Published in final edited form as:

Stat Med. 2015 June 15; 34(13): 2104–2115. doi:10.1002/sim.6474.

Bayesian sequential monitoring design for two-arm randomized clinical trials with noncompliance

Weining Shen^{*,a}, Jing Ning^{*,a}, and Ying Yuan^a

^a Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, U.S.A.

Abstract

In early-phase clinical trials, interim monitoring is commonly conducted based on the estimated intent-to-treat effect, which is subject to bias in the presence of noncompliance. To address this issue, we propose a Bayesian sequential monitoring trial design based on the estimation of the causal effect using a principal stratification approach. The proposed design simultaneously considers efficacy and toxicity outcomes, and utilizes covariates to predict a patient's potential compliance behavior and identify the causal effects. Based on accumulating data, we continuously update the posterior estimates of the causal treatment effects, and adaptively make the go/no-go decision for the trial. Numerical results show that the proposed method has desirable operating characteristics and addresses the issue of noncompliance.

Keywords

Bayesian design; causal effect; continuous monitoring; noncompliance; principal stratification

1. Introduction

A major goal in early-phase clinical trials is to determine the promise of a new treatment for further large-scale development based on safety and efficacy evaluations. For ethical considerations, interim monitoring is commonly conducted to ensure an early termination of the trial when the new treatment shows inadequate efficacy or when related safety concerns arise. In this regard, various two-stage and multiple-stage designs have been proposed in the literature for phase II clinical trials. Examples include two-stage designs [1, 2], in which the trial is terminated at the first stage if the number of favorable responses is not sufficient. Simon [3] proposed the optimal two-stage design that minimizes the expected sample size, and which was further extended to more complicated scenarios such as multi-center trials [4], optimal three-stage designs [5] and adaptive designs [6].

Continuous monitoring has been proposed to further improve the efficiency of phase II trial designs. Many such methods have been developed within the Bayesian framework due to its flexibility. A practical Bayesian guideline proposed by Thall and Simon [7] declares the new

treatment as promising if the posterior probability of efficiency exceeds pre-specified thresholds. This idea has been extended for the evaluation of a multivariate outcome situation [8]. Heitjan [9] proposed a “persuasion probability” criteria based on the posterior probability of the new treatment being superior to the standard one. Recently a few flexible Bayesian designs have been proposed [10, 11] based on predictive probability and hypothesis testing instead of using the posterior probability. Almost all existing phase II trial designs assume that full compliance exists and that, in the case of randomized phase II trials, the estimate of the treatment effect used for implementing the monitoring rule is a consistent estimate of the true or causal treatment effect.

The issue of noncompliance, however, is not uncommon in practice [12] and complicates the assessment of the treatment effect and monitoring of the clinical trial. In this article, we focus on two-arm randomized phase II clinical trials that evaluate the efficacy of a new treatment versus that of the control. In the presence of noncompliance, the standard intent-to-treat (ITT) analysis may result in biased estimates of the true treatment effect, which may lead to early termination of the trial when the treatment actually has a significant treatment effect. Some could argue that the ITT analysis measures the actual treatment effect in the target population and that such early termination of the trial may not be of concern. Even so, for many research projects, it is still of scientific interest to disentangle the true treatment effect from noncompliance. For example, if an intervention has a superior (causal) treatment effect, further effort may be warranted to study and increase its compliance rate. The pros and cons of the ITT analysis and causal inference have been discussed by Heitjan [13].

The noncompliance issue has been extensively researched in the statistical literature. Popular approaches include the instrumental-variable method [14], structural equation models [15], and a variety of methods built under the potential-outcome framework [16, 17]. A comprehensive review has been given by Pearl [18]. The principal stratification approach [19] has proven to be very useful in handling noncompliance by making adjustment based on post-treatment variables to produce estimates of the causal effects. The key idea is to compare treatment outcomes with control outcomes for the subpopulation who will always comply with the treatment assignment (i.e., “compliers”). The estimated difference is called the principal effect and has a causal interpretation [16]. In recent years, there has been an active line of research that aims to increase the precision of the identification of compliers using pre-treatment covariates [20, 21]. Despite the rich body of literature on estimating causal effects with noncompliance, to the best of our knowledge, the noncompliance issue has not been addressed in the context of clinical trial design.

In this paper, we propose a Bayesian continuous monitoring design that accommodates noncompliance. Our approach is based on the estimation of complier average causal effect (CACE) under the principal stratification framework, where patients are classified into different strata or subpopulations based on their potential compliance behavior. The proposed design simultaneously considers efficacy and toxicity outcomes, and utilizes baseline covariates to predict the patient’s potential compliance behavior and to identify the causal effects. Based on the accumulating data, we continuously update the posterior estimates of the causal treatment effects, and adaptively make the go/no-go decision for the trial. In addition, the proposed method may help identify useful baseline covariates to

determine the complying subgroup for future drug development. Note that by “Bayesian sequential monitoring”, we mean that the monitoring process is updated as we collect patients’ outcomes sequentially. This terminology is used in continuous monitoring clinical trial designs [7, 8, 9], and should not be confused with dynamic or sequential Bayesian updating models.

The rest of the paper is organized as follows. We introduce a motivating example to illustrate the necessity of considering noncompliance in clinical trial designs in Section 2. In Section 3, we propose our methodology based on Bayesian principal stratification modeling and discuss the corresponding stopping rules. Several simulation studies and sensitivity analyses are carried out in Section 4 to evaluate the operating characteristics of the proposed design. We conclude with a discussion in Section 5.

2. A motivating example

It is estimated that cigarette smoking is responsible for over four million deaths worldwide each year. The successful quitting rate is very low (7.3% after an average of 10 months of follow-up [22]) among smokers who have made serious attempts at cessation without any assistance. Several medicines have been developed and have been proven useful in the treatment of nicotine withdrawal symptoms, resulting in quitting rates between 15% to 25% after a one-year follow-up. In this example, we consider a two-arm, placebo-controlled randomized phase II clinical trial that aims to evaluate the toxicity and efficacy of a new agent. The study participants are patients with head and neck cancer who were recruited from the communities of Galveston and Houston, Texas. The inclusion criteria: a smoking history of at least three years, a daily consumption of 10 or more cigarettes, no current use of a smoking cessation treatment, and no current history of a psychiatric disorder or an uncontrolled systemic illness. Patients with Symptom Checklist–90–Revised (SCL-90-R) t scores exceeding 65 were also excluded.

The primary efficacy outcome of interest is defined as the biochemically verified 7-day point prevalence abstinence at the 1-month follow-up. That is, participants who self-report not smoking (not even a puff) in the past seven days and have a negative saliva cotinine test will be considered abstinent. This is a frequently used and well-accepted operational definition of smoking abstinence [23]. In this trial, we also monitor the adverse events (or toxicity), including increased blood pressure (1-3 mmHg), apathy or dry mouth, insomnia, difficulty urinating, urinary retention, loss of appetite, nausea, constipation, erectile dysfunction and reduced libido.

Based on the investigators’ previous experiences, the compliance rate of patients assigned to the new agent is expected to be around 70%. The compliance status of patients will be collected through questionnaires given to patients during the trial. Therefore, it is important to take into account the noncompliance issue when designing the trial. As the investigational agent is not publicly available, it is reasonable to assume that patients assigned to the control arm do not have access to that agent. In other words, the standard one-sided access assumption is reasonable for this trial. In addition, several previous studies have shown evidence of a strong association between the compliance of patients and their alcohol

consumption, anxiety level, age, gender and other characteristics [24, 25]. This motivates us to consider the approach of using the patient's baseline covariates to predict compliance, thereby circumventing the common non-identifiability issue that plagues causal inference.

3. Method

3.1. Principal stratification: notation and assumptions

Consider a two-arm randomized clinical trial in which a total of N patients are sequentially enrolled. For the i th patient, let $\mathbf{X}_i$ denote the baseline covariates that are observed before randomization, and $Z_i = \{0, 1\}$ denote the treatment assignment, where $Z_i = 0$ if the patient is assigned to the control, and $Z_i = 1$ if the patient is assigned to the new experimental treatment. We observe a pair of binary outcomes $(Y_{E,i}, Y_{T,i})$, where $Y_{E,i}$ is the treatment efficacy indicator (1 = efficacy, 0 = no efficacy) and $Y_{T,i}$ is the binary toxicity indicator (1 = toxicity, 0 = no toxicity). Let Y_i be a multinomial variable denoting four possible combinations of efficacy and toxicity outcomes, with $Y_i = 1$ if $(Y_{E,i}, Y_{T,i}) = (0, 0)$ (i.e., no efficacy, no toxicity), $Y_i = 2$ if $(Y_{E,i}, Y_{T,i}) = (0, 1)$ (i.e., no efficacy, toxicity), $Y_i = 3$ if $(Y_{E,i}, Y_{T,i}) = (1, 0)$ (i.e., efficacy, no toxicity), and $Y_i = 4$ if $(Y_{E,i}, Y_{T,i}) = (1, 1)$ (i.e., efficacy, toxicity).

Let $D_i(Z_i)$ be the treatment assignment that patient i actually received. Because of noncompliance, the treatment that the patient actually receives may be different from what he/she is assigned to, i.e., $D_i(Z_i)$ may be different from Z_i . The basic idea of the principal stratification approach is to classify patients into different strata based on their potential compliance status, and then make causal inference within the strata. Specifically, based on the values of $\{Z, D(Z)\}$, the principal stratification approach classifies patients into the following four principal strata (or subpopulations).

- **Compliers:** $D_i(0) = 0, D_i(1) = 1$. These are the patients who will comply with whatever treatment is assigned to them.
- **Never-takers:** $D_i(0) = 0, D_i(1) = 0$. These are the patients who will choose to take the control treatment (and not the experimental treatment) regardless of their randomized assignment.
- **Always-takers:** $D_i(0) = 1, D_i(1) = 1$. These are the patients who will choose to take the new treatment (and not the control) regardless of the assignment.
- **Defiers:** $D_i(0) = 1, D_i(1) = 0$. These are the patients who will not comply with whatever treatment is assigned to them;

In our setting, we focus on the case of one-sided access, i.e., the patients assigned to receive the new treatment can access the control treatment, but the patients assigned to the control treatment do not have access to the new treatment. The assumption of one-sided access is plausible in many randomized trials and has been widely adopted in causal inference [14]. In our motivating example, it is reasonable to assume one-sided access because the patients assigned to the control arm have no access to the new experimental agent, which is not available in the market. In this case, $D_i(0) \equiv 0$ and thus “always-takers” and “defiers” do not exist, and we need to consider only two principal strata, namely the “complier” and “non-

taker". Hereafter, we denote the principal stratum of patient i by S_i , with $S_i = c$ meaning "complier" and $S_i = n$ meaning "never-taker". Under the one-sided access assumption, the value of S_i is observable for patients who are assigned to the experimental treatment arm because we observe the value of $D_i(1)$ and $D_i(0) \equiv 0$. However, for patients assigned to the control, S_i is not directly observable because their $D_i(1)$ is unknown. Note that in the control, it is possible that some patients may not take the placebo. Because the placebo should not have any causal treatment effect, for these patients, we still assume $D_i(0) = 0$.

In order to identify the causal effect, we make the following assumptions,

- (A1) Stable unit treatment value assumption (SUTVA): This assumption has two components. The first component assumes no interference, i.e., the response of a particular patient depends on only his/her own treatment, not the treatments of other patients. The second component requires that there is only one version of each treatment, i.e., no variation in each treatment.
- (A2) Randomization: The treatment assignment Z is independent of the patients' potential outcomes, compliance status and baseline covariates $(Y, D, S, \mathbf{X})$.
- (A3) Exclusion restriction: $Y(D=0, Z=1) = Y(D=0, Z=0)$. In other words, noncompliant patients who are assigned to the treatment arm are effectively the same as the patients assigned to the control arm.
- (A4) One-sided access: $D(Z=0)$ is always 0.

Assumptions (A1)–(A3) are standard assumptions that have been commonly used in the causal inference literature (e.g., [14]). In the motivating example, (A1) is reasonable because the patients are recruited from different areas and we expect that the interaction between participants is negligible. That is, one patient's smoking status will not be affected by another patient's treatment assignment. Also, the patients in the same arm will receive the same dose of pills. Assumption (A2) is also reasonable since the treatment is assigned randomly. For Assumption (A3), we assume that patients who are assigned to the new experimental agent but do not take the agent (i.e., noncompliant) have similar outcomes with the patients who are assigned to the control arm, that is, there is no placebo effect. In this regard, our proposed method can also be viewed as an instrumental variable approach [14]. Assumption (A4), as we have discussed above, is also plausible in our example. It has been widely used in the literature [26, 27]. This assumption is sometimes relaxed to the monotonicity assumption [28], which will include another stratum, "always-takers". Under the monotonicity assumption, it is still possible to model the compliance using baseline covariates [29, 30]. For convenience in this paper, we consider only the stronger one-sided access assumption.

Under assumptions (A1)–(A4), the causal efficacy effect is defined as

$$\theta_E = Pr(Y_E = 1 | S=c, Z=1) - Pr(Y_E = 1 | S=c, Z=0), \quad (1)$$

that is, the difference between the average probability of observing efficacy in the treatment arm and the control arm within the complier stratum. Similarly, the causal toxicity effect is defined as

$$\theta_T = Pr(Y_T=1|S=c, Z=1) - Pr(Y_T=1|S=c, Z=0) \quad (2)$$

Based on the estimates of these causal effects, we can adaptively make the decision of continuing the trial or not.

The major difficulty of the inference of θ_E and θ_T , lies in the fact that the stratum S is not fully observed for the individuals who are assigned to the control arm. In order to identify their compliance stratum, we adopt a prediction model using baseline covariates for simultaneous identification of the principal strata and estimation of the causal effect. This idea has been explored in the principal stratification literature [20, 29, 30], and has a natural connection with the commonly used propensity-score method that identifies study participants in the control group who are likely to comply with the treatment [21, 31, 32, 33].

3.2. Compliance and response models

We propose to use the baseline covariate information of patients to assist in identifying the principal strata. In particular, we consider the following model that links the probability of compliance with the covariates:

$$Pr(S_i=c) = \rho(\mathbf{X}_i\beta), \quad i=1, \dots, N, \quad (3)$$

where ρ is a function taking values between 0 and 1, and indexed by a vector of parameters $\beta \in \mathbb{R}^d$. In practice, we may choose ρ as a logistic link function [20] or the cumulative distribution function of a standard normal distribution [30].

This compliance model plays several important roles in our design. First, it overcomes the identification problem of the principal strata for individuals who are assigned to the control arm, so that the CACE can be estimated based on the observed data. Second, it can be used to identify the compliant and noncompliant subgroups, which may benefit differently from the treatment. As pointed out by a referee, the question we address here can be viewed as a subgroup identification problem: the trial's target population consists of two subgroups (compliant/noncompliant) that may have different responses to the treatment. In addition to determining whether the drug is effective in the compliant group, it is also of interest to identify the compliant group for future drug development. For example, in the motivating trial, if we found that older patients with no alcohol consumption had better compliance and the treatment was effective for the compliant group, we might use that subpopulation as the target population for future treatment development, e.g., a therapy that combines the experimental drug with a behavioral intervention. Last, the compliance model can inform us how compliance can be increased in future studies to improve the benefit to the ITT population. For example, if depression is a strong predictor for noncompliance, we may combine the experimental (smoking cessation) drug with an anti-depression drug or intervention to achieve better treatment effects. Although investigators may take measures to minimize noncompliance in the trial, in many cases, the mechanism of noncompliance is not clear and it is difficult to develop effective strategies to avoid noncompliance at the stage of designing the trial. Identifying the factors that are predictive of noncompliance can be useful

for developing strategies to decrease noncompliance and improve the benefit to the ITT population.

Conditional on the (latent) strata S_i , we assume that the response Y_i follows a multinomial distribution

$$Y_i | S_i, Z_i \stackrel{i.i.d}{\sim} \text{Multi}(\theta(S_i, Z_i)), \quad i=1, \dots, N,$$

where parameters $\theta(S_i, Z_i) = (\theta_1(S_i, Z_i), \theta_2(S_i, Z_i), \theta_3(S_i, Z_i), \theta_4(S_i, Z_i))^T$, with $\theta_j(S_i, Z_i)$ denoting the probability of $Y=j$ in strata S_i under the assignment Z_i , where $j \in (1, 2, 3, 4)$.

The parameters satisfy the constraint $\sum_{j=1}^4 \theta_j(S_i, Z_i) = 1$ for every $S_i = c, n$ and $Z_i = 0, 1$. Under this model, the causal efficacy effects defined by equation (1) can be identified through the model parameters as follows,

$$\theta_E = \theta_3(c, 1) + \theta_4(c, 1) - \theta_3(c, 0) - \theta_4(c, 0).$$

Similarly, the causal toxicity effects θ_T can be obtained as

$$\theta_T = \theta_2(c, 1) + \theta_4(c, 1) - \theta_2(c, 0) - \theta_4(c, 0).$$

For the i th patient assigned to the control arm, since the compliance S_i is not observed, the associated likelihood is a two-component mixture given by

$$L_{c,i}(Y_i | Z_i, \mathbf{X}_i, \theta, \beta) = \rho(\mathbf{X}_i | \beta) \prod_{j=1}^4 \theta_j(c, 0)^{1l(Y_i=j)} + \{1 - \rho(\mathbf{X}_i | \beta)\} \prod_{j=1}^4 \theta_j(n, 0)^{1l(Y_i=j)}$$

where $1l(\cdot)$ is an indicator function and $\theta = (\theta(c, 0), \theta(c, 1), \theta(n, 0), \theta(n, 1))^T$. For the i th patient assigned to the treatment arm, his/her compliance S_i is observed and the corresponding likelihood is given by

$$\begin{aligned} L_{t,i}(Y_i | Z_i, \mathbf{X}_i, \theta, \beta) \\ = 1l(S_i=c) \rho(\mathbf{X}_i | \beta) \prod_{j=1}^4 \theta_j(c, 0)^{1l(Y_i=j)} + 1l(S_i=n) \{1 - \rho(\mathbf{X}_i | \beta)\} \prod_{j=1}^4 \theta_j(n, 0)^{1l(Y_i=j)} \end{aligned}$$

Let $p(\theta)$ and $p(\beta)$ denote independent prior distributions of θ and β , and define I_{ctl} and I_{trt} as the collection of patients assigned to the control and treatment arms, respectively. The posterior distribution of θ and β is given by

$$p(\theta, \beta | Y, \mathbf{X}, Z) \propto p(\theta) p(\beta) \prod_{i \in I_{\text{ctl}}} L_{c,i}(Y_i | Z_i, \mathbf{X}_i, \theta, \beta) \prod_{i \in I_{\text{trt}}} L_{t,i}(Y_i | Z_i, \mathbf{X}_i, \theta, \beta).$$

In our simulation, we take a vague normal prior for β , i.e., $p(\beta) = N(0, \sigma^2 \mathbf{I}_d \times d)$, where σ^2 is a large constant (e.g., 100), and assign $\theta(S, Z)$, $S = c, n$ and $Z = 0, 1$, independent Dirichlet priors $p(\theta(S, Z)) = \text{Dir}(1, 1, 1, 1)$.

3.3. Posterior sampling

We sample the posterior distribution using Gibbs sampling. To facilitate the posterior sampling, we treat the missing values of S_i as unknown parameters and assign independent Bernoulli priors on them, $S_i \stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p_s)$, $i \in I_{\text{ctl}}$, where p_s is a hyperparameter that takes a value of 0.5 in our simulation study. Then we can sample the parameters of interest from their full conditional distribution. One iteration of the Gibbs sampling is given as follows.

- For each patient in the control arm, we draw missing S_i , for every $i \in I_{\text{ctl}}$, from Bernoulli distributions $\text{Bern}(p_{i,u})$, where

$$p_{i,u} = \frac{p_s \rho(\mathbf{X}_i \beta) \theta_{Y_i}(c, 0)}{p_s \rho(\mathbf{X}_i \beta) \theta_{Y_i}(c, 0) + (1 - p_s) \{1 - \rho(\mathbf{X}_i \beta)\} \theta_{Y_i}(n, 0)}.$$

- Draw θ from its conditional distribution

$$\theta(s, z) | \beta, \mathbf{X}, Y, S, Z \sim \text{Dir}(1 + n_1(s, z), 1 + n_2(s, z), 1 + n_3(s, z), 1 + n_4(s, z)),$$

where $n_j(s, z) = \sum_{i=1}^N \mathbb{1}(Y_i = j, S_i = s, Z_i = z)$ is the number of patients with response $Y_i = j$ under principal stratum s and treatment z .

- Draw β from its full conditional distribution using the Metropolis-Hasting algorithm as follows: draw a candidate β^{new} from a normal distribution $N(\beta^{\text{old}}, \mathbf{I}_d \times d)$, where β^{old} is the value of β in the last step, and set $\beta = \beta^{\text{new}}$ with acceptance probability

$$\min \left\{ 1, \frac{\phi_{\sigma}(\beta^{\text{new}}) \prod_{i=1}^N \exp(\mathbf{X}_i^T \beta^{\text{new}})^{\mathbb{1}(S_i=1)} (1 + \exp(\mathbf{X}_i^T \beta^{\text{old}}))}{\phi_{\sigma}(\beta^{\text{old}}) \prod_{i=1}^N \exp(\mathbf{X}_i^T \beta^{\text{old}})^{\mathbb{1}(S_i=1)} (1 + \exp(\mathbf{X}_i^T \beta^{\text{new}}))} \right\},$$

where ϕ_{σ} is the probability density function of $N(0, \sigma^2 \mathbf{I}_d \times d)$.

3.4. Stopping rules

Depending on the design goals, various Bayesian stopping rules can be constructed based on the posterior distribution of the casual effects θ_E and θ_T . To be consistent with practice, we assume that patients are treated in cohorts and the stopping rule takes effect after the first n_0 cohorts are treated. A commonly used stopping rule is that we stop the trial if, with respect to the control, the causal efficacy effect of the experimental treatment is lower than a lower bound $\frac{\pi}{E}$, or the toxicity of the experimental treatment is higher than an upper limit $\frac{\pi}{T}$. That is,

After treating the first n_0 cohorts of patients, at any time during the trial, we stop the trial early for futility if $\Pr\left(\theta_E \lll \pi_{-E}\right) ggg C_E$, or stop the trial for toxicity if $\Pr\left(\theta_T ggg \bar{\pi}_T\right) ggg C_T$, where C_E and C_T are pre-specified cutoffs obtained by simulation calibration; otherwise we continue the trial until it reaches the maximum sample size $n_{\max}$.

Other stopping rules are certainly possible. For example, in some trials, it is more suitable to monitor the absolute toxicity rate of the experimental treatment rather than the relative toxicity rate (with respect to the control). In this case, we can modify the toxicity monitoring

rule as follows: We stop the trial for toxicity if $\Pr\left(\theta_{T1} ggg \bar{\pi}_T\right) ggg C_T$, where $\theta_{T1} = \Pr(Y_T = 1|S = c, Z = 1)$ is the marginal toxicity rate in the treatment arm. In other situations, we may prefer monitoring the trial based on the tradeoff between toxicity and efficacy. We can elicit a tradeoff function $g(\theta_E, \theta_T)$ to measure the desirability of the treatment, where $g(\cdot)$ maps the bivariate variables (θ_E, θ_T) into the real line; and then monitor the trial as follows: stop the trial if $\Pr\left(g(\theta_E, \theta_T) \lll \pi_{-U}\right) ggg C_U$, where π_{-U} is the lower bound of the utility and C_U is a cutoff.

4. Numerical results

4.1. Operating characteristics

We carry out a simulation study to evaluate the performance of the proposed method under the setting of the smoking cessation example in Section 2. We consider a two-arm randomized trial in which $n_0 = 30$ initial patients enter the study and then we start our sequential monitoring procedure with a cohort size of 5. The maximum sample size is set to be $n_{\max} = 150$. Outcomes Y_E and Y_T are generated independently from Bernoulli distributions with $\Pr(Y_E = 1|\text{control}) = 20\%$, $\Pr(Y_T = 1|\text{control}) = 10\%$ in the control arm, and with $\Pr(Y_E = 1|\text{treatment}) = p_E$, $\Pr(Y_T = 1|\text{treatment}) = p_T$ in the treatment arm, where p_E takes values in $\{0.2, 0.3, 0.4, 0.5\}$ and p_T takes values in $\{0.1, 0.4\}$. We generate two baseline covariates X_1 from a Bernoulli distribution, with success probability 0.4 and X_2 from a uniform distribution ranging between 0 and 2. We assume that the true compliance model is given by

$$\Pr(S=c) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}. \quad (4)$$

We consider the following three choices of $\beta = (\beta_0, \beta_1, \beta_2)^T$:

- Low noncompliance: $\beta = (2.3, -5, 6)^T$; the average noncompliance rate is 10%.
- Medium noncompliance: $\beta = (1, -6.8, 5)^T$; the average noncompliance rate is 25%.
- High noncompliance: $\beta = (1.2, -8, 3)^T$; the average noncompliance rate is 40%.

We apply the proposed Bayesian principal stratification (BPS) approach to monitor the trial with the following stopping rule: we stop the trial if either (1) $\Pr(\theta_E \leq 7.5\%|\text{data}) \geq 0.9$, that is, the data show that there is a high probability that the improvement in the efficacy rate for the new treatment is lower than 7.5%, or (2) $\Pr\{\theta_T \geq 20\%|\text{data}\} \geq 0.9$, that is, the data show that there is high probability that the toxicity rate of the new treatment is higher than 20%. We use 8000 MCMC iterations to fit the model, with 3000 as burn-in iterations. The acceptance ratio in Metropolis-Hastings sampling is controlled between 30-50%. We compare our results with the results obtained from the following three additional methods.

- (1) ITT method, where we ignore the compliance information and estimate θ in both treatment and control groups.
- (2) The gold standard (gold), where we use the true compliance stratum of each patient for the estimation of θ . The gold standard is not available in practice because the compliance stratum cannot be observed for patients in the control group, but it can serve as the “optimal” benchmark for comparison.
- (3) The local average treatment effect method (LATE) proposed by Imbens and Angrist [28]. More specifically, we consider an estimator for the efficacy effect based only on the observed compliance status,

$$\theta_E^{\text{LATE}} = \frac{E(Y_E|Z=1) - E(Y_E|Z=0)}{E(D|Z=1) - E(D|Z=0)} = \frac{\theta_E^{\text{ITT}}}{Pr(D=1|Z=1)},$$

where θ_E^{ITT} is the ITT efficacy estimator. Note that LATE will always produce a higher efficacy estimate than ITT. For toxicity, since we are monitoring the marginal proportion in the new drug, there is no need to consider its causal effect.

Table 1 shows the operating characteristics of the proposed design, including the percentage of early terminations and the average sample size (i.e., the number of patients enrolled in the study) based on 1000 replications. Our method performs uniformly better than the ITT in the sense that the percentage of early termination and the average sample sizes are closer to those under the gold standard. In particular, the difference in the percentage of early termination between our method and the gold standard is below 3% in most scenarios. In contrast, the performance of the ITT method varies across scenarios. When the noncompliance rate is low (10%), the performance of the ITT method is close to that of the gold standard. However, when the noncompliance rate gets higher (25% or 40%) and the toxicity level is low (10%), the ITT method tends to overly terminate the trial (e.g., termination proportion is higher by 5-24%) when the toxicity level is low (10%); and when the toxicity level is high (40%), the ITT method fails to stop the trial in a timely manner and treats 10-20 more patients in many situations.

Comparing ITT with LATE, we find that when the trial should not be terminated early (i.e., the new drug has high efficacy and low toxicity values), ITT tends to mistakenly stop the trial early because the efficacy effect is underestimated. In contrast, LATE can successfully correct the bias. LATE does not work well when the toxicity rate is high because it uses the

ITT estimate for toxicity. Comparing LATE with BPS, we see a significant improvement in considering the baseline covariate information in the causal effect modeling process.

4.2. Sensitivity analysis

We first investigate the robustness property of our method when the compliance model is mis-specified. We generate the data from the following true compliance model

$$\Pr(S=c) = \frac{\exp(1 - 6.8X_1 + 5X_2 + X_3)}{1 + \exp(1 - 6.8X_1 + 5X_2 + X_3)} \quad (5)$$

while the fitting compliance model is given by (4). That is, we miss a covariate X_3 in our model. We consider the following two cases:

- Case 1: X_3 is generated from $N(0, 4)$.
- Case 2: X_3 is generated from a uniform distribution on $(-2.5, -0.5)$.

In both cases, X_3 is generated independently with X_1 and X_2 , and the noncompliance rate is controlled at 25%. We also consider the situation where the logistic link function in the compliance prediction model (4) is incorrect. The data are generated from the following model:

$$\text{Case 3: } \Pr(S=c) = \Phi\{(1 - 6.8X_1 + 5X_2)/2\} \quad (\text{average noncompliance rate } 18\%),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$. The absolute values of the differences in the compliance probability under the above three cases with the original “medium noncompliance” model are 9%, 10% and 14%, respectively. We summarize the average sample size and the early termination rate based on 1000 replications in Table 2. We also compare the results with those obtained from the proposed method under the correctly specified model (BPS*). We find that the results of BPS are very close to those of BPS* (termination difference < 2%) in all scenarios, which indicates that the proposed method is robust. In most cases, the numbers from the BPS* are between those of the gold standard and the BPS.

Next, we consider the consequence when the one-sided access assumption (A4) is violated. We let the patients in the control group take the new treatment with probabilities of 0 (no violation), 0.05, 0.1 and 0.2, and compare their performances under the proposed BPS method. Relevant results based on 1000 replications are summarized in Table 3. In theory, when the efficacy is higher in the new treatment group, the violation of the one-sided access assumption will lead to an underestimate of the efficacy and hence increase the early termination percentage. This is verified by most of the entries in the table. A small violation (5% or 10%) of the assumption has little effect (the average sample size decreases by at most 8 patients) on the results. When the toxicity is high, i.e., $p_T = .40$, then the assumption violation essentially has no influence on the results.

We also consider the consequence when the exclusion restriction assumption (A3) is violated. For noncompliant patients assigned to the treatment arm (i.e., $D_i(Z_i = 1) = 0$),

instead of generating their outcomes in the same way as the outcomes of the patients assigned to the control arm, we consider the following four situations,

- No change: $\text{Pr}(\text{efficacy}) = .20$ and $\text{Pr}(\text{toxicity}) = .10$.
- Low efficacy: $\text{Pr}(\text{efficacy}) = .10$ and $\text{Pr}(\text{toxicity}) = .10$.
- High efficacy: $\text{Pr}(\text{efficacy}) = .30$ and $\text{Pr}(\text{toxicity}) = .10$.
- High toxicity: $\text{Pr}(\text{efficacy}) = .20$ and $\text{Pr}(\text{toxicity}) = .30$.

We evaluate the performance of the proposed BPS method under these situations and summarize the results based on 1000 replications in Table 4. In theory, when a noncompliant patient experiences higher treatment efficacy than a patient assigned to the control arm, the efficacy advantage of the new treatment relative to that of the control will be smaller, which results in a higher termination rate. Similarly, for the “low efficacy” situation, the termination rate will be smaller. This is verified by most of the entries in the table. Nevertheless, the effect of violation of the exclusion restriction assumption on the stopping rules seems minor, and the early termination rate is changed by 4% at most.

5. Discussion

In this paper, we propose a sequential monitoring design that handles noncompliance of patients based on a flexible Bayesian principal stratum prediction model. Numerical results show improvement over the existing ITT-based approaches by modeling the compliance strata of the patients. Such results may also help identify sub-populations that have a low adherence rate in the clinical trial. Therefore, the proposed method may be extended to future work by considering adaptive enrichment of the inclusion/exclusion criteria in patient recruitment, and hence increase the compliance rate in future trials.

Here we emphasize the trade-off that arises from using the CACE instead of ITT in sequential monitoring. Although using CACE provides greater accuracy in making early-termination decisions in the presence of compliance and identifying the compliant subgroup among the patients, it would require additional assumptions (e.g., (A3) and (A4)), and the performance would also rely on correct specification of the compliance model and the inclusion of informative covariates, which may be difficult to test and verify in practice. Some previous work has shown that the estimation of the causal effect may be sensitive to the choice of covariates in the model [34]. To address this concern, we discussed the validity of the listed assumptions for the motivating smoking cessation example, and carried out several sensitivity analyses. Notably, the performance of the proposed method is not sensitive to the choice of compliance model, i.e., the functional form of ρ and the covariates included in the model, and the assumption of one-sided access and exclusion restriction. The simulation results are also not sensitive to the choice of the hyper-parameters in the prior distribution. For example, we have tried values of .7 and .9 for p_s in addition to the non-informative choice of .5, and found that the results do not vary much under these choices. In practice, we recommend that the users consult with experts about the *a priori* information regarding noncompliance and the trade-off associated with using the causal effect instead of ITT.

It is of interest to extend the current work in a few directions. The one-sided access assumption can be relaxed to the monotonicity assumption, in which case the stratum prediction model will be generalized to incorporate a stratum of “always-takers” [30]. Our method can also be generalized to trials with two active arms [29] and to multi-arm trials, though a more complicated definition of compliance stratum is required.

Acknowledgment

Yuan's research is partially supported by the National Cancer Institute grants CA154591, CA016672 and 5P50CA098258. Ning's research is partially supported by the grant R21 HL109479.

References

1. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases*. 1961; 13:346–353. [PubMed: 13704181]
2. Fleming TR. One-sample multiple testing procedure for phase ii clinical trials. *Biometrics*. 1982; 38:143–151. [PubMed: 7082756]
3. Simon RM. Optimal two-stage designs for phase ii clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
4. Green SJ, Dahlberg S. Planned versus attained design in phase ii clinical trials. *Statistics in Medicine*. 1992; 11:853–862. [PubMed: 1604065]
5. Chen TT. Optimal three-stage designs for phase ii cancer clinical trials. *Statistics in Medicine*. 1997; 16:2701–2711. [PubMed: 9421870]
6. Lin Y, Shih WJ. Adaptive two-stage designs for single-arm phase iia cancer clinical trials. *Biometrics*. 2004; 60:482–490. [PubMed: 15180674]
7. Thall PF, Simon R. Practical bayesian guidelines for phase iib clinical trials. *Biometrics*. 1994; 50:337–349. [PubMed: 7980801]
8. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine*. 1995; 14:357–379. [PubMed: 7746977]
9. Heitjan DF. Bayesian interim analysis of phase ii cancer clinical trials. *Statistics in Medicine*. 1997; 16:1791–1802. [PubMed: 9280033]
10. Lee JJ, Liu DD. A predictive probability design for phase ii cancer clinical trials. *Clinical Trials*. 2008; 5:93–106. [PubMed: 18375647]
11. Johnson VE, Cook JD. Bayesian design of single-arm phase ii clinical trials with continuous monitoring. *Clinical Trials*. 2009; 6:217–226. [PubMed: 19528131]
12. Pullar T, Kumar S, Feely M. Compliance in clinical trials. *Annals of the Rheumatic Diseases*. 1989; 48:871–875. [PubMed: 2684057]
13. Heitjan DF. Causal inference in a clinical trial: a comparative example. *Controlled clinical trials*. 1999; 20:309–318. [PubMed: 10440558]
14. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*. 1996; 91:444–455.
15. Robins JM. Correcting for non-compliance in randomised trials using structural nested mean models. *Communications in Statistics*. 1994; 23:2379–2412.
16. Imbens GW, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*. 1997; 25:305–327.
17. Frangakis CE, Rubin DB. Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment- noncompliance and subsequent missing outcomes. *Biometrika*. 1999; 86:365–379.
18. Pearl J. Causal inference in statistics: An overview. *Statistics Surveys*. 2009; 3:96–146.
19. Frangakis CE, Rubin DB. Principal stratification in causal inference. *Biometrics*. 2002; 58:21–29. [PubMed: 11890317]

20. Hirano K, Imbens GW, Rubin DB, Zhou X. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*. 2000; 1:69–88. [PubMed: 12933526]
21. Jo B, Stuart EA. On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*. 2009; 28:2857–2875. [PubMed: 19610131]
22. Baillie AJ, Mattick RP, Hall W. Quitting smoking: estimation by meta-analysis of the rate of unaided smoking cessation. *Australian and New Zealand Journal of Public Health*. 1995; 19:129–131.
23. SRNT Subcommittee on Biochemical Verification. Biochemical verification of tobacco use and cessation. *Nicotine & Tobacco Research*. 2002; 4:149–159. [PubMed: 12028847]
24. Kahler CW, Spillane NS, Metrik J, Leventhal AM, Monti PM. Sensation seeking as a predictor of treatment compliance and smoking cessation treatment outcomes in heavy social drinkers. *Pharmacology Biochemistry and Behavior*. 2009; 93:285–290.
25. Sokolovsky AW, Mermelstein RJ, Hedeker D. Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine and Tobacco Research*. 2014; 16:351–358. [PubMed: 24097816]
26. Zelen M. A new design for randomized clinical trials. *New England Journal of Medicine*. 1979; 300:1242–1245. [PubMed: 431682]
27. Efron B, Feldman D. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*. 1991; 86:9–26.
28. Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica*. 1994; 62:467–475.
29. Roy J, Hogan JW, Marcus BH. Principal stratification with predictors of compliance for randomized trials with 2 active treatments. *Biostatistics*. 2008; 9:277–289. [PubMed: 17681993]
30. Zigler CM, Belin TR. The potential for bias in principal causal effect estimation when treatment received depends on a key covariate. *The Annals of Applied Statistics*. 2011; 5:1876–1892. [PubMed: 22308190]
31. Hill JL, Brooks-Gunn J, Waldfogel J. Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychobiology*. 2003; 39:730–744.
32. Follmann DA. On the effect of treatment among would-be treatment compliers: An analysis of the multiple risk factor intervention trial. *Journal of the American Statistical Association*. 2000; 95:1101–1109.
33. Joffe MM, Small D, Hsu CY. Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Statistical Science*. 2007; 22:74–97.
34. Jo B. Model misspecification sensitivity analysis in estimating causal effects of interventions with non-compliance. *Statistics in Medicine*. 2002; 21:3161–3181. [PubMed: 12375297]

Table 1

Simulation results of the proposed method (BPS), intent-to-treat (ITT) analysis, local average treatment effect method (LATE), and gold standard method (Gold) under different noncompliance (NC) rates. The efficacy and toxicity rate for the control are 0.2 and 0.1, respectively.

NC rate	Treatment		% of early termination					Sample size				
	p_T	p_E	BPS	ITT	LATE	Gold	BPS	ITT	LATE	Gold		
10%	.10	.20	92.6	93.8	90.6	93.1	59.6	56.6	62.5	58.4		
		.30	53.5	60.4	53.3	54.0	99.8	93.6	100.0	98.5		
		.40	19.6	25.8	21.7	20.1	129.8	124.2	127.4	129.4		
		.50	7.5	8.5	6.8	7.7	141.9	140.7	142.4	141.5		
		.40	.20	100	100	100	100	32.0	32.8	33.4	31.9	
	.20	.30	99.8	99.8	99.8	99.9	33.7	34.9	35.2	33.7		
		.40	99.8	99.6	99.5	99.8	35.1	36.6	36.9	35.0		
		.50	99.9	99.6	99.6	99.9	34.8	36.4	36.5	34.7		
		25%	.10	.20	88.8	94.0	84.2	89.8	64.4	55.7	70.6	62.4
				.30	51.6	66.3	50.8	56.1	103.0	88.1	104.4	98.4
.40	20.9			32.8	20.2	24.6	129.0	118.3	130.4	125.4		
.50	9.5			15.2	8.8	10.7	139.4	134.0	140.4	138.3		
.40	.20			99.9	99.6	99.3	100	33.6	35.6	38.5	33.2	
.20	.30		99.9	99.2	98.2	99.9	34.7	37.6	40.0	34.5		
	.40		99.8	98.7	98.3	99.8	35.7	41.4	43.5	35.7		
	.50		99.8	97.9	97.6	99.7	36.1	42.7	44.1	35.8		
	40%		.10	.20	84.8	92.0	74.1	86.1	69.9	57.6	82.7	67.6
				.30	52.3	76.9	47.1	55.2	102.1	77.4	106.8	98.4
.40		23.0		48.4	22.7	24.8	127.2	104.5	128.7	125.1		
.50		12.2		24.7	10.3	12.9	136.6	125.7	139.0	135.7		
.40		.20		99.8	98.7	95.9	99.8	34.9	39.7	48.2	34.7	
.20		.30	99.7	96.7	93.6	99.7	36.1	44.5	52.8	35.8		
		.40	98.7	92.4	89.9	98.7	37.8	50.8	56.1	37.7		
		.50	99.3	90.0	87.6	99.2	38.4	55.7	59.3	38.6		

Table 2

Sensitivity analysis under mis-specified compliance models (cases 1–3) for the proposed method (BPS), the intent-to-treat analysis (ITT), gold standard method (Gold) and the BPS with correctly specified compliance model (BPS*).

Case	Treatment		% of early termination					Sample size		
	p_T	p_E	BPS	ITT	Gold	BPS*	BPS	ITT	Gold	BPS*
1	.10	.20	86.0	92.8	88.7	86.9	67.9	57.1	64.3	66.1
		.30	48.1	67.7	52.6	50.2	106.6	86.7	100.7	105.2
		.40	22.4	38.1	26.2	24.0	127.6	114.3	124.1	126.5
	.40	.50	9.6	15.2	10.4	9.2	139.1	133.8	138.4	139.6
		.20	100	99.8	100	100	33.4	35.4	33.0	33.4
		.30	99.5	98.9	99.6	99.5	35.7	39.8	35.3	35.5
2	.10	.40	99.4	97.4	99.6	99.6	36.6	43.3	36.4	36.4
		.50	99.1	96.4	99.1	99.1	37.8	46.1	37.8	37.8
		.20	85.5	92.9	86.3	85.7	70.2	57.1	66.9	70.2
	.40	.30	51.8	75.4	56.1	52.9	103.0	80.9	98.2	102.7
		.40	21.3	42.5	24.6	21.9	128.1	110.0	125.0	127.6
		.50	10.5	21.2	10.9	10.7	138.7	128.8	137.9	138.2
3	.10	.20	99.7	99.3	99.8	99.8	34.6	37.7	34.5	34.5
		.30	99.5	98.4	99.4	99.4	35.7	42.0	35.9	35.8
		.40	99.2	94.9	99.1	99.1	36.7	47.8	36.8	36.7
	.40	.50	99.8	94.5	99.7	99.7	37.5	50.5	37.7	37.5
		.20	86.6	93.4	88.5	86.6	67.3	56.5	63.7	67.7
		.30	51.2	68.3	56.1	49.9	103.1	85.7	98.2	104.0
		.40	20.3	35.3	24.4	20.5	129.3	117.5	125.6	129.4
		.50	8.6	14.7	9.8	8.5	140.3	134.3	139.2	140.5
		.20	100	99.7	100	100	33.9	36.2	33.8	33.9
		.30	99.8	98.7	99.8	99.8	35.4	39.6	35.0	35.3
		.40	99.3	97.6	99.4	99.3	36.8	42.6	36.7	37.0
		.50	99.5	96.7	99.5	99.5	36.7	45.3	36.5	36.7

Table 3

Sensitivity analysis under the violation of the one-sided access assumption, with different noncompliance (NC) rates.

NC rate	Treatment		% of violation					% of violation				
	p_T	p_E	0	5	10	20	0	5	10	20		
<hr/>												
			% of termination					Sample size				
<hr/>												
10%	.10	.30	56.4	56.1	57.7	65.5	98.2	99.4	98.4	90.2		
		.40	19.0	20.9	27.2	31.8	130.9	128.9	122.8	119.1		
		.50	8.0	7.4	10.1	12.7	141.2	142.0	139.3	136.8		
	.40	.20	100	100	100	100	32.5	32.5	32.1	32.1		
		.30	100	99.9	99.7	99.8	33.2	33.2	34.4	33.4		
	.40		99.9	100	99.9	99.7	34.2	33.9	34.7	34.8		
	.50		99.7	99.8	99.9	99.9	35.3	35.0	34.7	34.7		
<hr/>												
25%	.10	.30	51.9	51.4	55.4	58.4	104.1	103.6	97.9	96.3		
		.40	20.5	22.1	26.3	29.2	129.0	127.9	124.0	122.0		
		.50	8.8	12.0	12.7	16.7	140.3	137.1	136.4	132.7		
	.40	.20	99.9	99.9	99.9	99.9	32.6	33.4	33.5	33.8		
		.30	99.7	99.8	100	99.7	34.6	35.2	34.5	34.8		
	.40		99.6	99.4	99.6	99.7	36.8	35.6	36.1	36.3		
	.50		99.4	99.7	99.6	99.6	37.0	35.8	36.9	36.1		
<hr/>												
40%	.10	.30	52.9	52.4	53.7	59.0	102.1	102.8	100.7	96.0		
		.40	22.1	27.6	27.8	33.8	127.7	122.8	121.5	117.5		
		.50	12.1	13.9	16.7	20.6	136.7	134.8	132.1	128.4		
	.40	.20	100	99.9	99.9	99.8	34.1	34.4	34.9	34.3		
		.30	99.7	99.7	99.4	99.7	36.5	35.8	36.6	36.5		
	.40		99.0	99.4	99.1	99.2	38.0	37.0	38.0	38.0		
	.50		98.9	99.2	99.3	98.8	39.0	38.4	38.0	39.1		

Table 4

Sensitivity analysis under the violation of the exclusion restriction assumption, with different noncompliance (NC) rates.

NC rate	Treatment		Type of violation				Type of violation					
	p_T	p_E	None	Low eff	High eff	High tox	None	Low eff	High eff	High tox		
Sample size												
10%	.10	.20	92.6	92.6	92.0	91.3	59.6	59.2	57.8	58.7		
		.30	53.5	54.9	54.8	53.7	99.8	99.7	99.9	101.0		
		.40	19.6	17.7	21.2	20.9	129.8	132.0	129.0	129.4		
		.50	7.5	6.7	5.9	8.4	141.9	142.5	143.2	140.4		
		.40	20	100	100	100	32.0	32.0	32.4	32.0		
	.40	.30	99.8	99.8	100	100	33.7	33.9	33.7	33.5		
		.40	99.8	99.8	99.9	99.8	35.1	34.5	34.7	34.4		
		.50	99.9	99.9	99.8	99.7	34.83	34.4	34.4	35.2		
		25%	.10	.20	88.8	88.0	87.9	86.6	64.4	65.4	66.0	67.8
				.30	51.6	47.8	52.1	52.7	103.0	106.9	102.1	102.2
.40	20.9			20.7	21.9	20.6	129.0	129.7	128.2	129.6		
.50	9.5			9.9	7.0	9.5	139.4	139.3	141.9	139.5		
.20	99.9			100	100	100	33.6	32.7	33.4	33.1		
.40	.30		99.9	99.9	99.5	99.9	34.7	34.5	35.1	34.5		
	.40		99.8	99.6	99.7	99.5	35.7	36.0	36.1	36.4		
	.50		99.8	99.8	99.8	99.5	36.1	36.3	35.8	36.8		
	40%		.10	.20	84.8	86.2	84.2	85.0	69.9	69.3	70.8	69.9
				.30	52.3	52.6	49.4	52.6	102.1	101.9	104.9	100.3
.40		23.0		22.7	26.4	24.5	127.2	127.5	122.4	125.3		
.50		12.3		12.4	12.7	12.7	136.6	136.5	136.0	136.3		
.40		20		99.8	100	100	34.9	34.1	34.3	33.9		
.40		.30	99.7	99.4	99.7	99.3	36.1	35.8	35.9	36.7		
		.40	98.7	99.7	99.2	99.5	37.8	37.8	38.3	38.4		
		.50	99.3	99.2	99.6	99.2	38.4	39.0	37.7	37.3		