

RESEARCH ARTICLE

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

Evan Kwiatkowski*¹ | Eugenio Andraca-Carrera² | Mat Soukup² | Matthew A. Psioda¹

¹Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina, USA

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

Correspondence

*Evan Kwiatkowski, Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB#7420, Chapel Hill, North Carolina, USA. Email: ekwiatkowski@unc.edu

Summary

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 through mixing an a priori skeptical prior with an enthusiastic prior using a weight that can be fixed or adaptively estimated based on the degree to which observed data are better supported by a skeptical versus an enthusiastic perspective. **In particular, we introduce the adaptive skeptical monitoring prior which dynamically weights skeptical and enthusiastic components in a mixture prior for data monitoring, where the priors used in the mixture are constructed using summaries from external data. The adaptive skeptical monitoring prior allows for prospective and pre-specified use of external data into the monitoring procedure.** The proposed method is applied in two examples: (1) a single-arm proof-of-concept trial and (2) a two-arm randomized controlled trial **including a retrospective analysis of the real trial data under 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 of each trial design is performed to illustrate that the proposed Bayesian approaches provide reasonable frequentist operating characteristics without having that explicit focus.

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 trials¹. 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 to account for interim analyses are 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 product². 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 penalty³.

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, etc.). 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 pediatric disease. 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 Complex Innovative Designs (CIDS) - a performance goal set forth under PDUFA VI legislation⁴. Moreover, in settings where pertinent preexisting data are available, Bayesian methods provide a natural approach for incorporating that information via a prior distribution into the design and analysis of a future trial. The potential use of Bayesian methods is discussed throughout the draft guidance⁵, created in response to the 21st Century Cures Act⁶, including the use of Bayesian methods to extrapolate information from adult patients to pediatric settings as one type of possible CID (see Section III, Part C).

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, when available.

A key contribution of this work is to give structured definitions for 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 (Section 2.2.2). A conditional-marginal prior factorization is proposed for settings where there are one or more nuisance parameters (Section 2.2.5) and we illustrate how prior information can be used in both marginal distribution for the treatment effect and conditional distribution for the nuisance parameters, if desired.

The structured definitions of skeptical and enthusiastic perspectives form the basis for adaptive mixture priors to be used prospectively. The prospective use of external data in a pre-specified design provides novelty beyond routine sensitivity analyses with different priors, and our structured framework makes these ideas more rigorous. The prospective incorporation of external data is not routinely used at the Center for Drug Evaluation and Research at the FDA, and this framework could lead to increased use of these methods.

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, the expected sample size and trial duration, the average posterior mean, and the coverage probabilities for 95% credible intervals for the treatment effect. 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^{7,8}. Prior work has shown such restrictions can result in Bayesian and frequentist designs that have stopping rules which are nearly identical^{9,10,11}. 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 analysis¹².

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 using the generalized normal distribution and for incorporating prior information into the monitoring priors, as well as construction of inference 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.

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 (e.g., placebo).

Throughout the paper we will let 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, $D = \{y_i, z_i : i = 1, \dots, n\}$ 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 $p(D|\theta, \eta)$ to reflect the density or mass function for the collective data 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

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

where $p(D|\theta)$ is commonly referred to as the likelihood for θ and $\pi(\theta)$ is its prior distribution. We will also refer to $P(\theta \in \Theta_i | D)$ as the posterior probability of hypothesis H_i . See Web Appendix A for a brief discussion of the appropriateness of referring to $P(\theta \in \Theta_i | D)$ 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 $P(\theta > \theta_0 | D)$ 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

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

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 $H_1 : \theta > \theta_0$ 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) $\arg\max_{\theta} \pi_E(\theta) = \theta_1$ 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) $\arg\max_{\theta} \pi_S(\theta) = \theta_0$ 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 $p(H_0) = p(H_1)$ (see Web 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 herein for sequential monitoring are appropriate regardless of how frequently data can be monitored.

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 D when $P_S(\theta > \theta_0 | D) > 1 - \epsilon$. 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) $\arg\max_{\theta} \pi_E(\theta) = \theta_1$ 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) $\arg\max_{\theta} \pi_S(\theta) = \theta_0$ 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)¹³ 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 (hypothetical) 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 parameter¹⁴. 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. Web 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 enthusiastic 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 C 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. Although we have focused on an enthusiastic prior in our discussion here, the same ideas apply to skeptical priors as well. 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 author's choices of 1.5 and 0.67 are relatively extreme perturbations from that afforded by the normal

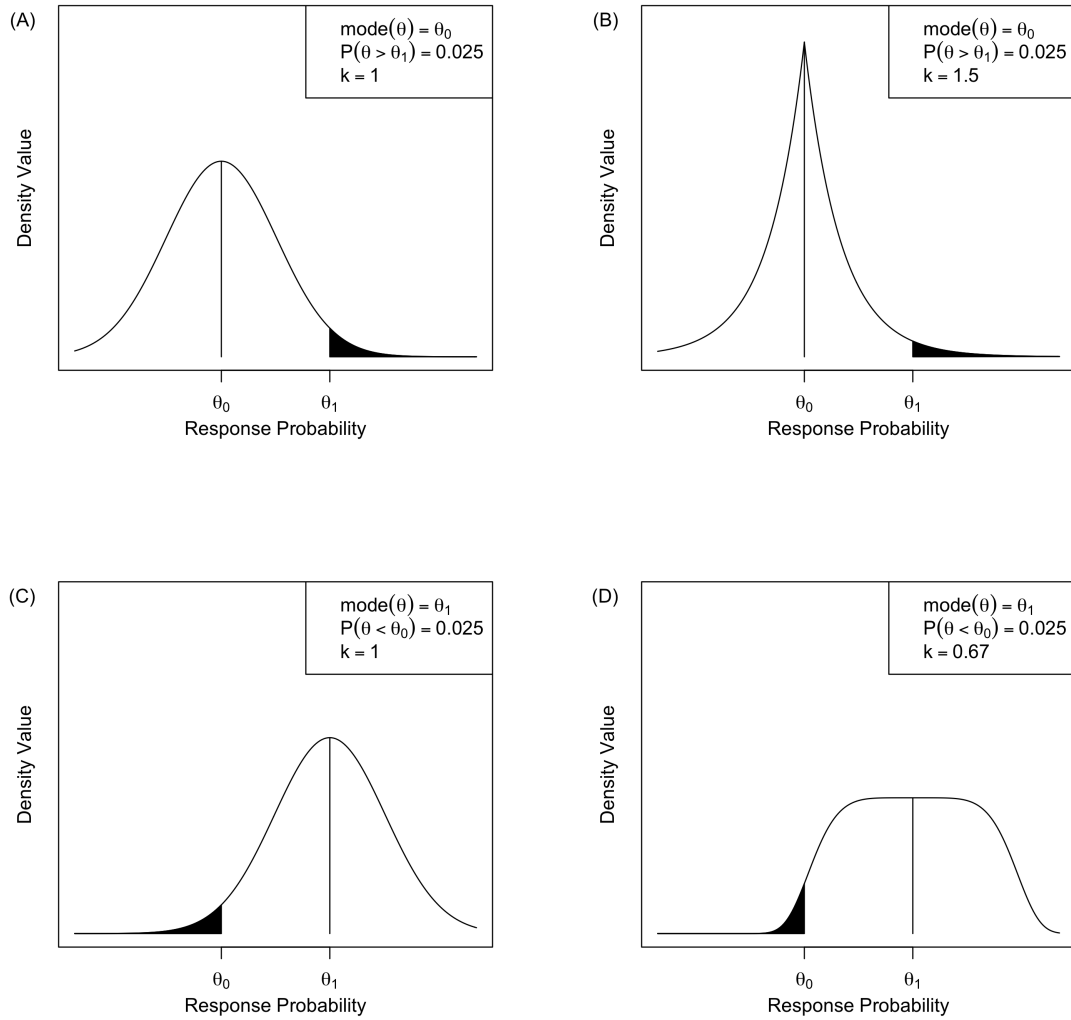


FIGURE 1 A, Default skeptical prior. B, Concentrated skeptical prior. C, Default enthusiastic prior. D, Flattened enthusiastic prior.

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 measures of compatibility between the observed data and the skeptical and enthusiastic priors. We define the *adaptive skeptical monitoring prior* as the mixture distribution

$$\pi_{AS}(\theta) \pi(\theta) = \omega_S \cdot \pi_S(\theta) + \omega_E \cdot \pi_E(\theta), \quad (3)$$

where $\omega \in [0, 1]$ is an adaptively determined mixing weight. Fixed choices of ω will be used as comparisons for the performance of the adaptive weight monitoring prior in Section 3.2.

The adaptive mixing weight ω is determined by an assessment of prior-data conflict, proposed by¹⁵, derived using the prior predictive distribution of the data which is defined (in our case) using the skeptical and enthusiastic monitoring priors. The prior-predictive distribution for data \mathbf{D} (also called the marginal likelihood) reflects the probability of observing \mathbf{D} given the assumed prior distribution for θ and is defined formally as

$$p(\mathbf{D}) = \int p(\mathbf{D}|\theta)\pi(\theta)d\theta. \quad (4)$$

Let \mathbf{D}_{obs} be the observed data at some point in time in an ongoing trial. *Box's p-value* is defined as the following:

$$\psi(\mathbf{D}_{\text{obs}}) = \int p(\mathbf{D})1[p(\mathbf{D}) \leq p(\mathbf{D}_{\text{obs}})]d(\mathbf{D}) \quad (5)$$

where $1[A]$ is an indicator that the event A is true.

Box's p-value can be interpreted as the probability of observing data as or more extreme than \mathbf{D}_{obs} , given the predictive distribution. Small values of $\psi(\mathbf{D}_{\text{obs}})$ indicate a lack of compatibility or congruency between the prior and the data. We propose using the skeptical and enthusiastic priors $\pi_S(\theta)$ and $\pi_E(\theta)$ to compute the quantities in (4) and (5) to create compatibility measurements $\psi^{(S)}(\mathbf{D}_{\text{obs}})$ and $\psi^{(E)}(\mathbf{D}_{\text{obs}})$ which are used to determine the mixing weight in (3).

Consider the mixing weight ω_E given to the *enthusiastic* prior:

$$\omega_E = (1 - \delta)f_\beta(\psi^{(E)}(\mathbf{D}_{\text{obs}})) \quad (6)$$

The minimum possible mixing weight assigned to the *skeptical* prior is achieved when $\psi^{(E)}(\mathbf{D}_{\text{obs}}) = 1$ and is equal to δ .

f_β indicates an upwards scaling of the value 0.5 such that $f_\beta(0.5) = \beta$ through a $\text{beta}(1, s_\beta)$ transformation. For $\beta = \{0.5, 0.6, 0.7, 0.8, 0.9\}$, $s_\beta = \{1, 1.32, 1.74, 2.32, 3.32\}$

2.2.4 | Inference Priors

The purpose of the inference prior is to synthesize posterior inferences from the disparate skeptical and enthusiastic perspectives to facilitate interpretation of the data once it has been obtained. The skeptical and enthusiastic monitoring priors defined in Section 2.1.3 represent extreme but plausible beliefs about θ . While analysis with these priors provides a rational perspective from which one can determine whether interim data are sufficient to cease enrolling patients, the belief of most stakeholders will likely fall somewhere between the two perspectives. Thus, when interpreting the final data once in hand, intermediate perspectives should be considered. To that end, we define an inference prior through mixing the two monitoring priors. One may take a fixed value of $\omega_S = \omega_E = 1/2$ in (3) to obtain an *agnostic* inference prior since it gives equal weight to the skeptical and enthusiastic components (see Section 3.1).

2.2.5 | 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. $\text{mode}(\eta|\theta = \theta_0) = \eta_0$), 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.

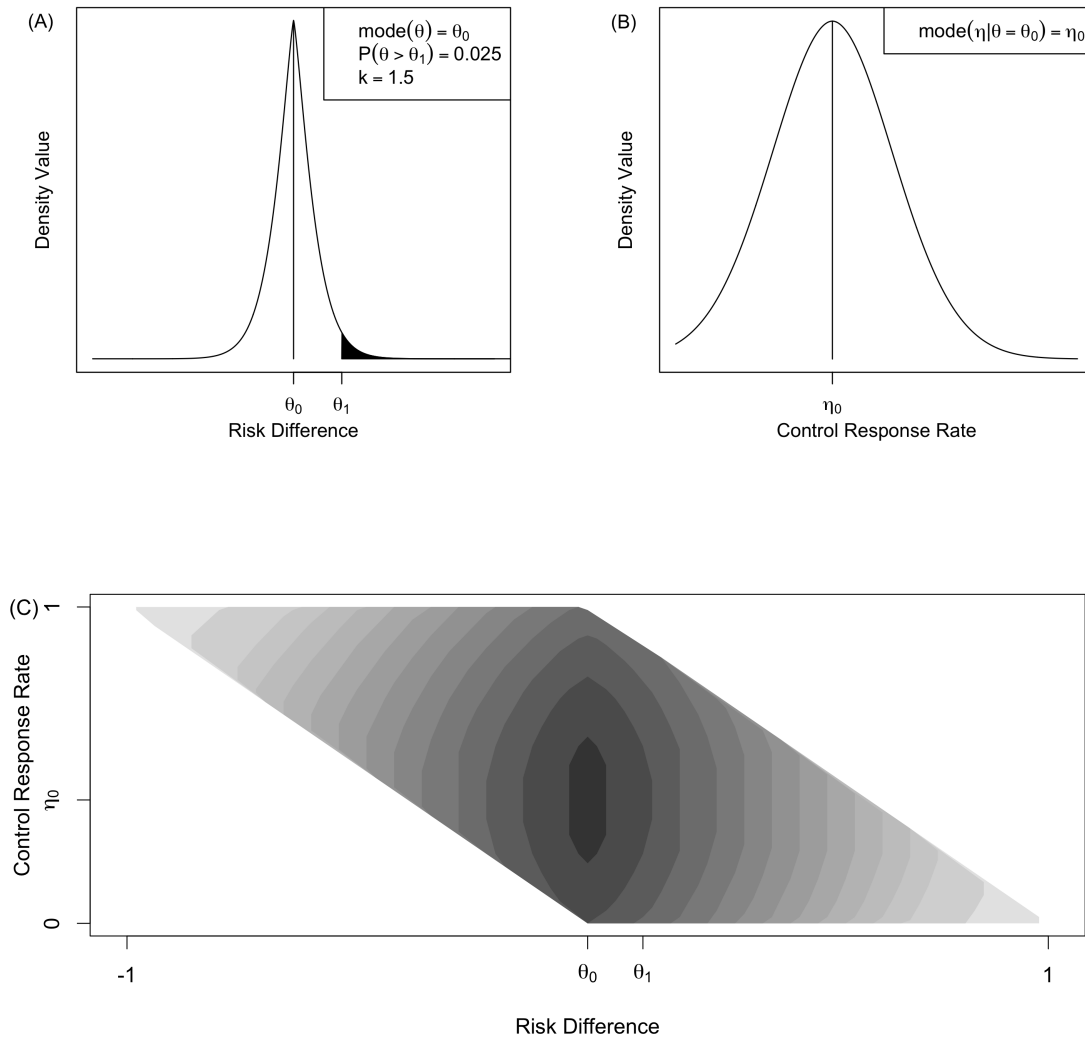


FIGURE 2 A, Concentrated skeptical prior $\pi_S(\theta)$ truncated to $[-1, 1]$. B, Conditional prior $\pi(\eta|\theta = \theta_0)$. C, Joint prior $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ truncated based on the conditions $-1 < \theta < 1$ and $0 < \theta + \eta < 1$.

3 | EXAMPLES

3.1 | Single-Arm Proof-of-Activity 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)¹⁶ 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

infiximab 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 trials¹⁷ 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 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. A comparison of design properties based on various combinations of the monitoring priors is given in Web Appendix D. In this example, we considered an inference prior defined as the mixture (??) using the values $\omega_S = \omega_E = 0.5$ (i.e., a non-adaptive inference prior), which was used to determine the posterior mean and 95% credible intervals. 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 R¹⁸ using the `pracma` package¹⁹.

3.1.3 | Example Paths

Figure 3 presents violin plots to illustrate the monitoring process for two hypothetical instances of the trial. For each instance, the monitoring priors (left-most set of distributions), and the posterior distributions at three selected interim analyses (middle sets) and from the final analysis (right-most set) are shown. 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 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. The final data matches the last interim analysis since enrolled patients would be transitioned off the IP.

3.1.4 | Preposterior Analysis of Operating Characteristics

The operating characteristics presented in this section are estimated using 100,000 simulated trials per value of θ using the trial design as described in Section 3.1.2. 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 . At $\theta_0 = 0.4$, the posterior mean estimates using the mixture prior are slightly greater than the true underlying value, and at $\theta_1 = 0.67$, the posterior estimates using the mixture prior are slightly less than the true underlying value. This slight bias towards the interval $[\theta_0, \theta_1]$ is because those are the values determined to be the most likely a priori by the mixture prior. The 95% credible intervals are shown to have coverage probabilities exceeding their nominal level for all values of θ considered.

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. These analyses show that even if the skeptical prior was used for the final determination of efficacy rather than an inference prior, the final posterior probability of the alternative would be similar to that which triggered stoppage of enrollment, and the level of similarity increases as the underlying response probability θ increases.

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 2018²⁰. The study population was comprised of patients ages 5 through 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 constraints 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 response-adaptive sequential monitoring from Section 2.2.3. Response-adaptive sequential monitoring 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 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 probability of concluding efficacy at an interim analysis is made using a mixture prior with dynamic weight of the form (3) 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} .

3.3 | Mixture Inference Prior Weights

Figure 5(A) shows the prior-data compatibility assessments $\psi^{(S)}$ and $\psi^{(E)}$ by observed risk difference. As expected, the skeptical and enthusiastic priors show highest compatibility when the observed risk difference matches the corresponding prior mode, and the non-informative prior shows high compatibility for a wide range of θ . Figure 5(B) shows the mixture inference prior weights ω_E by observed risk difference and choice of δ , β . Recall our goal is to create a mixture prior which favors the enthusiastic components in areas where high compatibility is demonstrated for this components, and favors the skeptical prior if the enthusiastic component shows low compatibility.

3.3.1 | Preposterior Analysis of Operating Characteristics

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 6 shows the probability of stopping early for efficacy using a fixed weight mixture prior of the form (3) for efficacy monitoring with a fixed choice of ω chosen at the outset to be in the set $\{0, 0.5, 1\}$ and the associated sample sizes, along with adaptive weight mixtures with different choices of δ and β .

4 | REAL DATA EXAMPLE

We consider applying the adaptive mixture prior of (3) to the real outcomes of “The Pediatric Lupus Trial of Belimumab Plus Background Standard Therapy (PLUTO)” (NCT01649765)²⁰. The skeptical and enthusiastic priors are defined in the same way as Section 3.2.2, which are centered around response differences of 0 and 0.12 respectively. There were responses available from 92 patients that were used for this analysis (one patient in the placebo group dropped out before any outcome was observed). Sequential monitoring after every two completed outcomes was conducted after a minimum sample size of 50 had been reached.

TABLE 1 Interim sample size at stoppage for efficacy by choice of minimum skeptical mixture weight δ and upwards scaling of Bayesian p -value of 0.5 (β) using concentrated skeptical prior and default enthusiastic prior.

δ / β	0.5	0.6	0.7	0.8	0.9
0.00	62	54	54	50	50
0.05	62	62	54	54	54
0.10	72	62	62	54	54
0.15	92	72	64	64	64
0.20	92	92	78	78	78
0.25	92	92	92	92	92

When δ is equal to zero, that is, the weight of the enthusiastic component is allowed to approach 1 in (3), then the efficacy criteria is satisfied at the first monitoring opportunity with 52 completed outcomes. At this point the data were shown to be more compatible with the enthusiastic prior than the skeptical prior, so the skeptical weight was reduced the $\delta = 0$ in the mixture. As the value of δ increases, so does the interim sample size where the efficacy criteria is achieved.

5 | DISCUSSION

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. Even though frequentist operating characteristics are not an explicit focus of the design, we demonstrate that the Bayesian approaches proposed provide good operating characteristics across a wide range of data monitoring frequencies and enrollment patterns.

Our results in Section 3.2 can be compared to a post-hoc Bayesian hierarchical analysis which used data from two studies of the use of belimumab in adults²⁰. 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. Our analyses show that for such a trial to have any chance of early stopping, it is necessary to borrow information for the skeptical monitoring prior as frequent data monitoring with a default skeptical prior has limited potential to conclude efficacy (see Figure 6). In fact, the adaptive weight skeptical prior used in this example may be too conservative for this setting, and adaptive methods that include more liberal information borrowing could be used.

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.

References

1. U.S. Food and Drug Administration . Establishment and Operation of Clinical Trial Data Monitoring Committees. 2006.
2. Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC . 2000.
3. Spiegelhalter DJ, Freedman LS, Parmar MKB. Applying Bayesian ideas in drug development and clinical trials. *Statistics in Medicine* 1993; 12(15-16): 1501–1511. doi: 10.1002/sim.4780121516
4. U.S. Food and Drug Administration . PDUFA Reauthorization Performance Goals and Procedures Fiscal Years 2018 through 2022. 2017.
5. U.S. Food and Drug Administration . Interacting with the FDA on Complex Innovative Trial Designs for Drugs and Biological Products. 2019.
6. U.S. Congress . 21st Century Cures Act (Pubic Law 114-255, 130 STAT 1033-1344). 2016.
7. Ventz S, Trippa L. Bayesian designs and the control of frequentist characteristics: A practical solution. *Biometrics* 2015; 71(1): 218–226. doi: 10.1111/biom.12226
8. Zhu H, Yu Q. A Bayesian sequential design using alpha spending function to control type I error. *Statistical Methods in Medical Research* 2015; 26(5): 2184–2196. doi: 10.1177/0962280215595058
9. Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. *BMC Medical Research Methodology* 2020; 20(1): 4. doi: 10.1186/s12874-019-0892-8

10. Kopp-Schneider A, Calderazzo S, Wiesenfarth M. Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. *Biometrical Journal* 2020; 62(2): 361–374. doi: 10.1002/bimj.201800395
11. Zhu L, Yu Q, Mercante DE. A Bayesian Sequential Design for Clinical Trials with Time-to-Event Outcomes. *Statistics in biopharmaceutical research* 2019; 11(4): 387–397. doi: 10.1080/19466315.2019.1629996
12. Psioda MA, Ibrahim JG. Bayesian design of a survival trial with a cured fraction using historical data. *Statistics in Medicine* 2018; 37(26): 3814–3831. doi: 10.1002/sim.7846
13. Le Cam L, Yang GL. *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer . 2000.
14. Nadarajah S. A generalized normal distribution. *Journal of Applied Statistics* 2005; 32(7): 685–694. doi: 10.1080/02664760500079464
15. Box GEP. Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)* 1980; 143(4): 383–430. doi: 10.2307/2982063
16. Hyams J, Damaraju L, Blank M, et al. Induction and Maintenance Therapy With Infliximab for Children With Moderate to Severe Ulcerative Colitis. *Clinical Gastroenterology and Hepatology* 2012; 10(4): 391 – 399.e1. doi: <https://doi.org/10.1016/j.cgh.2011.11.026>
17. Rutgeerts P, Sandborn WJ, Feagan BG, et al. Infliximab for Induction and Maintenance Therapy for Ulcerative Colitis. *New England Journal of Medicine* 2005; 353(23): 2462–2476. doi: 10.1056/NEJMoa050516
18. R Core Team . R: A Language and Environment for Statistical Computing. 2017.
19. Borchers HW. *pracma: Practical Numerical Math Functions*. 2019.
20. Brunner HI, Abud-Mendoza C, Viola DO, et al. Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial. *Annals of the Rheumatic Diseases* 2020; 79(10): 1340 LP – 1348. doi: 10.1136/annrheumdis-2020-217101
21. U.S. Food and Drug Administration . Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials. 2018.

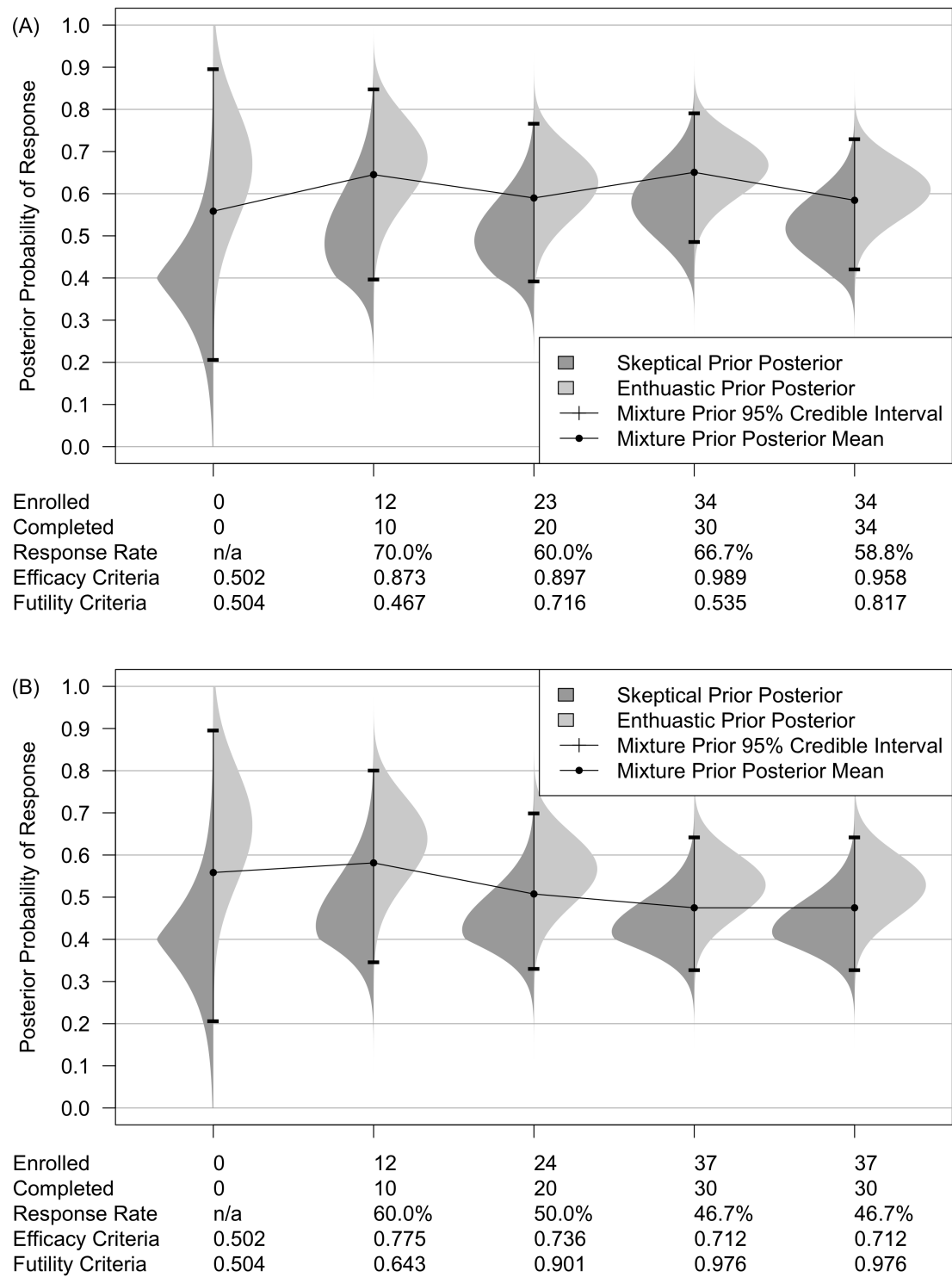


FIGURE 3 Example paths for the trial described in Section 3.1.2. A, Early stoppage for efficacy. B, Early stoppage for futility.

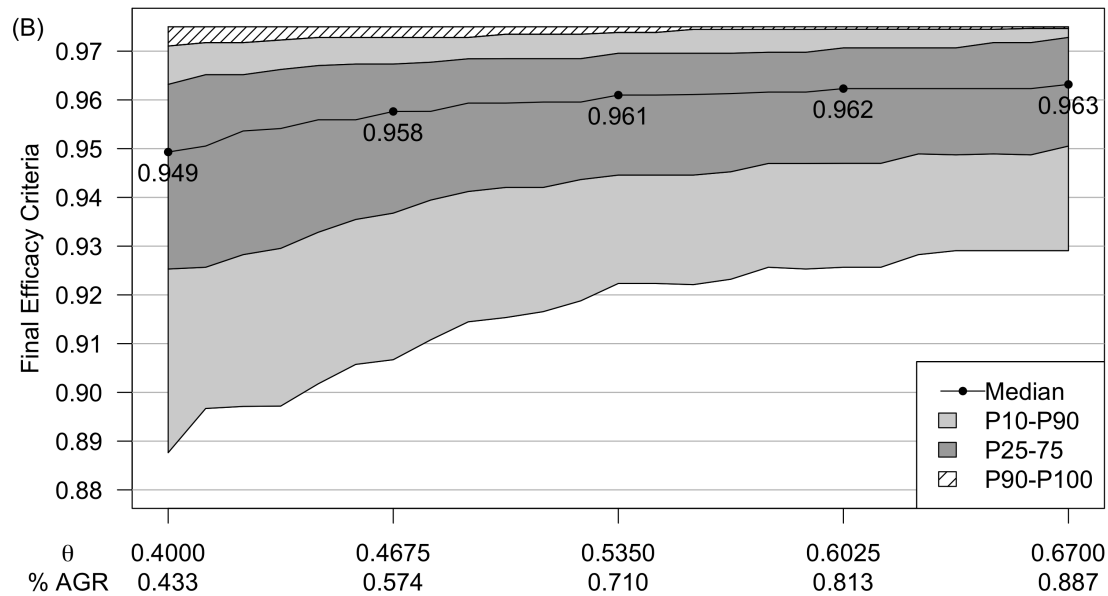
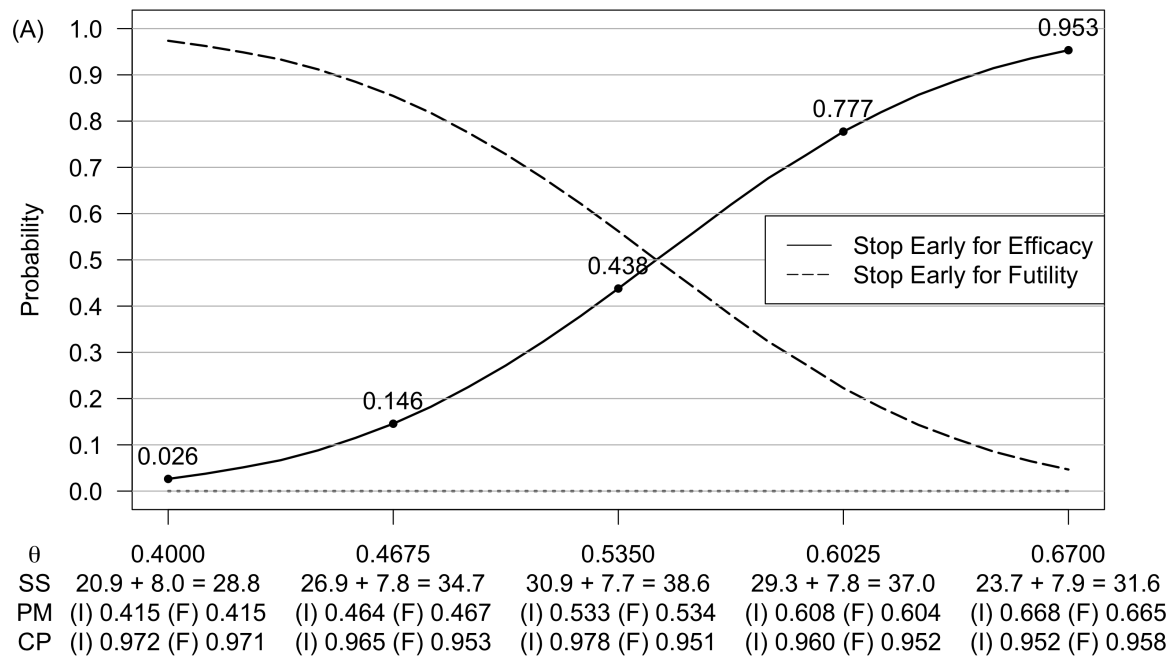


FIGURE 4 A, Sequential design properties. (SS; mean sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis). B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; percent of agreement between final and interim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

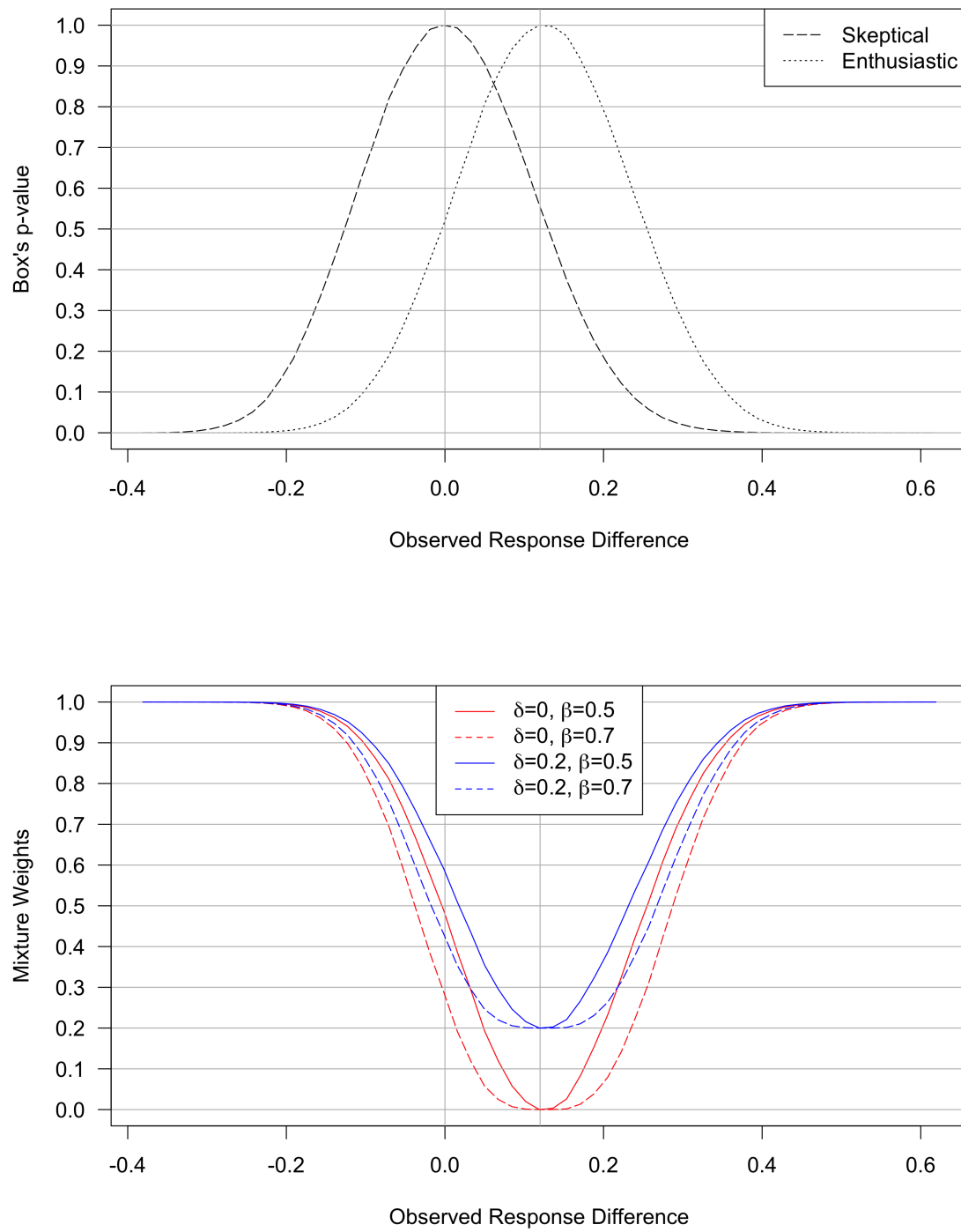


FIGURE 5 A; Prior-data compatibility assessments $\psi^{(S)}, \psi^{(E)}$ by observed risk difference. B; Mixture inference prior weights ω_E by observed risk difference and choice of δ, β .

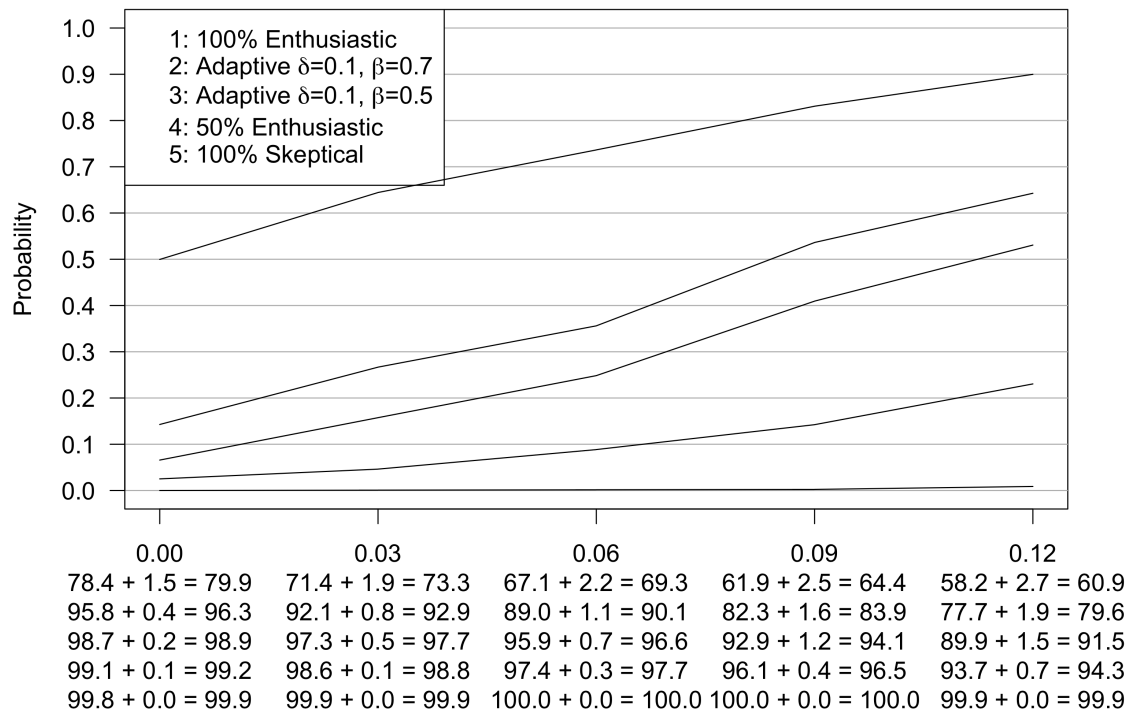


FIGURE 6 Probability of stopping for efficacy and associated sample sizes by true IP response probability θ with different choices of efficacy monitoring prior.

