A Structured Framework for Adaptively Incorporating External Evidence in Sequentially Monitored Clinical Trials

Evan Kwiatkowski[1], Eugenio Andraca-Carrera[2], Mat Soukup[2], Matthew A. Psioda[1]

[1] Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB#7420

[2]Division of Biometrics VII, Office of Biostatistics, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA

Abstract: We present a Bayesian framework for sequential monitoring that allows for use of external data, and that can be applied in a wide range of clinical trial applications. The basis for this framework is the idea that, in many cases, specification of priors used for sequential monitoring and the stopping criteria can be semi-algorithmic byproducts of the trial hypotheses and relevant external data, simplifying the process of prior elicitation. Monitoring priors are defined using the family of generalized normal distributions which comprise a flexible class of priors, naturally allowing one to construct a prior that is peaked or flat about the parameter values thought to be most likely. External data are incorporated into the monitoring process though mixing an a priori skeptical prior with an enthusiastic prior using a weight that can be fixed or adaptively estimated. In particular, we introduce an adaptive monitoring prior for efficacy evaluation which dynamically weighs skeptical and enthusiastic prior components based on the degree to which observed data are consistent with an enthusiastic perspective. The proposed approach allows for prospective and pre-specified use of external data in the monitoring procedure. We illustrate the method for both single-arm and two-arm randomized controlled trials. For the latter case, we also

include a retrospective analysis of actual trial data using the proposed adaptive sequential monitoring procedure. Both examples are motivated by completed pediatric trials, and the designs incorporate information from adult trials to varying degrees. Preposterior analysis and frequentist operating characteristics of each trial design are discussed.

Keywords: Adaptive Trial Design, Bayesian Sequential Monitoring, Information Borrowing, Pediatric Trials, Skeptical Prior.

1 Introduction

In the United States, sponsors of trials evaluating new drugs, biologics, and devices are required to monitor these trials (U.S. Food and Drug Administration 2006). While monitoring of trials takes on various forms, one form of monitoring is to assess the safety and efficacy of a product in an ongoing trial at pre-defined intervals (i.e. interim analyses), typically through an independent data monitoring committee. The most commonly used statistical approach for interim analyses uses frequentist group sequential methods in which Type I error for testing a null and alternative hypothesis is distributed across the set of interim analyses to ensure overall Type I error control for establishing the efficacy or futility of a product (Jennison & Turnbull 2000). Alternatively, in Bayesian sequential monitoring, data can be monitored on a continual basis and the evidence in favor (or against) a hypothesis can be evaluated against a single standard without penalty (Spiegelhalter et al. 1993).

In the clinical development of a therapeutic product, external forces may impede a clinical trial from reaching its objective (e.g. difficulty in enrolling, limited patient populations, long latency in observing the outcome of interest). This challenge is especially apparent in cases where the disease for which the investigational product (IP) is an intended treatment is rare or where the focus is on a pediatric population. In settings where patients are difficult to enroll, and therefore meaningful numbers of patients will complete follow-up prior to the trial reaching full enrollment, the concept of frequently monitoring interim data to determine whether a trial (or enrollment) can be stopped becomes appealing. As such, the use of Bayesian sequential methods, rooted in the likelihood principle and thus completely consistent with frequent or even continual data monitoring, provide an ideal basis from which to develop novel designs - a goal under the 21st Century CURES Act (U.S. Congress 2016).

For traditional frequentist approaches, pre-existing information (e.g., historical data) are often used to determine a plausible, clinically meaningful value for the treatment effect to use for power calculations, but is not generally used in analysis once data are actually available. In contrast, the Bayesian paradigm provides a natural framework for incorporating

information into both the design and analysis of a future trial. For analysis purposes, pre-existing information is typically translated into a prior distribution that characterizes what is currently believed about the treatment effect. See, for example, Psioda & Ibrahim (2018) and the references therein for recent work on general Bayesian methods for trial design and analysis using historical data.

In this paper, we propose a strategy for designing sequentially monitored clinical trials that entails eliciting priors used to monitor enrollment and/or data collection (i.e., monitoring priors) and stopping criteria that can be derived in a semi-automatic fashion based on standard inputs that are required for trial planning. These inputs include (1) the boundary null value for the treatment effect, (2) a plausible, clinically meaningful value for the treatment effect, and (3) a criteria for what constitutes a compelling demonstration of efficacy. In principle, the plausible, clinically meaningful value for the treatment effect should be informed by relevant external data.

A key contribution of this work is the provision of structured definitions of skeptical and enthusiastic perspectives that can be used to inform early stopping decisions in favor of efficacy and futility, respectively. Skeptical and enthusiastic priors are developed using the generalized normal family of distributions. This flexible family includes the normal distribution as a special case, and provides the capacity to construct monitoring priors that reflect nuanced prior opinion about the treatment effect. A conditional-marginal prior factorization is proposed for settings where there are one or more nuisance parameters, and we illustrate how prior information can be used in both marginal distribution for the treatment effect and conditional distribution for the nuisance parameters. The structured definitions of skeptical and enthusiastic perspectives form the basis for an adaptive monitoring prior to be used for efficacy evaluations monitoring when there is a desire to incorporate prior information into the monitoring process. The prospective use of external data in a pre-specified design provides novelty beyond conducting sensitivity analyses with different priors, and provides a pathway for innovative designs.

We perform simulation-based preposterior analysis to examine a variety of operating characteristics for the proposed design framework, and to understand how key operating characteristics are influenced by the frequency of monitoring. Specifically, we estimate the probability of stopping early at an interim analysis due to a compelling demonstration of efficacy or futility, and the sample size at the interim and final analyses. In most cases, patients will be ongoing in the trial at the time interim data are obtained that lead to ending enrollment as a result of a compelling demonstration of efficacy. It is our assumption that in most cases these patients will complete the study protocol and, accordingly, we also explore the degree that interim evidence changes on average, once final data are available.

Bayesian sequential designs are often restricted to have explicit frequentist properties (Ventz & Trippa 2015, Zhu & Yu 2015). Prior work has shown such restrictions can result in Bayesian and frequentist designs that have stopping rules which are nearly identical (Stallard et al. 2020, Kopp-Schneider et al. 2020, Zhu et al. 2019). While it is possible to calibrate a Bayesian design to have specific frequentist operating characteristics, we do not advocate for that strategy. Instead, we propose a Bayesian framework that leverages what the authors argue is an intuitive criteria for stopping enrollment and/or data collection at any point (based on posterior inference using a consistent criteria for a compelling demonstration) without explicit focus on strict type I error control – something that is not achievable when prior information is incorporated into the analysis (Psioda & Ibrahim 2018).

This paper is organized as follows: Section 2.1 reviews Bayesian hypothesis testing using posterior probabilities and the use of skeptical and enthusiastic priors for efficacy and futility monitoring. Section 2.2 presents a method for parameterizing monitoring priors the generalized normal distribution and for incorporating prior information into the monitoring priors, and a method to specify priors for nuisance parameters. Examples are given in Section 3, with Section 3.1 presenting an example based on a single-arm trial and Section 3.2 presenting an example based on a two-arm randomized, controlled trial. The purpose of Section 3.1 is to demonstrate using skeptical and enthusiastic priors that are peaked or flat about the

parameter values thought to be most likely, and the purpose of Section 3.2 is to demonstrate the adaptive monitoring prior in a more complicated setting that involves specifying a prior for a nuisance parameter. Section 3.3 contains a comparison of the adaptive monitoring prior with a standard design using a non-informative prior. Section 4 and 5 are not even mentioned. This was an oversight the first time.

2 Methods

2.1 Preliminaries

2.1.1 Bayesian Hypothesis Testing

Consider a clinical trial application where the primary objective is to test a hypothesis about an unknown quantity of interest which we denote by θ, with possible values for θ falling in the parameter space Θ. For example, in a single-arm trial with a binary response endpoint, $\theta \in (0, 1)$ may be the response probability associated with patients receiving the IP. In a two-arm trial with a binary response endpoint, $\theta \in (-1, 1)$ may be the difference in response probabilities between patients receiving the IP and those receiving the control treatment.

Throughout the paper we will let $\mathbf{D}$ represent the data collected in a trial at some point in time. For example, for the two-arm trial example above and assuming no covariates other than the treatment indicator, $\mathbf{D} = \{y_i, z_i : i = 1, ..., n\}$ where y_i is an indicator of response for patient i and z_i is an indicator for whether patient i was assigned the IP. We use the generic representation $p(\mathbf{D}|\theta, \eta)$ to reflect the density or mass function for the collective data $\mathbf{D}$ as a function of θ and potential nuisance parameters η, which could be multi-dimensional. For the two-arm trial example, η might correspond to the response probability for patients receiving the control treatment or some transformation thereof. For ease of exposition, for the remainder of Section 2.1 we will focus on the case where θ is the only unknown parameter.

Consider the hypothesis $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$. The posterior probability that

$\theta \in \Theta_i$ is given by

$$P(\theta \in \Theta_i|\mathbf{D}) = \frac{\int_{\Theta_i} p(\mathbf{D}|\theta)\pi(\theta)d\theta}{\int_{\Theta} p(\mathbf{D}|\theta)\pi(\theta)d\theta} \tag{1}$$

where $p(\mathbf{D}|\theta)$ is commonly referred to as the likelihood for θ and $\pi(\theta)$ is its prior distribution. We will also refer to $P(\theta \in \Theta_i|\mathbf{D})$ as the posterior probability of hypothesis H_i. See Appendix A for a brief discussion of the appropriateness of referring to $P(\theta \in \Theta_i|\mathbf{D})$ as the posterior probability of hypothesis H_i.

2.1.2 Formalizing the Statistical Concept of a Compelling Demonstration

Consider the one-sided hypotheses $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ for fixed θ_0, which we refer to as the boundary null value. Often in Bayesian hypothesis testing, one rejects the null hypothesis when $P(\theta > \theta_0|\mathbf{D})$ exceeds a prespecified threshold. Let ϵ represent *insignificant residual probabilistic uncertainty* regarding a claim. Define $1 - \epsilon$ to be the threshold for posterior probabilities in favor of the claim (e.g., that $\theta > \theta_0$), such that posterior probabilities above $1 - \epsilon$ are considered as providing *a compelling demonstration* that the claim is true. Leveraging common practice, we will use $\epsilon = 0.025$ for the examples presented herein so that $1 - \epsilon = 0.975$ is the threshold that determines when evidence of a claim is compelling. Our purpose in this paper is not to debate the appropriateness of using 0.975 as a threshold for defining a compelling demonstration, but rather to develop a strategy for prior elicitation that leverages an accepted threshold to simplify prior elicitation for sequentially monitored trials in hopes that this may facilitate the use of sequential monitoring more broadly and consistently.

Formally, we say that an individual whose belief is summarized by the distribution $\pi(\theta)$ is *all but convinced* that H_i is true if

$$P_\pi(\theta \in \Theta_i) = 1 - \epsilon, \tag{2}$$

where the subscript π in (2) is simply to indicate that the probability is calculated based on $\pi(\theta)$ which could be either a prior or posterior distribution.

2.1.3 Skeptical and Enthusiastic Monitoring Priors

Monitoring priors are used for interim analyses of the data, and the purpose of monitoring priors is to help answer the question "Is the evidence compelling enough to stop enrollment for the trial, or possibly end it altogether?" A promising interim analysis that provides a compelling demonstration of efficacy may justify ending enrollment, while enrolled patients may continue to receive the treatment for the pre-planned period of exposure. A discouraging interim analysis that provides a compelling demonstration of futility may justify ending enrollment, and may call for enrolled patients who are ongoing in the trial to be transitioned off the IP (i.e., termination of investigation of the treatment). For the Bayesian, the question becomes "From what prior perspective must the evidence be compelling to justify one of the two actions described above?" This motivates skeptical and enthusiastic monitoring priors, which represent two extreme but plausible beliefs about the quantity of interest θ relative to the hypotheses considered.

Having formalized concepts for *a compelling demonstration* and being *all but convinced* of a claim, we now can develop a structured framework for constructing skeptical and enthusiastic monitoring priors which will be used to determine early stopping rules for efficacy and futility, respectively. Consider again the hypotheses $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta > \theta_0$ where θ_0 represents a treatment effect of interest and let $\theta_1 > \theta_0$ represent a plausible, clinically meaningful effect. Define an enthusiastic prior, denoted as $\pi_E(\theta)$, as a prior consistent with θ_1 being the most likely value of θ (i.e., the prior mode) and that reflects the belief of an observer who is *all but convinced* that H_1 is true a priori. Formally, this is defined as satisfying (i) $\text{argmax}_\theta\ \pi_E(\theta) = \theta_1$ and (ii) $P_E(\theta > \theta_0) = 1 - \epsilon$, where the subscript E indicates that the probability is based on $\pi_E(\theta)$. Similarly, define a skeptical prior, denoted as $\pi_S(\theta)$, as a prior consistent with θ_0 being the most likely value of θ and that reflects the belief of

an observer who is *all but convinced* that $\theta < \theta_1$ is true a priori. Formally, this is defined as the prior $\pi_S(\theta)$ satisfying (iii) $\text{argmax}_\theta\ \pi_S(\theta) = \theta_0$ and (iv) $P_S(\theta < \theta_1) = 1 - \epsilon$. In what follows we refer to (i) and (iii) as *mode value constraints* and (ii) and (iv) as *tail-probability constraints*, respectively.

Note that the proposed development of the skeptical prior does not generally reflect skepticism regarding whether the alternative hypothesis is true. Indeed, assuming a symmetric skeptical prior is elicited (as we propose), the *induced* prior probabilities on the hypotheses satisfy $p(H_0) = p(H_1)$ as shown in Appendix A. Thus, the skeptical prior simply reflects skepticism regarding the possibility of large treatment effects but is otherwise consistent with clinical equipoise regarding the two hypotheses.

The totality of evidence in favor of a hypothesis is influenced by the prior distribution used for analysis. It is natural that one would stop a trial early in favor of efficacy or futility when the evidence in favor of the appropriate claim is compelling to a sufficiently skeptical or enthusiastic observer, respectively, as defined above. For example, if at any point data sufficiently convince an observer whose prior belief is in accordance with $\pi_S(\theta)$ that the alternative is true, then any less skeptical observer would also be convinced. Therefore, ceasing enrollment and possibly collection of additional data in order to assess whether the treatment is beneficial would be a reasonable action from almost any rational perspective. Similarly, if at any point data sufficiently convince an observer whose prior belief is in accordance with $\pi_E(\theta)$ that the effect of interest is significantly less than what was originally believed, then any less enthusiastic observer would be similarly convinced and ceasing the collection of data altogether would be a reasonable action from almost any rational perspective.

2.1.4 Maximum Sample Size and Formal Stoppage Criteria

In this section we formalize stopping criteria for futility and efficacy and give general advice for specifying a maximum sample size for the trial. Although sequentially monitored trials in principle require no fixed sample size, in practice due to resource constraints it will almost

always be the case that a maximum sample size exists. We recommend that (resources permitting) the maximum sample size, denoted by $n_{\max}$, should be chosen so that there is a high probability that the trial generates a compelling demonstration from the perspective of the skeptic when in fact $\theta \approx \theta_1$ in a scenario where the data are only examined once when the full set of outcomes are ascertained. The rationale behind this strategy is that one would want to ensure the trial's sample size is sufficient so that there is high probability the data collected will provide a compelling demonstration of treatment benefit to observers having relatively extreme skepticism regarding the magnitude of treatment benefit a priori.

For a sequentially monitored trial, observed data are analyzed as often as is feasible in accordance with the cost and/or logistical challenges of assembling the necessary data. For example, if an outcome requires adjudication by a committee of clinical experts, it may not be possible to reanalyze the data after each new patient's outcome is obtained due to scheduling or other constraints on the adjudication panel. In other scenarios, a patient's outcome may be based on a laboratory parameter's change after a fixed period of time and the rate limiting factor for sequential monitoring will be how quickly samples can be shipped, processed, and entered into a database for analysis. The strategies presented here for sequential monitoring are appropriate regardless of how frequently data can be monitored.

Stopping criteria are based on whether the posterior probability that the treatment effect is in a particular region is sufficiently large. For region Θ_i, this is formalized as

$$P_\pi(\theta \in \Theta_i | \mathbf{D}) > 1 - \epsilon, \tag{3}$$

where we note the inequality in (3) compared to the equality in (2). Stopping criteria for efficacy are defined from the perspective of a skeptical observer. The skeptic becomes convinced that a treatment is effective if at some point the observed data suggest there is a compelling demonstration that the alternative hypothesis is true. Formally, the early stopping criteria are met based on data $\mathbf{D}$ when $P_S(\theta > \theta_0 | \mathbf{D}) > 1 - \epsilon$. Note that the evidence

must *exceed* the threshold for what defines it as being compelling. When the evidence in favor of the alternative surpasses this threshold, it may no longer be necessary to enroll patients for the purpose of proving treatment efficacy.

Stopping criteria for futility monitoring are defined from the perspective of the enthusiastic observer. At first thought it may seem appealing to stop the trial when the enthusiast becomes convinced that the null hypothesis is true, that is, that $P_E(\theta \leq \theta_0|\mathbf{D}) > 1-\epsilon$. However, when $\theta = \theta_0$, $P_E(\theta \leq \theta_0|\mathbf{D})$ approaches 0.5 for large sample sizes. Therefore this potential futility criteria would not be satisfiable unless the observed data were consistent with values of θ much less than θ_0. For this reason, we consider a different approach. Recalling that θ_1 represents a plausible, clinically meaningful treatment effect, the early stopping criteria are met based on data $\mathbf{D}$ when $P_E(\theta < \theta_1|\mathbf{D}) > 1-\epsilon$. In this case the trial may be stopped due to there being a compelling demonstration that the treatment effect is much less than hypothesized (i.e., θ_1).

2.2 Specifying Monitoring Priors

2.2.1 Default Monitoring Priors

The skeptical and enthusiastic monitoring priors defined in Section 2.1.3 have mode value and tail-probability constraints. However, these constraints alone do not uniquely determine the priors. There are infinitely many distributions which satisfy these conditions. However, the mode and tail constraints do uniquely determine a pair of normal distributions which might serve as a default set of monitoring priors. A default enthusiastic monitoring prior satisfying (i) $\text{argmax}_\theta\ \pi_E(\theta) = \theta_1$ and (ii) $P_E(\theta > \theta_0) = 1-\epsilon$ is the normal distribution with location θ_1 and standard deviation $\sigma = \frac{\theta_1-\theta_0}{\Phi^{-1}(1-\epsilon)}$, where Φ^{-1} denotes the quantile function of a standard normal. The specification of μ and σ completely determine the density at all points, including the value of the density at the mode which is $f(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma}$. The skeptical monitoring prior is similarly defined, satisfying (i) $\text{argmax}_\theta\ \pi_S(\theta) = \theta_0$ and (ii) $P_S(\theta < \theta_1) = 1-\epsilon$.

Use of normal distributions for the monitoring priors can be motivated by the Bayesian Central Limit Theorem (CLT) (Le Cam & Yang 2000) which states that, under general conditions, the posterior distribution for θ approaches normality as the sample size increases, regardless of the initial choice of prior. Therefore, a normally distributed monitoring prior is consistent with belief derived from a sufficiently large dataset with maximum likelihood estimate equal to the mode value required by the prior.

2.2.2 Generalized Normal Distribution

Despite the aforementioned justification of normally distributed priors, it may be desirable to construct a monitoring prior with different behavior about the mode than what is possible when using the normal distribution. Choosing a flattened distribution is appropriate when one wishes to reflect more uncertainty regarding the likelihood that θ is near θ_1 (relative to what is permitted by the normal distribution), while maintaining the same residual uncertainty that $\theta < \theta_0$. Similarly, choosing a concentrated distribution is appropriate when one wishes to reflect a higher degree of certainty that θ is near θ_1, while maintaining residual uncertainty that $\theta < \theta_0$.

The family of generalized normal distributions, which contains the normal distribution as a special case, is able to accommodate changes in the density value at the mode while still satisfying the mode value and tail probability constraints. The density for a generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ is

$$f(\theta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|\theta - \mu|}{\alpha}\right)^{\beta}\right\}$$

where μ is a location parameter, $\alpha > 0$ is a scale parameter, and $\beta > 0$ is a shape parameter (Nadarajah 2005). Fixing the location parameter to be the mode value and changing the shape and scale parameters in conjunction can maintain the tail probability constraint while also changing the density's behavior near the mode. Recall the density at the mode

for a default enthusiastic prior is $f(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma}$. An enthusiastic monitoring prior in the generalized normal family of distributions can have density at the mode equal to $k \times \frac{1}{\sqrt{2\pi}\sigma}$, with $k < 1$ indicating a more flattened distribution and $k > 1$ indicating a more peaked distribution at the mode, relative to the default normal distribution. Appendix B details a procedure for parameterizing these flattened and concentrated monitoring priors. Flattened and concentrated distributions for different choices of k are shown in Figure 1.

Panel B of Figure 1 presents a concentrated skeptical prior that satisfies the mode value and tail probability constraints given in Section 2.1.3, and that has $k = 1.5$ times the density value at the mode as compared to the default normal distribution. Increasing the density value at the mode translates to a more peaked distribution about the mode as compared to the default normal distribution (shown in Panel A). Panel D of Figure 1 presents a flattened enthusiastic prior that satisfies the mode value and tail probability constraints given in Section 2.1.3, and that has $k = 0.67$ times the density value at the mode as compared to the default normal distribution. This translates to a distribution that is significantly more flat about the mode value than the default normal distribution (shown in Panel C). The approach we have proposed results in a unique flattened or concentrated prior.

The use of the scale factor k corresponds to beliefs that reflect useful and nuanced perspectives for clinical trial decision making. A concentrated skeptical prior views θ_0 as being more likely than a data-driven (i.e. normally distributed via Bayesian CLT) perspective while still reflecting residual uncertainty that $\theta > \theta_1$. This is a rational perspective for a monitoring prior since the skeptical viewpoint should give substantial preference to the null value. Similarly, a flattened enthusiastic prior views θ_1 as being less likely than a data-driven perspective while still reflecting residual uncertainty that $\theta < \theta_0$. This is a rational perspective for a monitoring prior since even from an enthusiastic viewpoint, one may wish to reflect increased uncertainty regarding the likelihood of values at and around θ_1. While there is no *correct* choice for the scale factor k for either a skeptical or an enthusiastic prior, the authors' choices of 1.5 and 0.67 are relatively extreme perturbations from that afforded by the normal

distribution and will be used henceforth to demonstrate the methodology proposed. Lastly, we note that the default normal, flattened, and concentrated priors all can be truncated while maintaining the mode and tail probability constraints. This will be necessary when the parameter of interest has bounded support (e.g., θ is a response probability).

2.2.3 Incorporating Prior Information in the Monitoring Priors

The monitoring priors are constructed based on the quantities θ_0 and θ_1, as well as the definition of a compelling demonstration. As described previously, prior information may be directly used in the construction of the enthusiastic prior (e.g., choice of θ_1). It also may be desirable to incorporate prior information into the monitoring process when making a determination of when to stop enrollment early for efficacy. To facilitate this, we introduce a procedure for modifying the monitoring process such that, if the enthusiastic prior is congruent with observed data, the degree of skepticism can be adaptively lessened. We propose incorporating prior information into the monitoring process for efficacy through constructing a mixture prior from the skeptical and enthusiastic priors using a mixing weight that is constructed from a measure of compatibility between the observed data and the enthusiastic prior. We define the *adaptive monitoring prior* for efficacy evaluations as the mixture distribution

$$\pi_{AE}(\theta) = \omega \cdot \pi_E(\theta) + (1 - \omega) \cdot \pi_S(\theta), \tag{4}$$

where $\omega \in (0, 1)$ is an adaptively determined mixing weight. The objective of the proposed approach is to create a mixture prior which favors the enthusiastic prior component in cases where high compatibility is observed between the trial data and the enthusiastic prior, and favors the skeptical prior component if the data observed are incompatible with the enthusiastic prior. This approach is motivated by the rationale that the enthusiastic prior reflects a plausible perspective about the treatment's effect, and one that we assume will typically be informed by data (e.g., from adult trials in the case of a planned pediatric trial). The adaptive monitoring prior $\pi_{AE}(\theta)$ is used to make determinations regarding treatment

efficacy for monitoring purposes and is a replacement for the traditional skeptical prior $\pi_S(\theta)$.

The adaptive mixing weight ω is determined by an assessment of prior-data conflict, proposed by Box (Box 1980), derived using the prior predictive distribution of the data which is defined (in our case) using the enthusiastic prior. The prior-predictive distribution for replicated data $\mathbf{D}_{\text{rep}}$ reflects the probability of observing hypothetical data $\mathbf{D}_{\text{rep}}$ given the assumed data generating mechanism and prior for θ, and is defined formally as

$$p(\mathbf{D}_{\text{rep}}) = \int p(\mathbf{D}_{\text{rep}}|\theta)\pi(\theta)d\theta. \tag{5}$$

Box's p-value is defined as the following:

$$\psi(\mathbf{D}) = \int p(\mathbf{D}_{\text{rep}})1[p(\mathbf{D}_{\text{rep}}) \leq p(\mathbf{D})]d(\mathbf{D}_{\text{rep}}) \tag{6}$$

where $1[A]$ is an indicator that the event A is true. Note that the expression in (6) can be viewed as an expectation of $1[p(\mathbf{D}_{\text{rep}}) \leq p(\mathbf{D})]$ with respect to random $\mathbf{D}_{\text{rep}}$ and is thus equal to the probability of observing a dataset $\mathbf{D}_{\text{rep}}$ as or less likely than $\mathbf{D}$. Thus, large values of $\psi(\mathbf{D})$ indicate the observed data are very well supported by the distribution in (5) whereas small values indicate the opposite. We propose using the enthusiastic prior $\pi_E(\theta)$ to compute the quantities in (5) and (6) to create a compatibility measurement $\psi^{(E)}(\mathbf{D})$ which is used to determine the mixing weight in (4). Use of Box's p-value as a measure of prior-data conflict has been considered previously (Psioda & Xue 2020), but not in the context of sequential monitoring or using a mixture prior framework as we proposed here.

Define the mixing weight ω given to the *enthusiastic* prior as

$$\omega = (1 - \delta) \cdot \psi^{(E)}(\mathbf{D}) \tag{7}$$

This mixture weight achieves the goal of favoring the enthusiastic component if the trial data are compatible with that prior, and otherwise assigning a higher weight to the skeptical

component. The minimum possible mixing weight δ assigned to the *skeptical* prior is achieved when $\psi^{(E)}(\mathbf{D}) = 1$ and is equal to δ. Choices of δ in $\{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$ are explored in Sections 3.2 and 4, and general advice for choosing δ is given in Section 5.

2.2.4 Prior Specification for Nuisance Parameters

Often there are additional parameters besides the treatment effect θ that are not of primary interest (i.e. nuisance parameters). It is necessary to elicit a prior distribution $\pi(\theta, \eta)$ for all unknown quantities. The marginal-conditional factorization of the joint prior $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ allows direct elicitation of the marginal prior on the treatment effect and provides the ability to incorporate prior information on the nuisance parameters through their conditional distribution given θ. The prior for $\pi(\theta)$ will be a generalized normal distribution that satisfies the aforementioned mode value and tail probability constraints. We propose to define $\pi(\eta|\theta)$ as a generalized normal distribution, with parameters chosen based on shape of the conditional distribution evaluated at the most likely value of $\pi(\theta)$. For example, if $\pi_S(\theta)$ is a skeptical prior, then the location, shape, and scale parameters for a generalized normal distribution for $\pi(\eta|\theta)$ will be chosen based on $\pi(\eta|\theta = \theta_0)$. The location parameter will be the most likely value of η when $\theta = \theta_0$ (e.g. mode$(\eta|\theta = \theta_0) = \eta_0$), and shape and scale parameters will be chosen to reflect a reasonable amount of uncertainty regarding η.

If the parameters θ and η are assumed to be independent, then the joint prior can be factored as $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta)$ and the priors $\pi(\theta)$ and $\pi(\eta)$ can be elicited separately. In some cases this is not possible. For example, suppose that θ is the risk difference between response probabilities of a treatment group and the control group, and denote the response probability in the control group by η. In this case θ and η are linked through constrained support (e.g. $0 \leq \theta + \eta \leq 1$). Such a prior specification is demonstrated in Figure 2, and Section 3.2.2 uses this representation of the joint prior. Panel A shows the marginal distribution $\pi(\theta)$, Panel B shows the conditional distribution $\pi(\eta|\theta = \theta_0)$, and Panel C shows

the joint prior $\pi(\theta, \eta)$. In this example, the conditional distribution $\pi(\eta|\theta)$ will look very similar to the marginal distribution of $\pi(\eta)$ except at the boundaries of the parameter space.

3 Examples

3.1 Single-Arm Trial with Binary Endpoint

3.1.1 Motivating Example

We consider the T72 pediatric trial "A Study of the Safety and Efficacy of Infliximab (REMICADE) in Pediatric Patients With Moderately to Severely Active Ulcerative Colitis" (NCT00336492) (Hyams et al. 2012) which was conducted between August 2006 and June 2010. The study population was patients ages 6 through 17 with moderate to severe ulcerative colitis defined as having a baseline Mayo score of 6 or above on a scale of 0-12, where higher scores indicate more severe disease activity. A 5mg/kg dose of infliximab was given to patients at weeks 0, 2, and 6. The primary endpoint was clinical response, corresponding to a 3-point or greater decrease in Mayo score from baseline to week 8. Patients were enrolled over approximately 33.5 months (approximately 1 patient enrolled per 17 days). The sample size of 60 patients was chosen so that a frequentist 95% two-sided confidence interval for the response probability would have a half-width of 0.12 if the true response probability is 0.67. The value 0.67 was the observed proportion of responders among adults with the same disease enrolled in the ACT 1 and ACT 2 trials (Rutgeerts et al. 2005) who received the same weight-based dose of 5mg/kg (N = 242). Obtaining a 95% confidence interval that excluded 0.40 was used as the criterion for classifying the results as clinically significant. Clinical response was observed in 44 of 60 (73.3%) pediatric patients.

3.1.2 Model Formulation & Prior Elicitation

We use this trial as a motivating example to demonstrate the proposed framework for sequential monitoring. The data $\mathbf{D}$ are assumed to be comprised of independent Bernoulli

random variables having common response probability θ. As mentioned above, the primary hypotheses evaluated in the trial were $H_0 : \theta \leq 0.4$ and $H_1 : \theta > 0.4$. For purposes of monitoring, we took $\theta_1 = 0.67$ consistent with the ACT 1 and ACT 2 trial data. The example presented in this section make use of a concentrated skeptical prior and a default enthusiastic prior for monitoring. For sequential monitoring, we consider analyzing the accumulating data after every two patients complete follow-up.

The early stopping criteria, as well as any quantity involving the posterior distribution of θ requires evaluating integrals of the dimension of θ (or the dimension of (θ, η) in the case of nuisance parameters). For the cases we consider in this paper, these quantities are $1-$ or $2-$dimensional integrals which are evaluated using numerical integration in R (R Core Team 2017) using the `pracma` package (Borchers 2019).

3.1.3 Example Paths

Figure 3 presents violin plots to illustrate the monitoring process for two hypothetical instances of the trial. Each instance shows the monitoring priors (left-most set of distributions), the posterior distributions at three selected interim analyses (middle sets), and the posterior distributions from the final analysis (right-most set). Panel A of Figure 3 shows the results of a trial with early stopping for efficacy once 30 outcomes are ascertained, and enrollment is henceforth terminated. The final data (i.e., the data after ongoing patients are followed-up) in this example path no longer meet the criteria for a compelling demonstration of efficacy. Panel B of Figure 3 shows the results of a trial with early stopping for futility once 30 outcomes are ascertained, and enrollment is henceforth terminated.

3.1.4 Choice of Monitoring Priors

Table 1 illustrates properties of the sequential monitoring procedure with different specifications of skeptical and enthusiastic monitoring priors, using the trial design as described in Section 3.1.2 with $100,000$ simulated trials per value of θ. When the skeptical prior is

concentrated, the probability of the efficacy criteria being satisfied when $\theta = 0.4$ is lower than when the skeptical prior is given the default specification. When the enthusiastic prior is flattened, the probability of futility criteria being satisfied when $\theta = 0.67$ is greater than when the enthusiastic prior is given the default specification, and expected sample sizes at $\theta = 0.4$ and $\theta = 0.535$ are lower than when the enthusiastic prior is given the default specification. General advice for choosing the specification of these priors is given in Section 5.

3.1.5 Preposterior Analysis of Operating Characteristics

The operating characteristics presented in this section are estimated using the concentrated skeptical prior and the default enthusiastic prior as shown in Table 1. As shown in Panel A of Figure 4, when the true response probability is $\theta_0 = 0.4$, the probability of stopping the trial early for efficacy is equal to 0.026, and at $\theta_1 = 0.67$ it is 0.953. Interim stoppage for either efficacy or futility occurred in each simulation (i.e., the maximum sample size was such that the trial could not reach the maximum). The expected sample sizes are the lowest when the true response probabilities are θ_0 or θ_1. This is because the trial is more likely to be stopped for futility or efficacy when the data are consistent with the skeptical or enthusiastic priors respectively, which have modes at θ_0 and θ_1.

Given that the early stopping criteria was satisfied as an interim analysis, it is of interest to compare the posterior probability of the alternative once patients in follow-up have completed outcomes using the same skeptical prior. It is of particular interest when the threshold for a compelling demonstration is satisfied for an interim analysis but is no longer satisfied once outcomes from patients in progress are ascertained, as was the case in Panel A of Figure 3. The probability of such an occurrence and the difference between the posterior probabilities evaluated at the different time points is shown in Panel B of Figure 4. The probability of these cases occurring is reflected by the percent agreement between interim and final results. Recall that when the generating value of theta is $\theta_0 = 0.4$, the trial is stopped

early for efficacy with probability 0.026. Among these cases, the posterior probability is still greater than $1 - \epsilon$ with probability 0.433. This means that the completed outcomes from patients in progress at time enrollment was terminated are likely to meaningfully diminish the evidence in favor of efficacy relative to the threshold for what is viewed as compelling. When the generating value of theta is $\theta_1 = 0.67$, the trial is stopped early for efficacy with probability 0.952, and among those cases the posterior probability is greater than $1 - \epsilon$ with probability 0.887.

The distribution of the posterior probability given the final data for these cases demonstrate that even in the cases where there is evidence decrease, the final posterior probability is still similar to the $1 - \epsilon = 0.975$ threshold. Consider $\theta_1 = 0.67$: in the 11.3% of situations where there is evidence decrease below the $1 - \epsilon = 0.975$ threshold, 90% of these cases have a final posterior probability of approximately 0.93 or greater.

3.2 Parallel Two-Group Design with Binary Endpoint

3.2.1 Motivating Example

We consider the trial "The Pediatric Lupus Trial of Belimumab Plus Background Standard Therapy (PLUTO)" (NCT01649765) which was conducted between September 2012 and January 2018 (Brunner et al. 2020). The study population was comprised of patients ages 5 though 17 with active systemic lupus erythematosus (SLE), defined as a baseline SELENA SLEDAI score of 6 or above on a scale of 0-105, where higher scores indicate more severe disease activity. Patients were randomized to monthly dosing of either belimumab 10mg/kg or placebo, while continuing to receive standard of care therapy regardless of assignment. The primary endpoint was a dichotomous variable reflecting a 4-point or greater reduction in SELENA SLEDAI score from baseline to week 52. The original study design included enrollment of 100 patients, the first 24 patients randomized in a 5:1 allocation ratio (belimumab:placebo) and the remaining 76 patients in a 1:1 ratio, resulting in 58 patients randomized to belimumab and 42 to placebo. The sample size was based on feasibility con-

straints rather than power considerations. Data from two studies of belimumab in adults having the same disease resulted in a placebo response probability of 0.39, and a 10mg/kg response probability of 0.51. Using these values for the null and hypothesized response probabilities for the treatment group and assuming a response probability of 0.39 for the control group, a frequentist two-sided hypothesis test with confidence level 95% and 80% power would require 266 patients per group. Ultimately, 93 patients were enrolled over approximately 52.5 months (approximately 1 patient enrolled per 17 days). Clinical response was observed in 28 of 53 (52.8%) of patients randomized to belimumab and in 17 of 40 (43.6%) of patients randomized to placebo.

3.2.2 Model Formulation & Prior Elicitation

We use this trial as a template to demonstrate our framework, in particular the performance of the adaptive monitoring prior defined in Section 2.2.3. The adaptive monitoring prior is necessary since the power analysis in Section 3.2.1 shows the need for many more patients than were available; therefore, a strategy for prospective incorporation of prior information must be implemented for the trial to have a chance of providing a compelling demonstration of efficacy through a pre-specified design. The data $\mathbf{D}$ are assumed to be independent Bernoulli random variables with response probability η_0 for the placebo group and η_1 for the treatment group, with $\theta = \eta_1 - \eta_0$ denoting the difference in response probabilities. This trial has a superiority hypothesis of treatment to control with null difference in response probabilities, denoted by $\theta_0 = 0$. An estimate for the pediatric response probability is denoted by $\eta_0 = 0.39$ (i.e. the sample proportion of responders from the pooled adult studies), and for purposes of monitoring, a plausible, clinically meaningful difference in response probabilities is $\theta_1 = 0.12$ (i.e. based on the pooled adult study's treatment response probability of 0.51).

The skeptical monitoring prior is $\pi_S(\theta, \eta_0) = \pi_S(\theta) \times \pi(\eta_0|\theta)$, where $\pi_S(\theta)$ is a concentrated skeptical prior. The enthusiastic monitoring prior is $\pi_E(\theta, \eta_0) = \pi_E(\theta) \times \pi(\eta_0|\theta)$,

where $\pi_E(\theta)$ is a default enthusiastic prior. The conditional prior for the nuisance parameter $\pi(\eta_0|\theta)$ is specified as a flattened prior around the conditional modal value of $\eta_0 = 0.39$. The probability of concluding efficacy at an interim analysis is made using the adpative monitoring prior as described in Section 2.2.3.

A maximum sample size of $n_{\max} = 100$ was chosen based on the original trial protocol. A minimum sample size of $n_{min} = 50$ was chosen to provide an adequate number of placebo controls to be enrolled given the initial 5:1 allocation to the treatment group. An interim analysis is completed after every two patients have outcomes beginning at n_{min}.

3.2.3 Preposterior Analysis of Operating Characteristics

Figure 5(A) shows the enthusiastic mixture weights ω by choice of δ in (7) for all combinations of response difference between the IP and PC groups, when the PC group is fixed at a 38% response rate (16/42 responses). Observe that the highest mixture weights ω are observed when the response differences are observed to be around 0.12, which was the mode value for the enthusiastic prior, and have maximum values of $1 - \delta$.

The operating characteristics presented in this section are estimated using 2,500 simulated trials per value of θ using the trial design as described in Section 3.2.2. The generating response probability in the placebo group was assumed to be 0.39, and the generating response probability in the treatment group was determined based on risk differences θ in $\{0, 0.03, 0.06, 0.09, 0.12\}$. Figure 5(B) shows the probability of stopping early for efficacy and the associated sample sizes when using the adaptive monitoring prior (4) with different choices of δ in (7). When $\delta = 0$ or $\delta = 0.05$, a conclusion of efficacy is made at an interim analysis 24% and 14% of the time respectively, while this value is 7% or lower when $\delta \geq 0.1$. Reductions in expected sample size are seen with lower choices of δ and higher generated risk differences. When $\delta = 0.1$, a demonstration of efficacy is observed in 53% of simulated trials, with an expected sample size of 90.1. Even though this is a modest reduction from the maximum sample size of 100 for this case, even more favorable reductions are possible when

enrollment is comparatively slower and/or when follow-up times are comparatively shorter.

3.3 Comparison to Single Analysis with Non-Informative Prior

To better understand the adaptive monitoring prior, we now compare a design based on it to one based on a non-informative (NI) prior. To simplify this illustration, we consider a design without interim analyses. The non-informative prior is taken to be a uniform prior over the joint parameter space of (θ, η) so that the risk difference θ is marginally uniform. We need the sample size, randomization scheme, etc. It is still not clear exactly what designs are being compared. What is assumed about the nuisance parameter in the data generation processes?

When using $\pi_{NI}(\theta)$, the probability of concluding treatment efficacy when $\theta = \theta_0$ is 0.025 (i.e., the type I error rate), and when $\theta = \theta_1$ it is 0.217 (i.e., the power). This shows that an analysis with the non-informative prior maintains the nominal type I error rate but, for the sample sizes used in the PLUTO trial, provides very low power. When using $\pi_{AE}(\theta)$ with $\delta = 0$, the probability of concluding efficacy is 0.189 when $\theta = \theta_0$ and is 0.675 when $\theta = \theta_1$. Thus, the design with the adaptive monitoring prior has substantially higher power and, necessarily, a substantially higher type I error rate. Note that if the design using the non-informative prior was allowed to have the same type I error rate (by decreasing the evidence threshold in equation (3)), the associated power would be similar, at 0.617. I am still concerned that the power is essentially 0.06 higher for the same type I error rate. I have not seen this before (I would have expected them to be exactly the same). We have to understand this, especially since the phenomenon is not present in the simulations described in blue. Also, is it possible to connect this set of results to Figure 5 to show how much the total type I error inflation in the Figure is attributable to sequential monitoring and how much is attributable to dynamic information borrowing? I think this connectivity would be more important than the part in blue (assuming we resolve the type I error rate issue).

Consider increasing the sample size to 161 subjects per group and changing the residual uncertainty to $\epsilon = 0.05$. When using a non-informative prior, the probability of concluding treatment efficacy when $\theta = \theta_0$ is 0.05 (i.e., the type I error rate), and when $\theta = \theta_1$ it is 0.70 (i.e., the power). When using $\pi_{AE}(\theta)$ with $\delta = 0$, the probability of concluding efficacy is 0.26 when $\theta = \theta_0$ and is 0.94 when $\theta = \theta_1$. If an analysis with the non-informative prior was permitted to have this same higher type I error rate by modifying the evidence threshold in equation (3), the associated power would identical, at 0.94. These analyses show that power is comparable across analyses that use different priors when the procedures are held to the same type I error control constraints.

Although possible to simply modify the critical value used for hypothesis testing within a traditional design framework, we do not advocate this approach. By doing so, one is implicitly deciding what level of evidence the pediatric data provide on their own, whereas a compelling demonstration of treatment effectiveness after synthesis of all pertinent information still requires additional post hoc evaluation. Instead, the proposed approach focuses on the end result, whether or not there is a compelling demonstration of treatment efficacy once all evidence has been synthesized and provides a clear and rigorous framework for synthesis that balances information borrowing with the need to act sensibly in the presence of prior-data conflict.

Reconsider modifications once other comments are addressed.

4 Real Data Example

We consider applying the adaptive monitoring prior of (4) to the observed outcomes of the PLUTO trial presented in Section 3.2. Responses were available for 92 patients (one subject in the placebo group had no outcome available due to a protocol violation). Sequential monitoring after every two completed outcomes was conducted after a minimum sample size of 50 had been reached. Results of this analysis by different choice of δ are shown in Table

2. When $\delta \leq 0.1$, a conclusion of efficacy is made before the maximum sample size of 92. Figure 6(A) shows Box's p-value at the observed data with 90 completed outcomes to be 0.965 which translates directly to the value of ω in the case that $\delta = 0$. Figure 6(B) shows the efficacy posterior probability of 0.979 when $\delta = 0$ so that $\omega = \psi^{(E)}(\mathbf{D})$.

We note that the final sample size is ≥ 90 for all choices of δ. Thus, in this application, due to the 52-week period of follow-up for the primary outcome and despite the slow enrollment, the impact of sequential monitoring would not have been substantial in terms of shortening the overall trial or reducing the number of patients enrolled. However, it would have nonetheless provided a mechanism for prospective incorporation of external evidence in a pre-specified manner for the trial. In situations where the time-to-outcome ascertainment is shorter and/or enrollment is slower relative to the time-to-outcome ascertainment, greater reductions in sample size would be expected.

5 Discussion

One comment was the significantly tighten up this section. Did we do anything? We have to do something to address the comment. Maybe this will free up space. "In my opinion, Section 5 of the paper can be improved by the authors taking a more critical view of their writing, sharpening the arguments, and including recommendations for the practitioners." In this paper, we present a structured framework for specifying monitoring priors and stoppage criteria for a Bayesian sequentially monitored clinical trial that is based on intuitive justification for the design quantities rather than being motivated by having pre-specified frequentist operating characteristics. Consequently, the choice of monitoring prior and stoppage criteria are the same regardless of the frequency of data monitoring and the number of patients in progress at enrollment termination, although these factors do impact the operating characteristics of the trial.

Our formulation of the enthusiastic prior enforces that there be residual uncertainty that

the null hypothesis is true; it demonstrates strong belief about effectiveness of the treatment yet is still consistent with a degree of equipoise. In the extreme case that interim data are observed to be perfectly consistent with the enthusiastic prior, the residual uncertainty that the null hypothesis is true reflected in the adaptive monitoring prior cannot be less than that reflected in the enthusiastic prior itself. This is a critical feature of the design as it enforces the requirement that observed data must demonstrate some degree of efficacy on their own to justify stopping enrollment early. Without maintaining residual uncertainty as we have done when constructing the enthusiastic prior, it would be possible to conclude benefit in cases where observed data are somewhat consistent with that prior (i.e., $\psi^{(E)}(\mathbf{D}) > 0$) but also consistent with no benefit (or even harm). This is particularly problematic when the observed data contain little information compared to the source that informs the enthusiastic prior (as is often the case in pediatric settings). Thus, the proposed approach provide a desirable assurance that evidence of efficacy must come, at least in part, from both the prior information and the trial data. A conclusion of treatment efficacy is possible only when there is overwhelming treatment benefit observed in the trial data so as to convince a skeptic on that data's own merit, or, in the more likely scenario, some evidence of benefit from the trial data along with reasonable compatibility with the enthusiastic prior.

Our results in Section 3.2 can be compared to a published post-hoc Bayesian hierarchical analysis (Brunner et al. 2020) which used data from two studies of the use of belimumab in adults. Patients in the pediatric trial had 1.5 times the odds of clinical response with 95% CI (0.6, 3.5), and a meta-analysis of the two adult studies showed an odds ratio of 1.6 with 95% CI (1.3, 2.1). The analysis used a mixture prior which was a weighted sum of a skeptical prior centered at null effect with effective sample size equal to two pediatric patients and an informative prior resulting from the meta-analysis. When the weight of the informative component was 0.55 and above, efficacy was concluded based on a 95% credible interval excluding one. The 0.55 weight of the informative component, interpreted as a 55% weight on the relevance of the adult information to the pediatric population, was determined

to be reasonable by the clinical team. Our method contrasts such a post-hoc analysis with the prospective use of a monitoring prior for efficacy which gives weight to the adult data at interim analyses, although both methods show the necessity of information borrowing.

Although the examples provided are based on superiority trials with binary endpoints and response probabilities as the parameter of interest, the framework applies to any type of data and parameter of interest. Future work will involve demonstrating the framework in Bayesian clinical trials with survival outcomes, such as large cardiovascular outcomes trials where frequent analysis of data may be useful to reduce excessive sample size requirements.

Software

A GitHub repository (https://github.com/psioda/Bayesian-Sequential-Monitoring) contains the programs and other resources needed to reproduce the analyses presented Examples 1 and 2 of this paper. Software was written using R 4.1.0.

Acknowledgments

To be written.

Declaration of Interest Statement

To be written.

References

Borchers, H. W. (2019), 'pracma: Practical Numerical Math Functions'.

URL: *https://cran.r-project.org/web/packages/pracma/*

Box, G. E. P. (1980), 'Sampling and Bayes' Inference in Scientific Modelling and Robustness', *Journal of the Royal Statistical Society. Series A (General)* **143**(4), 383–430.

Brunner, H. I., Abud-Mendoza, C., Viola, D. O., Calvo Penades, I., Levy, D., Anton, J. & et al. (2020), 'Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial', *Annals of the Rheumatic Diseases* **79**(10), 1340 LP – 1348.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on Scientific Computing* **16**(5), 1190–1208.

Griffin, M. (2018), 'Working with the exponential power distribution using gnorm'. **URL:** *https://cran.r-project.org/web/packages/gnorm/*

Hyams, J., Damaraju, L., Blank, M., Johanns, J., Guzzo, C., Winter, H. S. & et al. (2012), 'Induction and Maintenance Therapy With Infliximab for Children With Moderate to Severe Ulcerative Colitis', *Clinical Gastroenterology and Hepatology* **10**(4), 391 – 399.e1.

Jennison, C. & Turnbull, B. W. (2000), *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, Boca Raton.

Kopp-Schneider, A., Calderazzo, S. & Wiesenfarth, M. (2020), 'Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control', *Biometrical Journal* **62**(2), 361–374.

Le Cam, L. & Yang, G. L. (2000), *Asymptotics in Statistics: Some Basic Concepts*, Springer, New York.

Nadarajah, S. (2005), 'A generalized normal distribution', *Journal of Applied Statistics* **32**(7), 685–694.

Psioda, M. A. & Ibrahim, J. G. (2018), 'Bayesian clinical trial design using historical data

that inform the treatment effect', *Biostatistics* pp. kxy009–kxy009. 10.1093/biostatistics/kxy009.

Psioda, M. A. & Xue, X. (2020), 'A Bayesian Adaptive Two-Stage Design for Pediatric Clinical Trials', *Journal of Biopharmaceutical Statistics* .

R Core Team (2017), 'R: A Language and Environment for Statistical Computing'.

Rutgeerts, P., Sandborn, W. J., Feagan, B. G., Reinisch, W., Olson, A., Johanns, J. & et al. (2005), 'Infliximab for Induction and Maintenance Therapy for Ulcerative Colitis', *New England Journal of Medicine* **353**(23), 2462–2476.

Spiegelhalter, D. J., Freedman, L. S. & Parmar, M. K. B. (1993), 'Applying Bayesian ideas in drug development and clinical trials', *Statistics in Medicine* **12**(15-16), 1501–1511.

Stallard, N., Todd, S., Ryan, E. G. & Gates, S. (2020), 'Comparison of Bayesian and frequentist group-sequential clinical trial designs', *BMC Medical Research Methodology* **20**(1), 4.

U.S. Congress (2016), '21st Century Cures Act (Pubic Law 114-255, 130 STAT 1033-1344)'.

U.S. Food and Drug Administration (2006), 'Establishment and Operation of Clinical Trial Data Monitoring Committees'.

Ventz, S. & Trippa, L. (2015), 'Bayesian designs and the control of frequentist characteristics: A practical solution', *Biometrics* **71**(1), 218–226.

Zhu, H. & Yu, Q. (2015), 'A Bayesian sequential design using alpha spending function to control type I error', *Statistical Methods in Medical Research* **26**(5), 2184–2196.

Zhu, L., Yu, Q. & Mercante, D. E. (2019), 'A Bayesian Sequential Design for Clinical Trials with Time-to-Event Outcomes', *Statistics in biopharmaceutical research* **11**(4), 387–397.

Appendices

A: Bayesian Hypothesis Testing

Consider the hypotheses $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ where $\Theta_0 \bigcup \Theta_1 = \Theta$ and $\Theta_0 \bigcap \Theta_1 = \emptyset$. Formal Bayesian hypothesis testing requires the specification of prior probabilities on the hypotheses (e.g. $p(H_i)$ for $i = 0, 1$) and prior distributions for θ specified over the parameter space defined with respect to each of the hypotheses (e.g. $\pi(\theta \big| H_i)$ for $i = 0, 1$).

The posterior probability of hypothesis H_i is given by

$$p(H_i|\mathbf{D}) = \frac{p(\mathbf{D}|H_i) \cdot p(H_i)}{p(\mathbf{D}|H_0) \cdot p(H_0) + p(\mathbf{D}|H_1) \cdot p(H_1)}, \tag{8}$$

where $p(\mathbf{D}|H_i) = \int_{\Theta_i} p(\mathbf{D}|\theta)\pi(\theta|H_i)d\theta$ is the marginal likelihood associated with hypothesis H_i. In practice, most Bayesian hypothesis testing methods are based on the posterior probability of the *event defining* H_i. For this approach, one simply needs to specify a prior $\pi(\theta)$ representing belief about θ and compute the posterior distribution. The posterior probability that $\theta \in \Theta_i$ is given by

$$P(\theta \in \Theta_i|\mathbf{D}) = \frac{\int_{\Theta_i} p(\mathbf{D}|\theta)\pi(\theta|\theta \in \Theta_i)d\theta \cdot P(\theta \in \Theta_i)}{\sum_{j=0,1} \int_{\Theta_j} p(\mathbf{D}|\theta)\pi(\theta|\theta \in \Theta_j)d\theta \cdot P(\theta \in \Theta_j)} \tag{9}$$

where $P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta)d\theta$. We can readily see that the $P(\theta \in \Theta_i|\mathbf{D})$ is equal to $p(H_i|\mathbf{D})$ if one takes $p(H_i) = P(\theta \in \Theta_i)$ and $\pi\left(\theta \big| H_i\right) = \pi\left(\theta \big| \theta \in \Theta_i\right)$ for $i = 0, 1$. If in fact $\pi(\theta)$ does represent belief about θ, these choices are perhaps the most intuitive and thus we should have no reservation referring to $P(\theta \in \Theta_i|\mathbf{D})$ as the probability that hypothesis H_i is true.

B: Parameterizing Flattened and Concentrated Monitoring Priors

Recall the value of the normal density at the mode is $\frac{1}{\sqrt{2\pi}\sigma}$ and note that the value of a generalized normal density at the mode is $\frac{\beta}{2\alpha\Gamma(1/\beta)}$. These are equivalent when $\beta = 2$ and

$\alpha = \sqrt{2}\sigma$ (i.e. the normal density is a special case of the generalized normal density at these parameter values). Let $F_{\mu,\alpha,\beta}$ denote the cumulative distribution function of the generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$, which can be expressed as (Griffin 2018)

$$P(\theta \leq q|\mu, \alpha, \beta) = \frac{1}{2} + \frac{\text{sign}(q-\mu)}{2} \int_0^{|q-\mu|^\beta} \frac{w^{1/\beta - 1}}{\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{1}{\alpha}\right)^\beta w\right\} dw.$$

A flattened or concentrated enthusiastic monitoring prior in the generalized normal family of distributions has density at the mode equal to $k \times \frac{1}{\sqrt{2\pi}\sigma}$. The parameters for the generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ are derived as follows: μ remains equal to the mode value of θ_1 and α and β are determined to minimize the function

$$\left(F_{\mu,\alpha,\beta}(\theta_0) - \epsilon\right)^2 + \left(\frac{\beta}{2\alpha\Gamma(1/\beta)} - k\frac{1}{\sqrt{2\pi}\sigma}\right)^2$$

with box-constrained optimization (Byrd et al. 1995), where $\sigma = \frac{\theta_1 - \theta_0}{\Phi^{-1}(1-\epsilon)}$ is the standard deviation of the default normally distributed enthusiastic monitoring prior. The first term reflects the residual uncertainty that $\theta < \theta_0$, and the second term reflects the density at the mode value. Similarly, the parameters for a flattened or concentrated skeptical monitoring prior are as follows: μ remains equal to the mode value of θ_0 and α and β are determined to minimize the function

$$\left((1 - F_{\mu,\alpha,\beta}(\theta_1)) - \epsilon\right)^2 + \left(\frac{\beta}{2\alpha\Gamma(1/\beta)} - k\frac{1}{\sqrt{2\pi}\sigma}\right)^2,$$

where $\sigma = \frac{\theta_0 - \theta_1}{\Phi^{-1}(\epsilon)}$.

This parameterizing procedure is applicable to a generalized normal distribution truncated to an interval domain (e.g. when θ is a response probability with domain $[0, 1]$). In this case, the generalized normal distribution truncated to an interval domain $\Theta = (\theta_{min}, \theta_{max})$ has density equal to $f(\theta) = c \cdot \exp\left\{-\frac{|\theta-\mu|}{\alpha}^\beta\right\} I(\theta \in \Theta)$ where $c = \frac{\beta}{2\alpha\Gamma(1/\beta)}(F_{\mu,\alpha,\beta}(\theta_{max}) - F_{\mu,\alpha,\beta}(\theta_{min}))^{-1}$.

C: Step-by-Step Implementation Guide

Below are step-by-step instructions for implementing this method in the context of a non-inferiority trial.

1. Identify parameters for the trial design:

   (a) Specify null treatment effect θ_0 which is used to define the null and alternative hypotheses H_0 and H_1.

   (b) Specify threshold for a compelling demonstration $1 - \epsilon$.

   (c) Specify the plausible, clinically meaningful value for the treatment effect θ_1.

2. Create monitoring priors:

   (a) Choose prior shape for the skeptical and enthusiastic monitoring priors $\pi_S(\theta)$ and $\pi_E(\theta)$ (e.g. choose k as described in Section 2.2.1). Our recommendation is to use a concentrated specification (i.e., $k = 1.5$) for $\pi_S(\theta)$ and a default specification (i.e., a normal distribution which has asymptotic justification as belief arriving from a hypothetical dataset) for $\pi_E(\theta)$.

   (b) Solve for the parameters in the generalized normal distributions $\pi_S(\theta)$ and $\pi_E(\theta)$ as described in Section 2.2.1 and Appendix B. These parameters are determined by the quantities provided in Step 1 and Step 2(a). The code for this paper shows how these parameters were computed.

   (c) Repeat Steps 2(a,b) for nuisance parameters as described in Section 2.2.4.

   (d) If the adaptive monitoring prior $\pi_{AE}(\theta)$ is to be used, specify the minimum possible mixing weight δ assigned to the skeptical prior. Our recommendation is $\delta = 0.1$ so that a modest weight of at least 0.1 is always given to the skeptical prior.

3. Sequentially monitor the clinical trial:

(a) Iteratively conduct monitoring according to (3). If the adaptive monitoring prior is used, then compute the mixture weight ω using (7) at each iteration.

Tables

Table 1: Operating characteristics of Example 1 simulations based on REMICADE trial. $\pi_S(\theta)$ = skeptical prior, $\pi_E(\theta)$ = enthusiastic prior, Eff = probability of efficacy criteria satisfied at interim analysis, Fut = probability of futility criteria satsified at interim analysis, E(SS) = expected sample size at interim stoppage, Conc = concentrated prior specification, Def = default prior specification, Flat = flattened prior specification.

		$\theta = 0.4$			$\theta = 0.535$			$\theta = 0.67$		
$\pi_S(\theta)$	$\pi_E(\theta)$	Eff	Fut	E(SS)	Eff	Fut	E(SS)	Eff	Fut	E(SS)
Conc	Def	0.026	0.974	20.9	0.438	0.562	30.9	0.953	0.047	23.7
Conc	Flat	0.025	0.975	18.4	0.397	0.603	27.7	0.930	0.070	23.0
Def	Flat	0.033	0.967	18.3	0.428	0.572	26.3	0.934	0.066	21.3
Def	Def	0.037	0.963	20.7	0.478	0.522	29.2	0.957	0.043	21.9

Table 2: Summary characteristics of re-analysis of PLUTO trial. I/F = Interim/Final, $\psi^{(E)}(\mathbf{D})$ = Box's p-value using enthusiastic prior, ω = Enthusiastic mixing weight in adaptive monitoring prior, Efficacy Post Prob = Posterior probability of treatment efficacy.

δ	Sample Size (I/F)	$\psi^{(E)}(\mathbf{D})$ (I/F)	ω (I/F)	Efficacy Post Prob (I/F)
0.00	62 / 90	0.914 / 0.965	0.914 / 0.965	0.980 / 0.979
0.05	64 / 92	0.876 / 0.934	0.833 / 0.887	0.976 / 0.962
0.10	76 / 92	0.941 / 0.934	0.847 / 0.841	0.975 / 0.951
0.15	92 / 92	0.934 / 0.934	0.794 / 0.794	0.940 / 0.940
0.20	92 / 92	0.934 / 0.934	0.747 / 0.747	0.928 / 0.928
0.25	92 / 92	0.934 / 0.934	0.701 / 0.701	0.917 / 0.917

Figures

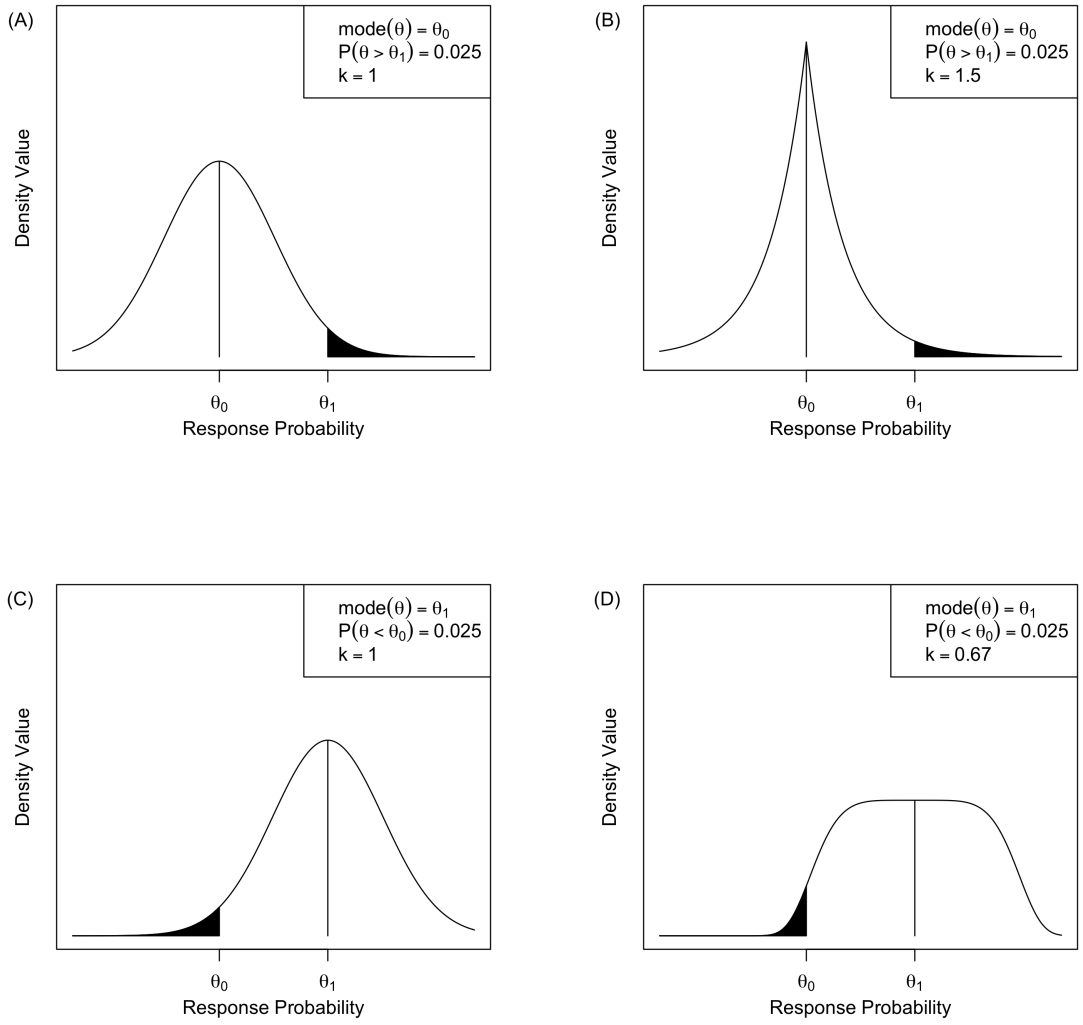


Figure 1: A, Default skeptical prior. B, Concentrated skeptical prior. C, Default enthusiastic prior. D, Flattened enthusiastic prior.

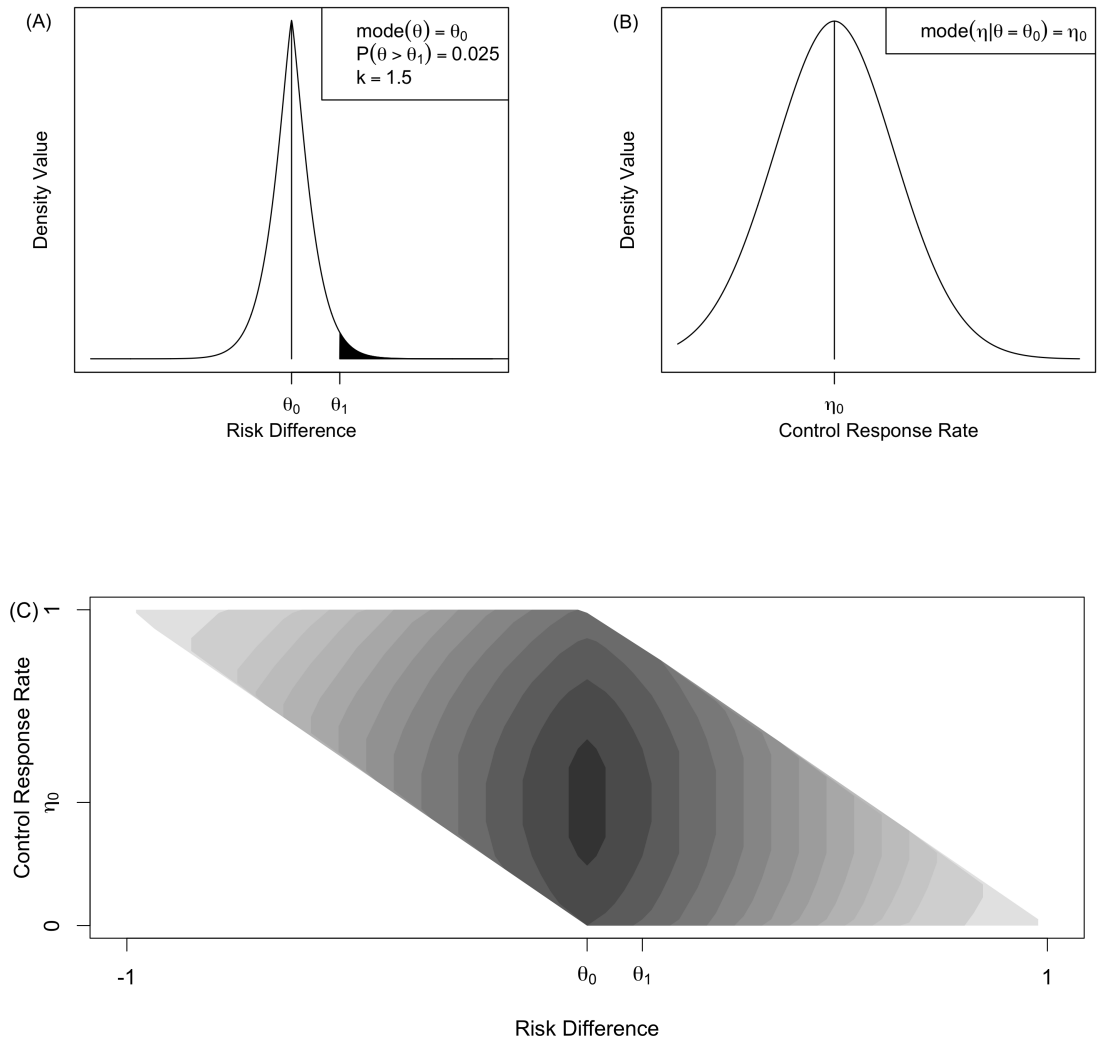


Figure 2: A, Concentrated skeptical prior $\pi_S(\theta)$ truncated to $[-1, 1]$. B, Conditional prior $\pi(\eta|\theta = \theta_0)$. C, Joint prior $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ truncated based on the conditions $-1 < \theta < 1$ and $0 < \theta + \eta < 1$.

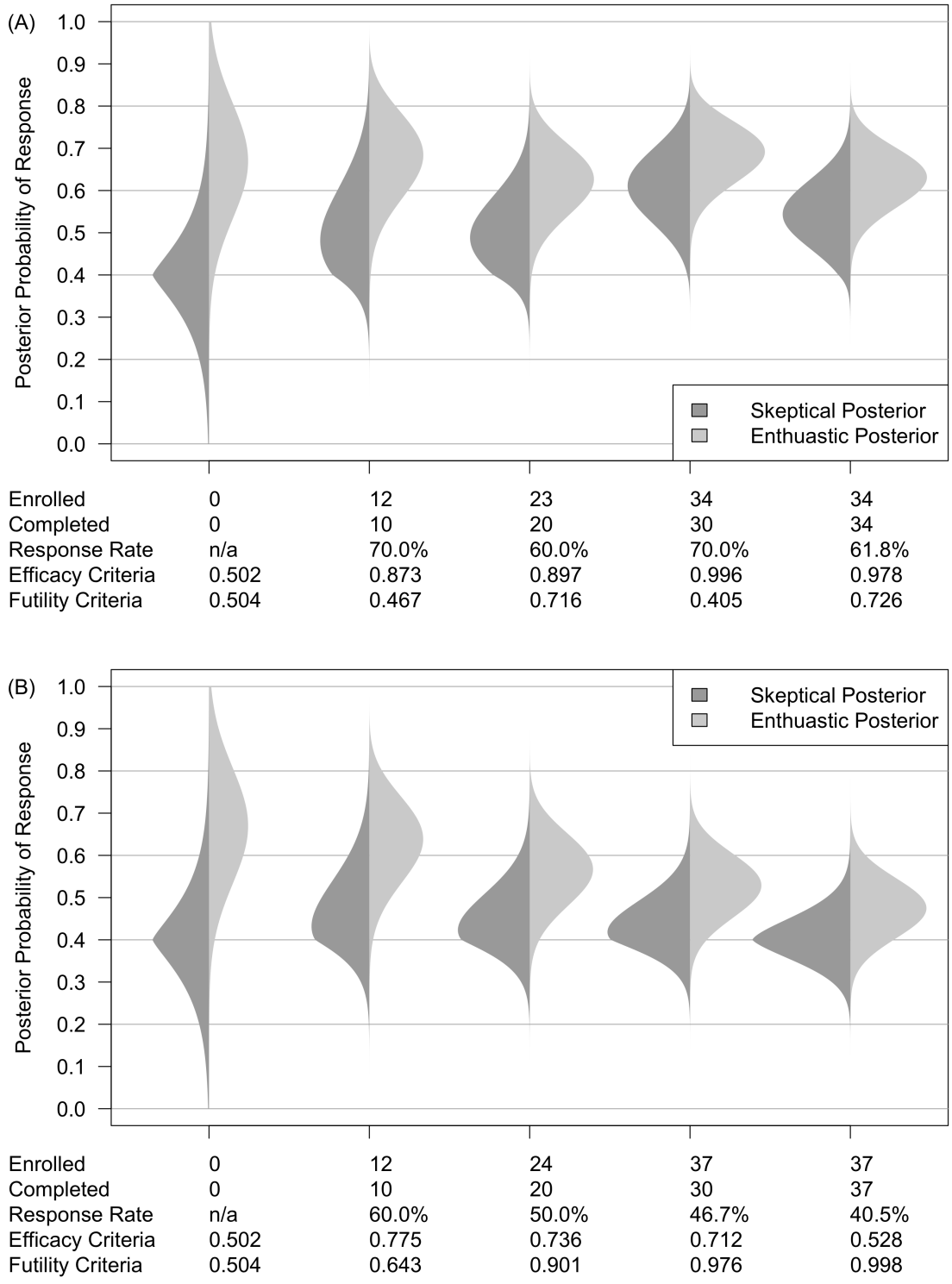


Figure 3: Example paths for the trial described in Section 3.1.2. A, Early stoppage for efficacy. B, Early stoppage for futility.

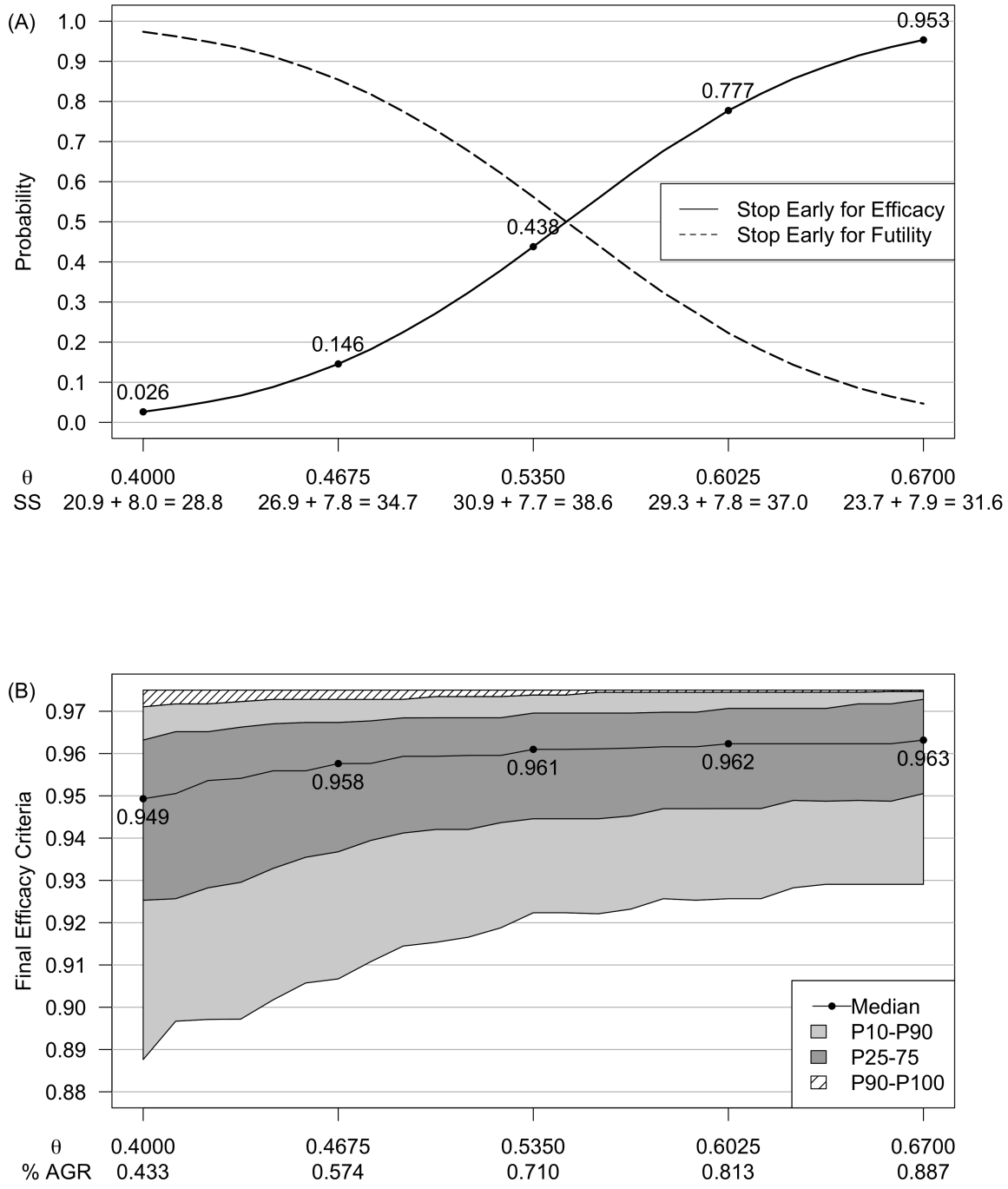
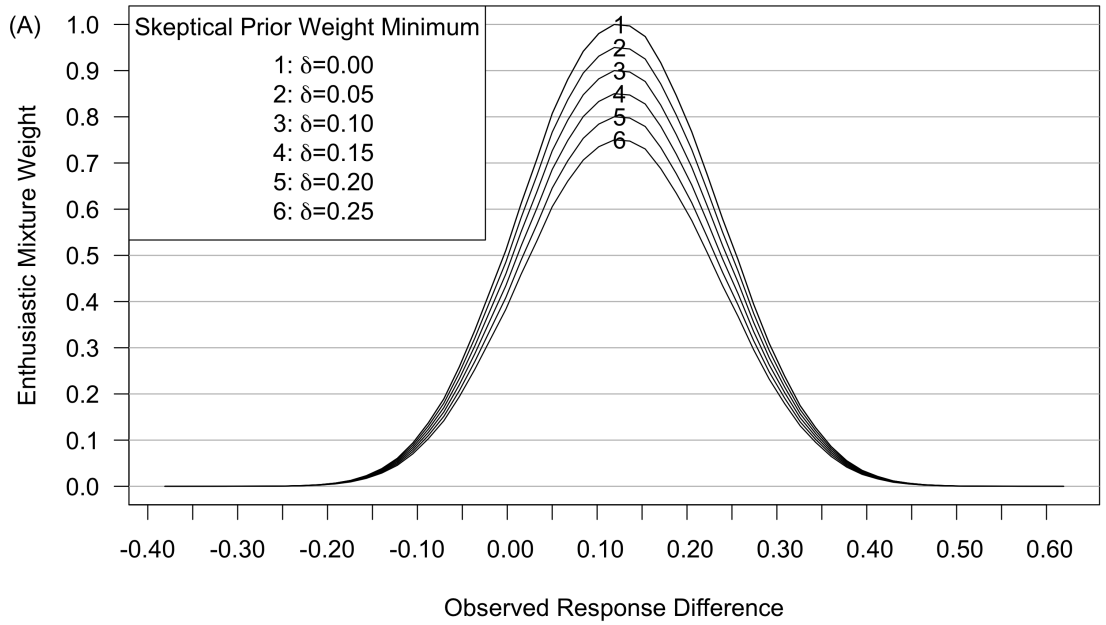
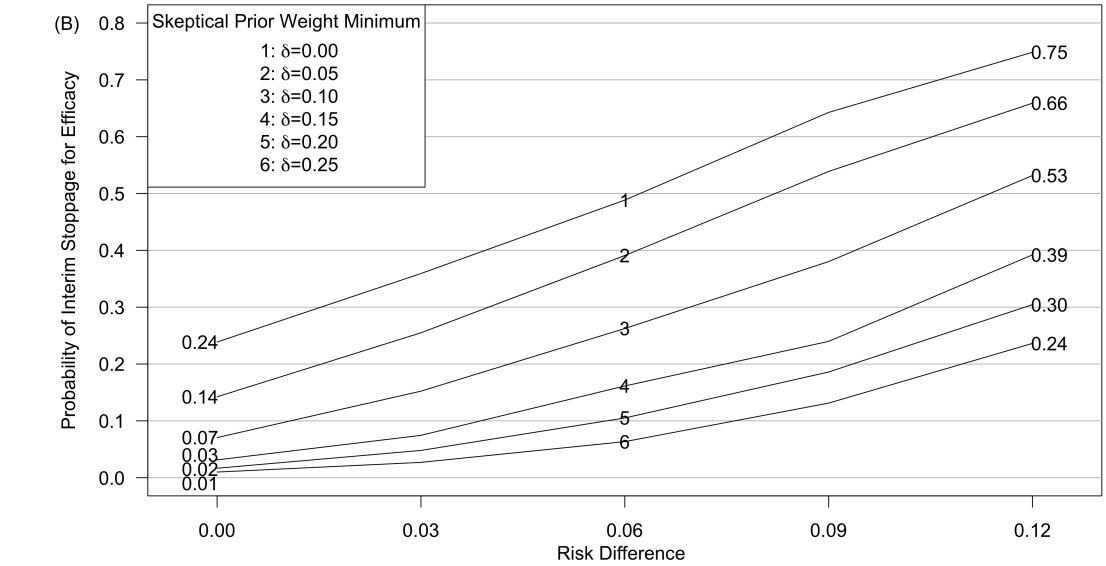


Figure 4: A, Sequential design properties. (SS; mean sample size, (I); interim analysis, (F); final analysis). B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; percent of agreement between final and interim posterior probabilities relative to $1 - \epsilon$ threshold).





Expected Sample Size (Interim/Final)					
1: δ=0.00	91.5 / 96.1	87.1 / 94.0	82.6 / 92.0	77.4 / 89.7	73.4 / 88.0
2: δ=0.05	95.5 / 98.3	92.6 / 97.3	89.3 / 96.1	84.5 / 94.3	80.2 / 92.8
3: δ=0.10	98.7 / 99.8	97.6 / 99.8	95.5 / 99.5	93.2 / 99.2	90.1 / 98.7
4: δ=0.15	99.6 / 100.0	99.2 / 100.0	98.2 / 100.0	97.1 / 99.9	94.9 / 99.9
5: δ=0.20	99.7 / 100.0	99.6 / 100.0	99.0 / 100.0	98.4 / 100.0	96.7 / 99.9
6: δ=0.25	99.8 / 100.0	99.6 / 100.0	99.5 / 100.0	99.0 / 100.0	97.9 / 100.0

Figure 5: A, Enthusiastic prior mixing weight ω associated with skeptical prior weight minimum δ in (7) by observed response difference between IP and PC groups, when the PC response rate is fixed at 38% (16/42 responses). B, Operating characteristics for designs having with skeptical prior weight minimum δ in (7) by true risk difference when the PC response rate generated at 39%.

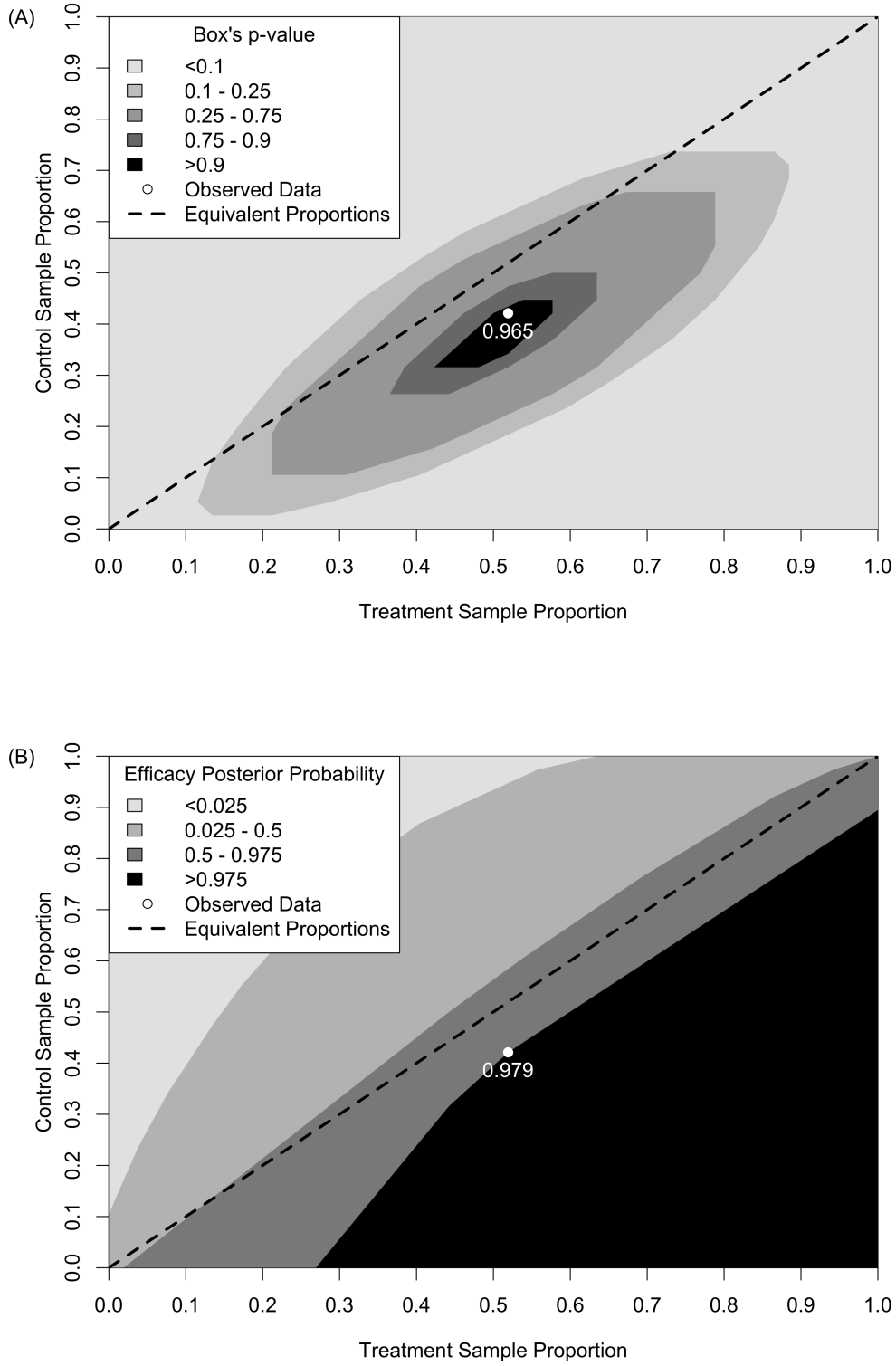


Figure 6: A, Box's p-value by control and treatment sample proportions at the final analysis with 90 subjects when $\delta = 0$ is used (7) for the adaptive monitoring prior. B, Posterior probability of efficacy by control and treatment sample proportions.

Figure Captions

Figure 1: A, Default skeptical prior. B, Concentrated skeptical prior. C, Default enthusiastic prior. D, Flattened enthusiastic prior.

Figure 2: A, Concentrated skeptical prior $\pi_S(\theta)$ truncated to $[-1, 1]$. B, Conditional prior $\pi(\eta|\theta = \theta_0)$. C, Joint prior $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ truncated based on the conditions $-1 < \theta < 1$ and $0 < \theta + \eta < 1$.

Figure 3: Example paths for the trial described in Section 3.1.2. A, Early stoppage for efficacy. B, Early stoppage for futility.

Figure 4: A, Sequential design properties. (SS; mean sample size, (I); interim analysis, (F); final analysis). B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; percent of agreement between final and interim posterior probabilities relative to $1 - \epsilon$ threshold).

Figure 5: A, Enthusiastic prior mixing weight ω associated with skeptical prior weight minimum δ in (7) by observed response difference between IP and PC groups, when the PC response rate is fixed at 38% (16/42 responses). B, Operating characteristics for designs with skeptical prior weight minimum δ in (7) by true risk difference when the PC response rate generated at 39%.

Figure 6: A, Box's p-value by control and treatment sample proportions at the final analysis with 90 subjects when $\delta = 0$ is used (7) for the adaptive monitoring prior. B, Posterior probability of efficacy by control and treatment sample proportions.