500

4.3. Adaptive enrichment design

We define a new adaptive enrichment design using the general framework developed by Ref. [10]. We consider two subpopulations denoted by S: $S = 1$ if the patient has no APOE ε4 allele, and $S = 2$ if the patient has one or more APOE ε4 allele. Denote by $S = 0$ the combined population. We consider an adaptive enrichment design with maximum number of stages $K = 5$. At each analysis $k \le K$, denote by $Z_{s,k}$ 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 $k \le K$, 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 ($s \in \{1,2\}$). Below are the steps that are followed at each analysis $k \le K$ to

determine the continuation (or stop) of enrollment and the results of hypothesis testing.

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. For $s \in \{0,1,2\}$, if H_{0s} is not rejected in the above steps, we fail to reject it.

Define the power of H_{01} to be the probability to reject H_{01} under effect setting (b), power of H_{02} to be the probability to reject H_{02} under effect setting (c), and power of H_{00} to be the probability to reject 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 $l_{s,k}$, $s \in \{1,2\}$, $1 \le k \le K$ 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 Ref. [19]; 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 of subpopulation 1 $p_1 = 0.47$ and the variances differ by subpopulation. In determining the values of efficacy boundaries $u_{s,k}$, $s \in \{0,1,2\}$, $1 \le k \le K$, we use the error spending approach as described in Rosenblum et al. (2016, Section 3.2), which extends the approach of [20,21] to multiple populations. The boundaries are numerically calculated to ensure that the test at each stage maintains its pre-specified Type I error, by assuming joint normal distribution of the test statistics; see the Supplementary Material for details. These efficacy boundaries depend on the covariance matrix of the estimator being used. As shown in Ref. [10]; 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 the 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 (Section 6). Faster information accrual can lead to earlier analyses in calendar time.

Since in a real trial 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 50,000 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 estimator and the adjusted estimator 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 estimator due to the former having a faster information accrual rate.

5. Results

We simulated 50,000 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.

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} are all about 80% under different simulation scenarios. This is as expected due to our method of determining the analysis timing 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_{00} and H_{01} under certain simulation scenarios can be much higher than 80%. For example, when the prognostic value in W (R^2_W) is over 0.3, the power of H_{01} can exceed 90%. This is because when adjusting for baseline variables, the ratio of information accrual rate between the two subpopulations is different than when the unadjusted estimator is used, which changes the covariance matrix of the test statistics. 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 estimator and the adjusted estimator, so

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)	4.4	5.6	6.4	9.1	12.3	14.3
CSS (Subpop. 1)	202 (+148)	299 (+135)	353 (+149)	544 (+157)	928 (+138)	1066 (+0)
CSS (Subpop. 2)	211 (+190)	329 (+175)	405 (+158)	620 (+170)	1040 (+183)	1223 (+0)
CSS (Comb. Pop.)	413 (+338)	628 (+310)	758 (+307)	1164 (+327)	1968 (+321)	2289 (+0)
Adjusted estimator						
Time (years)	3.7	4.8	5.5	7.7	10.2	12.2
CSS (Subpop. 1)	138 (+159)	219 (+158)	278 (+156)	453 (+158)	824 (+171)	995 (+0)
CSS (Subpop. 2)	150 (+164)	236 (+169)	295 (+182)	500 (+196)	916 (+186)	1102 (+0)
CSS (Comb. Pop.)	288 (+323)	455 (+327)	573 (+338)	953 (+354)	1740 (+357)	2097 (+0)

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*	1.1*	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	1.0*	**80**	0.9*	85
	(d) $\Delta_1 = \Delta_2 = \Delta_{\min}$	**84**	87	80	69	98
Unadjusted estimator	(a) $\Delta_1 = \Delta_2 = 0$	0.6*	1.1*	1.1*	0.0*	2.5*
	(b) $\Delta_1 = \Delta_{\min}, \Delta_2 = 0$	12	**81**	1.0*	0.9*	81
	(c) $\Delta_1 = 0, \Delta_2 = \Delta_{\min}$	15	1.1*	**81**	0.9*	82
	(d) $\Delta_1 = \Delta_2 = \Delta_{\min}$	**82**	82	81	67	97

we believe the current simulation setup makes more sense.

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

Fig. 1 illustrates how ESS and ED are affected when one of R_W^2 or R_L^2 varies. The performance of the unadjusted estimator remains the same when the prognostic value in W and L changes, providing a benchmark to compare against. The adjusted estimator performs similar to the unadjusted when there is no prognostic value in W or L, i.e. $R_W^2 = R_L^2 = 0$. As R_W^2 or R_L^2 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_W^2 is varied from 0 to 0.6; in simulation study 2, R_L^2 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_W^2 from 0 to 0.25 results in a 19% drop in ESS (1618–1314), whereas increasing R_L^2 from 0 to 0.25 only renders a 1% drop (1618–1608). 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 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. Theoretical justification of this based on semiparametric efficiency theory can be found in Ref. [22].

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 4 years with L being not prognostic at all, in order to separate the impact of d_Y. In simulation study 4, we set d_Y to several levels, and in each case vary d_L from 0 (immediate L) to d_Y; in this situation we set the prognostic value of L to default (same as in the ADNI data set).

Fig. 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.

Fig. 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 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. Such impact of d_L is modified by d_Y in that having a quickly-observable short-term outcome (i.e., smaller d_L) is slightly more beneficial when the delay of the final outcome is longer. For example, when $d_Y = 4$ years, decreasing d_L from d_Y to 0 results in a 2% drop in average duration (8.9 years–8.7 years); when $d_Y = 0.1$ years, changes in d_L have almost no effect on the trial. Of course, this is also because we are considering d_L on a relative scale of d_Y. Finally, note that even when $d_L = d_Y$, which implies no asymptotic precision gain from adjusting for L, the adjusted estimator still gains efficiency from adjusting for prognostic W.

5.3. Simulation study 5: effect of accrual rate

Fig. 4 illustrates how the ESS and ED are affected by accrual rate when the outcomes are observed with delay. 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 response Y observed, with faster accrual there will generally be more pipeline participants at interim analyses. These additional pipeline participants make ESS larger. Therefore, having fast accrual can have the negative consequence of increasing the overall study size when the primary outcome is measured with delay. For ED the result is intuitive: the duration of the trial gets shorter with faster accrual. We observe similar trends for both estimators.

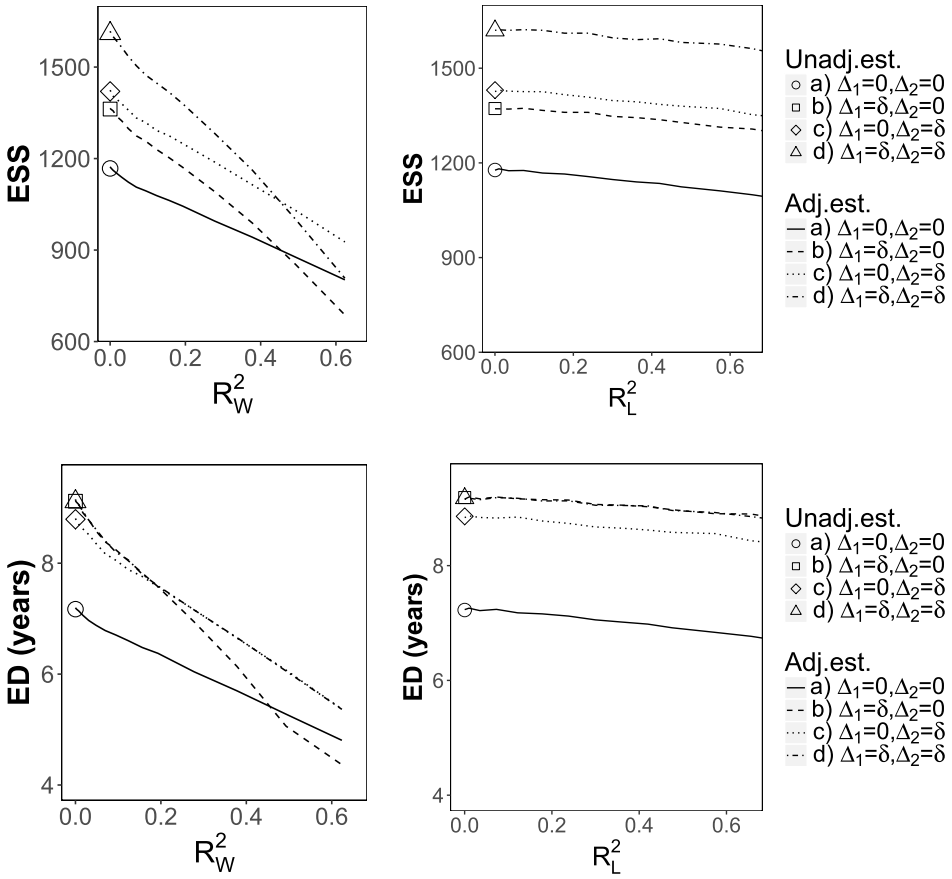


Fig. 1. Left: impact of R^2_W on ESS and ED in simulation study 1. Right: impact of R^2_L on ESS and ED in simulation study 2. Since the results corresponding to unadjusted estimator do not change as R^2_W and R^2_L are varied, they are marked only once next to the vertical axis using the circle, square,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

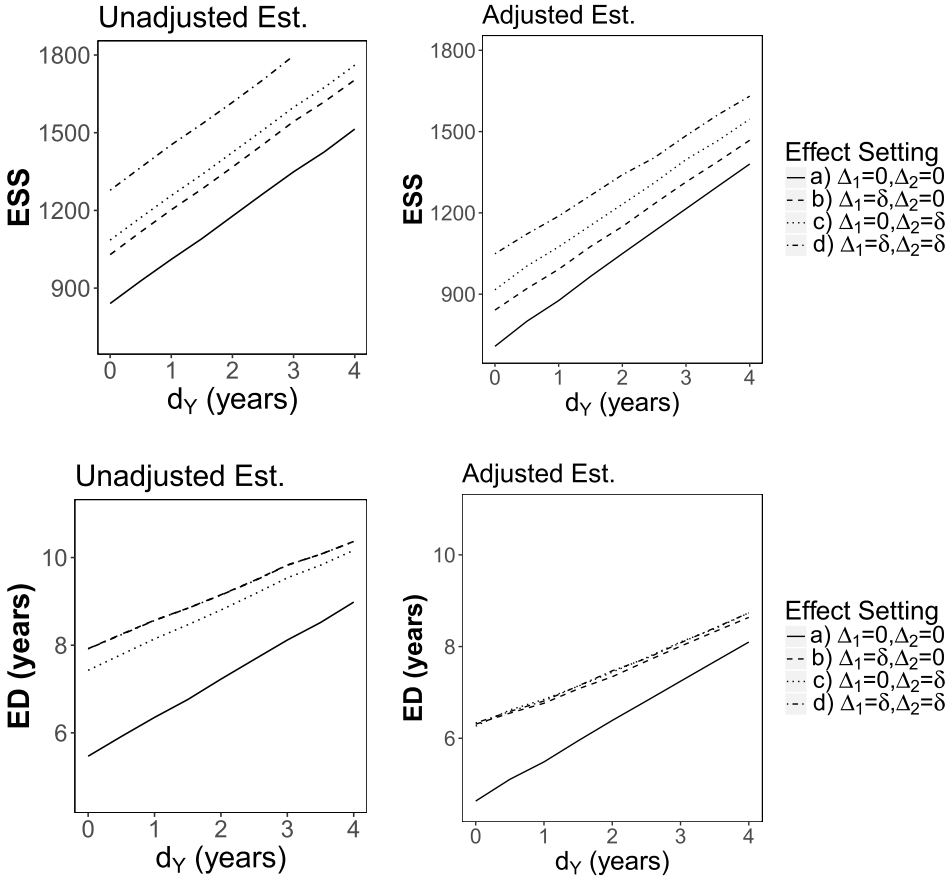


Fig. 2. Impact of d_Y on ESS and ED in simulation study 3. Different line types indicate the ESS and ED under four effect settings. For the unadjusted estimator, the lines for ED under effect settings (b) and (c) are clustered together. For the adjusted estimator, the lines for ED under effect settings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

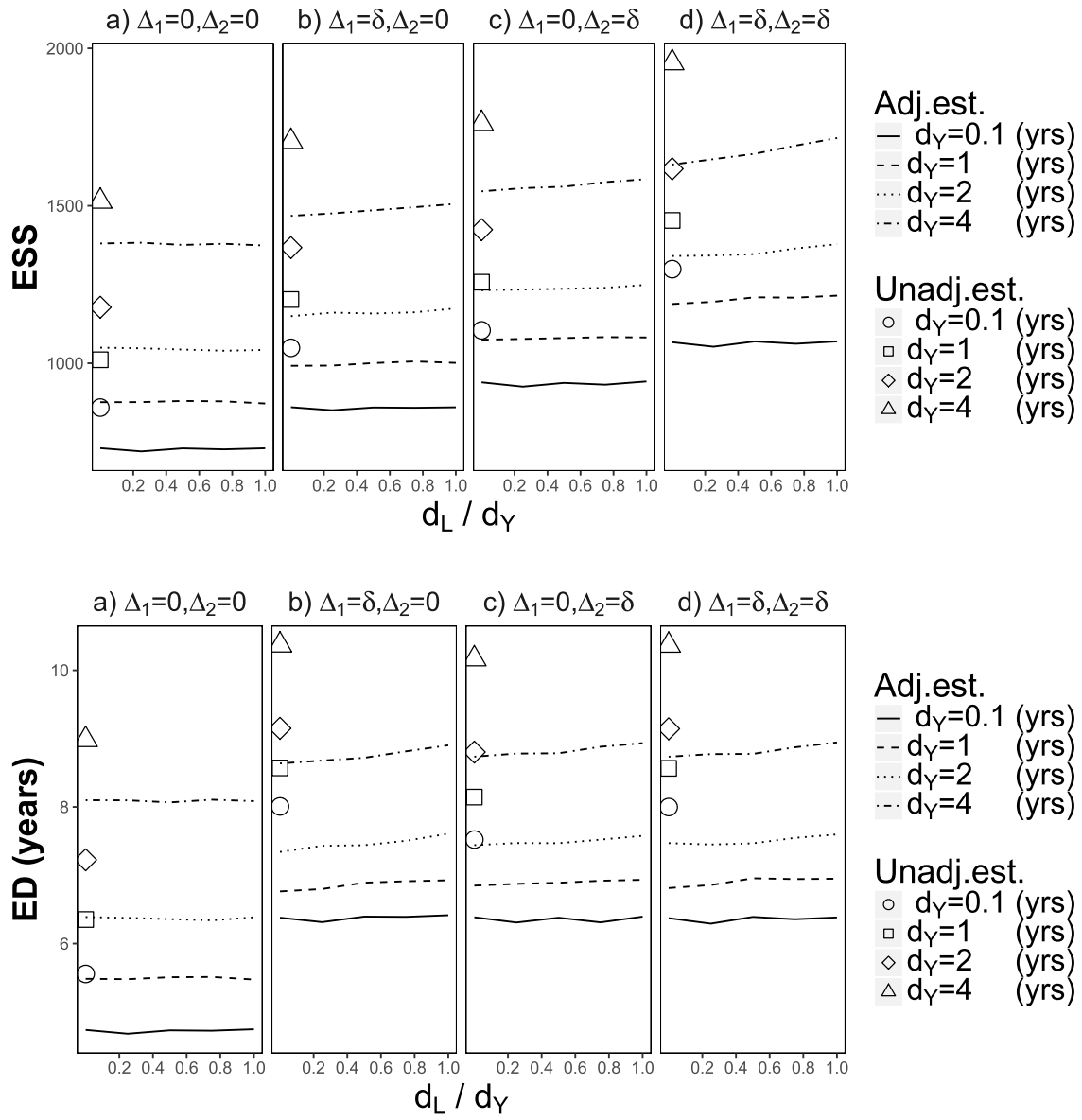


Fig. 3. Effect of d_Y and d_L on ESS and ED in simulation study 4. Since the results corresponding to unadjusted estimator do not change when d_L varies as long as d_Y is fixed, they are marked only once next to the vertical axis using the circle, square, diamond, and triangle symbols. δ refers to $\Delta_{\min}$.

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 time t, 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 time t, and the *wait-for-pipeline* information, i.e., the inverse of variance of the estimator using available data at time t plus the not yet observed L and Y of the pipeline participants at time t. In order words, the wait-for-pipeline information for time t is computed as if enrollment were stopped at time t and we wait till all pipeline participants finish the trial before calculating the estimator. 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 where we wait until all pipeline participants finish the trial and then test hypotheses.

Fig. 5(a) shows how the two types of information accrue over time for the two estimators under the default simulation scenario when enrollment for both subpopulations continues. 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 and 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)

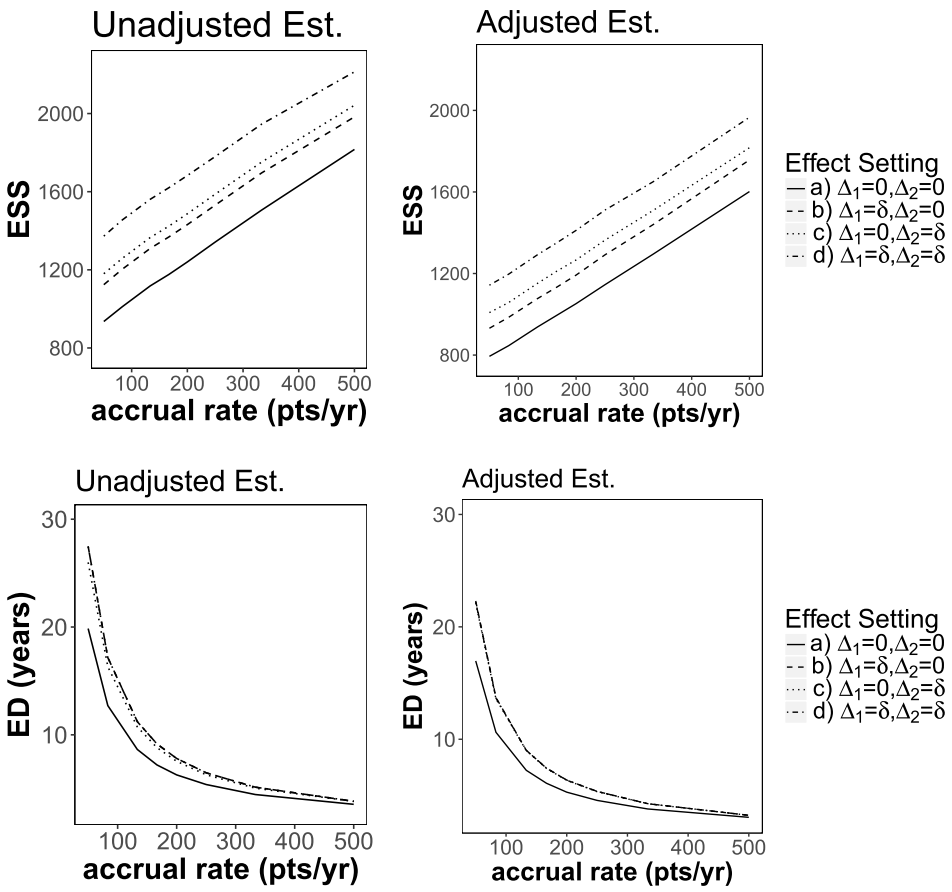


Fig. 4. Impact of accrual rate on ESS and ED. Different line types indicate the ESS and ED under four effect settings. For each estimator, the curves for ED under effect settings (b)–(d) are clustered together. δ refers to $\Delta_{\min}$.

for the TMLE estimator. In theory ICVE can be conservative in the sense that it may overestimate the variance [12]; in our simulation it approximates the variance quite well.

Fig. 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 inverse of variance estimator at 5 analyses assuming no early stopping (sample variance estimator for the unadjusted, ICVE for the adjusted). For the information of the adjusted estimator, the mean and the spread of the distribution 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. Conclusion and discussion

We conducted extensive simulation studies to examine the sensitivity of trial performance (measured by Type I error, power, expected sample size, and average duration) to different trial characteristics, including prognostic value of the baseline variable W and the short-term outcome L, delay time to observe the short-term outcome (d_L) and the primary outcome (d_Y), and the accrual rate. We constructed simulation distributions to mimic features of the ADNI data set. We used the full set of baseline variables in generating data, and only used a subset in the adjusted estimator to incorporate model misspecification in our simulation study. Throughout the paper, we do not assume that the short-term outcome L is a surrogate. That is, we are not using L as basis for stopping rules. L is only used for improving estimation precision of the treatment effect, due to its correlation with the primary outcome.

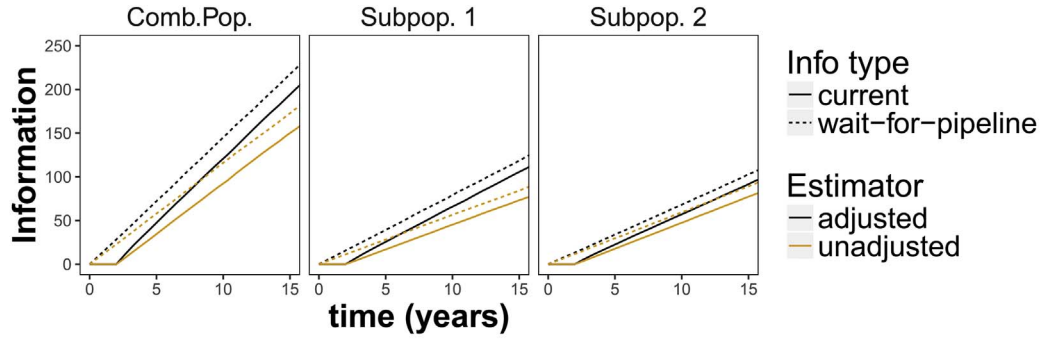
For both the unadjusted estimator and the adjusted estimator, expected sample size and trial duration increase with longer delay of the primary outcome (d_Y). Faster patient accrual results in shorter trial duration, but can have the negative consequence of increasing the overall study size when the primary outcome is measured with delay.

For trials using the adjusted estimator, with more prognostic W or L the power increases, and the expected sample size and average duration decrease. A prognostic W results in better trial efficiency compared to an equally prognostic L (measured in terms of R^2). Shorter d_L helps to slightly reduce expected sample size and average duration, when L is prognostic for the primary outcome.

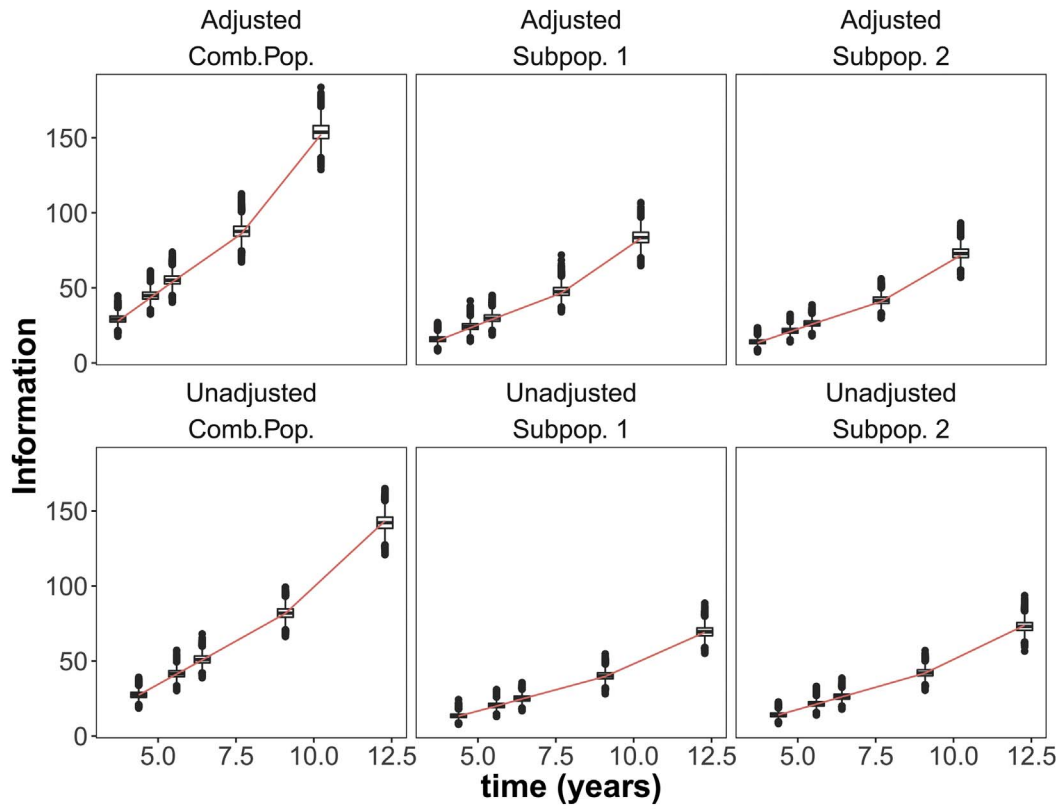
For trials using the unadjusted estimator, because it only uses information in the observed primary outcome, the performance is not affected by the prognostic value of W or L, or the delay to the short-term outcome d_L.

The adjusted estimator is especially useful when there are strongly prognostic baseline variables or short-term outcome available, or when the primary outcome is measured with considerable delay while a prognostic, short-term outcome is observed relatively soon after enrollment. Our simulation results can inform trial planning that involves delayed response.

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 become stronger. It is an area of future research to explore such simulation scenarios, in which there is a trade-off that



(a)



(b)

Fig. 5. Information accrual rates and box-plots of estimated variance for the adjusted and unadjusted estimators under the default simulation scenario. (a) Information accrual under the default simulation scenario. Yellow corresponds to unadjusted estimator and black corresponds to adjusted estimator. (b) Box-plots of estimated information level for adjusted estimator (using inuence-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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shorter d_L results in more pipeline participants with L observed, but such L is less prognostic for Y.

Open research problems include investigating the impact of the subpopulation proportion differing from the assumed value, and generalizing the findings to other trial 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 can be inconsistent [12].

Throughout, we considered a continuous-valued primary outcome Y. For the case of a binary outcome, the adjusted estimator of [23] can be used, in which case R^2 (computed in the ordinary least squares sense, as described in their Section 6) is directly related to the asymptotic variance reduction due to adjustment. For a time-to-event outcome, one can use the adjusted estimator of [24]; in this case, there may not be a simple formula such as R^2 for computing the asymptotic variance reduction due to adjustment.

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 the editor and two referees for constructive suggestions. 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 Bioengineering, 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 organizatioin 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.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.conctc.2017.08.003.

References

[1] S.-J. Wang, R.T. O'Neill, H.M. Hung, Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset, Pharm. Stat. 6 (2007) 227–244.
[2] N. Stallard, T. Hamborg, N. Parsons, T. Friede, Adaptive designs for confirmatory clinical trials with subgroup selection, J. Biopharm. Stat. 24 (2014) 168–187.
[3] W. Brannath, E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, A. Racine-Poon, Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology, Stat. Med. 28 (2009) 1445–1463.
[4] F. Bretz, H. Schmidli, F. König, A. Racine, W. Maurer, Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts, Biom. J. 48 (2006) 623–634.
[5] C. Jennison, B.W. Turnbull, Adaptive seamless designs: selection and prospective testing of hypotheses, J. Biopharm. Stat. 17 (2007) 1135–1161.
[6] H. Schmidli, F. Bretz, A. Racine, W. Maurer, Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations, Biom. J. 48 (2006) 635–643.
[7] T. Friede, N. Parsons, N. Stallard, A conditional error function approach for subgroup selection in adaptive clinical trials, Stat. Med. 31 (2012) 4309–4320.
[8] B.P. Magnusson, B.W. Turnbull, Group sequential enrichment design incorporating subgroup selection, Stat. Med. 32 (2013) 2695–2714.
[9] N. Stallard, Group-sequential methods for adaptive seamless phase II/III clinical trials, J. Biopharm. Stat. 21 (2011) 787–801.
[10] M. Rosenblum, T. Qian, Y. Du, H. Qiu, A. Fisher, Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches, Biostatistics. (2016), http://dx.doi.org/10.1093/biostatistics/kxw014.
[11] Biogen, 221AD302 phase 3 study of Aducanumab (BIIB037) in early Alzheimer's disease (EMERGE), ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2016 May 11], 2016 Available at: https://clinicaltrials.gov/ct2/show/nct02484547?term=mci+biogen&rank=1.
[12] M.J. van der Laan, S. Gruber, Targeted minimum loss based estimation of causal effects of multiple time point interventions, Int. J. Biostat. 8 (2012).
[13] S. Gruber, M. van der Laan, Targeted minimum loss based estimator that outperforms a given estimator, Int. J. Biostat. 8 (2012).
[14] J. Schwab, S. Lendle, M. Petersen, M. van der Laan, Ltmle: Longitudinal Targeted Maximum Likelihood Estimation, R package version 0.9-5 (2015).
[15] X. Lu, A.A. Tsiatis, Semiparametric estimation of treatment effect with time-lagged response in the presence of informative censoring, Lifetime Data Anal. 17 (2011) 566–593.
[16] A. Rotnitzky, Q. Lei, M. Sued, J.M. Robins, Improved double-robust estimation in missing data and causal inference models, Biometrika 99 (2012) 439–456.
[17] E. Colantuoni, M. Rosenblum, Leveraging prognostic baseline variables to gain precision in randomized trials, Stat. Med. 34 (2015) 2602–2617.
[18] S. Gruber, M. van der Laan, A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome, Int. J. Biostat. 6 (2010) 1–18.
[19] A. Fisher, M. Rosenblum, Stochastic Optimization of Adaptive Enrichment Designs for Two Subpopulations, Johns Hopkins University, Dept. of Biostatistics Working Papers, 2016 Working Paper 279 http://biostats.bepress.com/jhubiostat/paper279.
[20] K.G. Lan, D.L. DeMets, Discrete sequential boundaries for clinical trials, Biometrika 70 (1983) 659–663.
[21] E. Slud, L. Wei, Two-sample repeated significance tests based on the modified Wilcoxon statistic, J. Am. Stat. Assoc. 77 (1982) 862–868.
[22] T. Qian, M. Rosenblum, H. Qiu, Improving Power in Group Sequential, Randomized Trials by Adjusting for Prognostic Baseline Variables and Short-term Outcomes, Johns Hopkins University, Dept. of Biostatistics Working Papers, 2016 Working Paper 285 http://biostats.bepress.com/jhubiostat/paper285.
[23] K.L. Moore, M.J. van der Laan, Covariate adjustment in randomized trials with binary outcomes: targeted maximum likelihood estimation, Stat. Med. 28 (2009) 39–64.
[24] J.C. Brooks, M.J. van der Laan, D.E. Singer, A.S. Go, et al., Targeted minimum loss-based estimation of causal effects in right-censored survival data with time-dependent covariates: Warfarin, stroke, and death in atrial fibrillation, J. Causal Inference 1 (2013) 235–254.