A Structured Framework for Bayesian Sequential Monitoring in Clinical Trials

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

[1]Department of Biostatistics, University of North Carolina,

McGavran-Greenberg Hall, CB#7420,

Chapel Hill, North Carolina, USA

[2]Division of Biometrics VII, Office of Biostatistics,

Center for Drug Evaluation and Research, US Food and Drug Administration,

Silver Spring, Maryland, USA

*email: ekwiatkowski@unc.edu

SUMMARY:

The Bayesian paradigm is naturally suited for sequentially monitoring clinical trials and incorporating external data. We present a framework for Bayesian sequential monitoring that makes structured use of external data that can be consistently applied for a wide range of clinical trial applications. The basis for this framework is the requirement that parameterization of monitoring priors and specification of stopping criteria are algorithmic byproducts of the trial hypotheses and relevant external data summaries. The monitoring priors are defined using the family of generalized normal distributions which provide a unified approach for prior specification and interpretation for a wide array of data types. Mixtures of these monitoring priors using fixed or adaptive weights are explored for controlling the influence of the external data. The standardization of these fundamental concepts relating to the trial design provides for coherent inference on trial hypotheses. The proposed method is applied in two scenarios (1) a simple single-arm proof-of-activity trial and (2) a two-group randomized controlled trial. Both scenarios are modeled after completed trials in pediatrics and the designs incorporate information from adult trials to varying degrees. The frequentist properties of the trial designs using this framework are explored to illustrate that the highly interpretable Bayesian approaches proposed provide good frequentist operating characteristics without explicit focus on such criteria.

KEY WORDS:

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

In the clinical development of a therapeutic product, it is often a challenge to obtain sufficient outcomes for the primary objective (e.g. difficulty in enrolling, limited patient populations). Bayesian sequential methods are ideal to address this uncertainty in outcome ascertainment by adhering to the likelihood principle, which includes no intrinsic restrictions for frequent or even continual data monitoring. The advantages of sequential monitoring are especially apparent when there are additional difficulties in obtaining outcomes (e.g. long latency in observing the outcome, substantial financial cost associated with each patient). In settings where pertinent existing data are available, Bayesian designs offer a natural approach for incorporating that information into the design and analysis of future trials. An example where sequential Bayesian methods are valuable is in the area of pediatric clinical trials, where enrollment is often difficult and where relevant data from adult trials are often available. Utilizing information from the adult trials may result in pediatric trials that are more efficient (e.g., increased power, fewer subjects required, shorter trial duration).

We present a constructive method of designing Bayesian sequentially monitored clinical trials by eliciting monitoring priors and stopping criteria from the following: the null level for the treatment effect, a highly efficacious but plausible level for the treatment effect, and a threshold for determining substantial evidence. The highly efficacious but plausible level could be derived from a relevant external data summary. Skeptical and enthusiastic priors are defined with the generalized normal distribution which is applicable to a wide array of data types.

The framework includes the capacity to adjust the monitoring priors to reflect nuanced prior opinion about the treatment effect (Section 2.2.1) and to be adjusted dynamically based on observed responses (Section 2.2.2). A procedure for conditional-marginal specification is presented as the default option in multi-parameter settings (Section 2.2.4). This paper uses

posterior probabilities to highlight how the level of evidence at interim analyses is linked to monitoring prior specification.

The frequentist properties of trial designs using this framework are explored. Bayesian group sequential designs are often restricted to have explicit frequentist properties (Ventz and Trippa, 2015; Zhu and Yu, 2015). In fact, recent work has shown these restrictions can result in Bayesian and frequentist group sequential designs to have stopping rules that are nearly identical (Stallard et al., 2020; Kopp-Schneider et al., 2019; Zhu et al., 2019). While it is possible to calibrate a Bayesian design to have exact pre-specified frequentist properties, this re-introduces inflexibility (e.g. restrictions on interim analyses) that is unnecessary using a fully Bayesian method. Therefore, our objective is to present operating characteristics generally that are "well-calibrated" (Grieve, 2016) but not motivated by strict adherence to pre-specified thresholds.

This paper explores the operating characteristics of a sequentially monitored clinical trial including probability of stopping early for efficacy or futility at an interim or at the final analysis, expected sample size and trial duration, and posterior mean and coverage probability for the treatment effect. The distribution of the final posterior probability of efficacy given interim stoppage is explored in cases that final data dampens the conclusion for efficacy that was made at an interim analysis. Impact of frequency of data monitoring on determination of efficacy is shown.

This paper is organized as follows: Section 2 contains a brief review of Bayesian hypothesis testing using posterior probabilities, and describes the method for defining monitoring and inference priors. Examples are in Section 3, with Section 3.1 containing a one-arm study and Section 3.2 a two-arm study.

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 binary response endpoint, $\theta \in (0, 1)$ may be the response probability associated with patients receiving the investigational treatment. In a two-arm trial with binary response endpoint, $\theta \in (-1, 1)$ may be the difference in response probabilities between patients receiving the investigational treatment and those receiving the control treatment (e.g., placebo).

Throughout the paper we will let $\boldsymbol{D}$ 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, $\boldsymbol{D} = \{y_i, z_i : i = 1, ..., n\}$ where y_i is an indicator of response for patient i and z_i is an indicator for whether patient i was assigned the investigational treatment. We use the generic representation $p(\boldsymbol{D}|\theta, \eta)$ to reflect the density or mass function for the collective data $\boldsymbol{D}$ as a function of θ and potential nuisance parameters η which could be multi-dimensional. For example, for the two-arm trial example above η would correspond to the response probability for patients receiving the control treatment. For ease of exposition, for the remainder of Section 2.1, we will focus on the case where θ is the only unknown parameter and ignore η.

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

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

where $p(\boldsymbol{D}|\theta)$ is also referred to as the likelihood $\mathcal{L}(\theta|\mathbf{D})$. We will also refer to $P(\theta \in \Theta_i|\boldsymbol{D})$ as the posterior probability of hypothesis H_i (see (16) and (17) in Appendix Section 5.1).

2.1.2 *Formalizing the Statistical Concept of Substantial Evidence.*   For testing one-sided hypotheses such as $H_0 : \theta \leqslant \theta_0$ versus $H_1 : \theta > \theta_0$ using posterior probabilities, the null hypothesis is rejected when $P(\theta > \theta_0 | \boldsymbol{D})$ exceeds a predefined threshold. Let ϵ be the threshold for insignificant *residual uncertainty*, and define the posterior probability threshold $1 - \epsilon$ to be the threshold for what constitutes *substantial evidence* in favor of a claim (e.g., that $\theta > \theta_0$). Leveraging common practice, we will use $\epsilon = 0.025$ for the examples in this paper so that $1 - \epsilon = 0.975$ is the threshold for substantial evidence. Our purpose in this paper is not to debate the appropriateness of using 0.975 as a threshold for defining substantial evidence but rather to develop a strategy for prior elicitation that leverages an accepted threshold to make prior elicitation more structured for sequentially monitored trials in hopes that this added structure facilitates the use of sequential monitoring more broadly.

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.*   Having formalized concepts for *substantial evidence* and being *all but convinced* of a claim, we can now 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. The use of monitoring based on changing the opinion of skeptical and enthusiastic observers has been described as overcoming a handicap (Freedman and Spiegelhalter, 1989) and providing a brake (Fayers et al., 1997) on the premature termination of trials, and as constructing "an adversary who will need to be disillusioned by the data to stop further experimentation" (Spiegelhalter et al., 1994). Skeptical and enthusiastic monitoring priors represent two extreme but plausible beliefs

about the quantity of interest θ relative to the hypotheses to be tested. 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?"

Monitoring priors are used for interim analyses of the data. A promising interim analysis that provides substantial evidence of efficacy may justify ending enrollment, while enrolled patients would continue to receive the treatment for the pre-planned period of exposure. A discouraging interim analysis that provides substantial evidence of futility may justify ending enrollment, and may call for enrolled patients who are ongoing in the trial to be transitioned off the investigational treatment (i.e., termination of investigation of the treatment). For the Bayesian, the question becomes "From what prior perspective must the evidence be substantial to justify one of the two actions described above?". A key contribution of this work is to give rationale for definitions of a priori skeptical and enthusiastic perspectives that can be used for early stopping decisions in favor of efficacy and futility, respectively.

For ease of exposition, consider the hypotheses $H_0 : \theta \leqslant \theta_0$ 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 $\pi_E(\theta)$ as a prior that suggests θ_1 is the most likely value of θ (e.g. modal value) 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 the prior $\pi_S(\theta)$ satisfying

$$\mathrm{argmax}_\theta P_E(\theta) = \theta_1 \text{ and } P_E(\theta > \theta_0) = 1 - \epsilon, \tag{3}$$

where the subscript E indicates that the probability is based on $\pi_E(\theta)$.

Define a skeptical prior $\pi_S(\theta)$ as a prior that suggests θ_0 is 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

$$\mathrm{argmax}_\theta P_S(\theta) = \theta_0 \text{ and } P_S(\theta < \theta_1) = 1 - \epsilon. \tag{4}$$

In what follows we refer to (3) and (4) as *modal value and tail-probability constraints.*

Note that the development of the skeptical prior does not generally reflect skepticism that the alternative hypothesis is true. Indeed, in most cases the *induced* prior model probabilities based on (16) will be such that $p(H_0) \approx p(H_1)$. The skeptical prior simply reflects that $\theta > \theta_1$ is exceedingly unlikely and is therefore consistent with their being clinical equipoise about the two hypotheses.

Unlike the frequentist approach, the degree to which evidence in favor of a hypothesis is substantial is influenced by the prior distribution on the quantity of interest. It is natural that one would stop a trial early in favor of efficacy or futility when the evidence in favor of that claim is compelling to an a priori skeptic 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 in fact the alternative is true, then most any rational observer would also be convinced and therefore ceasing the collection of data to prove that claim would be reasonable. Similarly, if at any point data sufficiently convince an observer whose prior belief is in accordance with $\pi_E(\theta)$ that in fact the effect of interest is less than what was originally believed, then most any rational observer would also be convinced and therefore ceasing the collection of data altogether would be reasonable.

2.1.4 *Maximum Sample Size and Formal Stoppage Criteria.* In this section we formalize a 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. Resources permitting, the maximum sample size, denoted by $n_{\max}$, should be chosen so that there is a high probability that the trial generates substantial evidence 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 compelling evidence 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 analyze the data after each patient's outcome data are available 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 even if the motivation for sequential monitoring is scenarios where frequent monitoring is feasible.

Stopping criteria for efficacy is based on the skeptical viewpoint. The skeptic becomes convinced that a treatment is effective if at some point the observed data suggest there is substantial evidence that the alternative hypothesis is true.

Define the efficacy criteria for data $\boldsymbol{D}$ as

$$\text{eff}(\boldsymbol{D}) = P_S(\theta > \theta_0 | \boldsymbol{D}), \tag{5}$$

and the threshold for substantial evidence to be

$$\text{eff}(\boldsymbol{D}) > 1 - \epsilon. \tag{6}$$

Note that the evidence must exceed the threshold for what defines it being substantial. 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.

For futility monitoring, 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 $P_E(\theta \leqslant \theta_0|\mathbf{D}) > 1 - \epsilon$. However, notice that when $\theta = \theta_0$, $P_E(\theta \leqslant \theta_0|\boldsymbol{D})$ approaches 0.5 for large sample sizes. Therefore this potential futility criteria would reasonably be satisfied for values of θ near the null value. For this reason, we consider a different approach. Recalling that θ_1 represents a plausible, clinically meaningful treatment effect, we define $\theta_m = (\theta_0 + \theta_1)/2$ to be a less than desirable treatment effect such that if it is the case that $P_E(\theta < \theta_m|\boldsymbol{D}) > 1 - \epsilon$, the trial may be stopped due to having a low probability of reaching a meaningful treatment effect.

Define the futility criteria for data $\boldsymbol{D}$ as

$$\text{fut}(\boldsymbol{D}) = P_E(\theta \leqslant \theta_m|\boldsymbol{D}) \tag{7}$$

and the threshold for substantial evidence to be

$$\text{fut}(\boldsymbol{D}) > 1 - \epsilon. \tag{8}$$

2.1.5 *Example Paths.* Violin plots are used to show the results of simulated trials with the initial prior specification (first set of distributions), three interim analyses (middle sets), and a final analysis (last set), where interim analyses are conducted after every 10 completed outcomes.

Figure 1(a) shows the results of a trial where at the third interim analysis the efficacy condition is satisfied and enrollment is terminated. Note that in the final analysis the efficacy condition is no longer at the $1 - \epsilon$ threshold. A discussion of this type of "evidence decrease" is given in Section 2.3.4. Figure 1(b) shows the results of a trial where at the third interim analysis the futility condition is satisfied and enrollment is terminated.

[Figure 1 about here.]

2.2 *Monitoring Prior Parameterization*

2.2.1 *Generalized Normal Distribution.* Skeptical and enthusiastic monitoring priors defined in Section 2.1.3 have a required modal value and tail-probability constraint. There are many distributions which satisfy these conditions. The normal distribution will be used as a default choice. The specification of the mean and variance of a normal distribution completely specifies the modal value and tail-probability constraints, and is therefore sufficient for defining skeptical and enthusiastic priors (see Section 5.2.1). Using a normal distribution as the monitoring prior is also motivated by the Bayesian Central Limit Theorem which states under general conditions that 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 of the same family of distribution as a posterior which is the result of a large sample collection of data.

There is the possibility to change local behavior of the monitoring prior distribution around the modal value. The generalized normal distribution, which is an extension of the normal distribution, is able to change the local behavior around the modal value as compared to the default normally distributed monitoring prior, while still satisfying the modal value and tail probability constraints. The density for a generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ is $f(\theta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\{-(\frac{|\theta-\mu|}{\alpha})^\beta\}$ 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 modal value and changing the shape and scale parameters in conjunction can maintain the tail probability constraint while also changing the behavior around the modal value (see Appendix Section 5.2). A flattened monitoring is locally uniform around the modal value and a concentrated monitoring is peaked around the modal value. The flattened and concentrated priors have different levels of local dispersion around the modal value than what would be

expected from a large sample collection of data according to the Bayesian Central Limit Theorem.

An example of flattened and concentrated enthusiastic priors are shown in Figure 2. Choosing a flattened distribution is appropriate when θ is more likely to be in a wide range of values around θ_1, while maintaining the same residual uncertainty that $\theta < \theta_0$. Choosing a concentrated distribution is appropriate when prior belief reflects a higher degree of certainty that θ is in a narrow range around θ_1, while maintaining residual uncertainty that $\theta < \theta_0$.

[Figure 2 about here.]

The flattened and enthusiastic priors will be parameterized by inflating or deflating the probability mass in the interval between θ_0 and $\frac{\theta_0 + \theta_1}{2}$ as compared to a normal distribution. The inflation/deflation of the area in this interval is shaded gray in Figure 2. The flattened enthusiastic prior is defined as the generalized normal distribution that satisfies the modal value and tail probability constraints (3)-(4) that has 25% less mass in the interval between θ_0 and $\frac{\theta_0 + \theta_1}{2}$ as compared to a normal distribution which satisfies the same modal value and tail probability constraints. Reducing the mass in this interval forces a peak around the modal value. Similarly, the flattened enthusiastic prior has 50% more mass in the interval between θ_0 and $\frac{\theta_0 + \theta_1}{2}$, which forces a flattening around the modal value. These values of are chosen based on the intuitive graphical properties of the distribution (e.g. noticeably peaked around θ_1, relatively flat around θ_1). Likewise, a flattened skeptical prior has 25% less mass in the interval between $\frac{\theta_0 + \theta_1}{2}$ and θ_1, and a concentrated skeptical prior as 50% more mass in this same interval. These fixed inflation/deflation percentages provide for an unambiguous definition of a flattened or concentrated prior (see technical definition in Section 5.2.2). There are other inflation/deflation percentages which would achieve the same purpose, but these values will be used henceforth to demonstrate the concepts. Alternatively, the choice of shape and scale parameters could be specified to identify an additional probability constraint

exactly (e.g. $P(\theta \in (\theta_0, \frac{\theta_0+\theta_1}{2})) = 0.1$). The default, flattened, and concentrated priors can all be truncated while maintaining the mode and tail probability constraints.

A locally non-informative prior is a prior that is relatively flat over a wide range of plausible values of θ (in particular, relatively flat over the interval (θ_0, θ_1)). The locally non-informative prior can be viewed as a particular case of the enthusiastic prior. Define a locally non-informative prior $\pi_{NI}(\theta)$ as a prior that suggests $\frac{\theta_0+\theta_1}{2}$ is the most likely value of θ and that reflects the belief of an observer who is *all but convinced* that $\theta > \frac{\theta_0+\theta_1}{2} - 2(\theta_1 - \theta_0) = \frac{3\theta_0-\theta_1}{2}$ a priori. Formally, this is defined as the prior $\pi_{NI}(\theta)$ satisfying

$$argmax_\theta P_{NI}(\theta) = \frac{\theta_0 + \theta_1}{2} \text{ and } P_{NI}(\theta < \frac{3\theta_0 - \theta_1}{2}) = 1 - \epsilon. \tag{9}$$

This prior will be flattened such that there is 50% more mass in the interval $(\frac{3\theta_0-\theta_1}{2}, \theta_0)$ as compared to a normal distribution with the same modal value and tail probability constraint. The locally non-informative prior is shown in Figure 2, and the technical definition is in Section 5.2.2.

2.2.2 *Incorporating Prior Information in the Monitoring Priors.* Prior information can be incorporated into the monitoring priors resulting in a mixture distribution with mixing weight ω of the form

$$\pi(\theta) = \omega \cdot \pi_S(\theta) + (1 - \omega) \cdot \pi_E(\theta), \tag{10}$$

where $\omega \in [0, 1]$.

The mixing weight ω is determined by an assessment of prior-data conflict using the prior predictive distribution of the data (Box, 1980). The prior-predictive distribution of the data gives the probability of future observations of the data given initial assumptions about θ, and is defined as

$$p(\mathbf{D}) = \int \mathcal{L}(\theta|\mathbf{D})\pi(\theta)d\theta \tag{11}$$

Let $\mathbf{D}_{\text{obs}}$ be the observed data at some point in time. For each predictive density, we

compute the following:

$$\psi(\mathbf{D}) = \int 1[p(\mathbf{D}) \leqslant p(\mathbf{D}_{\text{obs}})]d(\mathbf{D}) \tag{12}$$

which in the case of discrete data is equal to $\psi(\mathbf{D}) = \sum_{\mathbf{D}} p(\mathbf{D})1[p(\mathbf{D}) \leqslant p(\mathbf{D}_{\text{obs}})]$. This can be interpreted as the probability of observing data as or more extreme given the predictive distribution. Small values of $\psi(\mathbf{D})$ indicate inconsistency between the prior and the data.

The skeptical and enthusiastic priors $\pi_S(\theta)$ and $\pi_E(\theta)$ will be used in (11) and (12) to create compatibility measurements $\psi^{(S)}(\mathbf{D})$ and $\psi^{(E)}(\mathbf{D})$ which will be used to determine the mixing weight in (10). If $\psi^{(E)}(\mathbf{D}) > \psi^{(S)}(\mathbf{D})$, then the data are more consistent with the enthusiastic prior, which should be given a greater weight in the mixture. Similarly, if $\psi^{(S)}(\mathbf{D}) > \psi^{(E)}(\mathbf{D})$ then the skeptical prior should be given a greater mixing weight.

A conservative choice of mixing weight which gives full weight to the skeptical prior (i.e. $\omega = 1$) whenever $\psi^{(S)}(\mathbf{D}) \geqslant \psi^{(E)}(\mathbf{D})$ is

$$\omega = \begin{cases} 1 & \text{if } \psi^{(S)}(\mathbf{D}) \geqslant \psi^{(E)}(\mathbf{D}) \\ 1 - (\psi^{(E)}(\mathbf{D}) - \psi^{(S)}(\mathbf{D})) & \text{if } \psi^{(S)}(\mathbf{D}) < \psi^{(E)}(\mathbf{D} \end{cases} \tag{13}$$

The weight given to the enthusiastic prior is $\psi^{(E)}(\mathbf{D}) - \psi^{(S)}(\mathbf{D})$ which is equal to how much more the data is compatible with the enthusiastic prior than it is the skeptical prior.

2.2.3 *Mixture Inference Prior.* The purpose of the inference prior is to synthesize the posterior inferences from the a priori diverse 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 a priori belief of most stakeholders will likely fall somewhere in 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 as a mixture prior constructed by mixing the monitoring priors.

Using the mixture prior specification of (10), a fixed value of $\omega = 1/2$ will be referred to as an agnostic inference prior since it gives equal weight to the skeptical and enthusiastic prior.

The inference prior can also be determined based on an assessment of prior-data conflict. Let $\pi_{NI}(\theta)$ denote a locally non-informative prior that is relatively flat over plausible values of θ. Consider a three-part mixture prior

$$\pi(\theta) = \omega_S \cdot \pi_S(\theta) + \omega_E \cdot \pi_E(\theta) + \omega_{NI} \cdot \pi_{NI}(\theta) \tag{14}$$

where $\omega_S + \omega_E + \omega_{NI} = 1$. The quantities $\psi^{(E)}$, $\psi^{(S)}$, and $\psi^{(NI)}$ will be derived from using the priors $\pi_S(\theta)$, $\pi_E(\theta)$, and $\pi_{NI}(\theta)$ in the expression for Box's p-value (12). Each of $\psi^{(E)}$, $\psi^{(S)}$, and $\psi^{(NI)}$ represent compatibility of the data with the prior.

Our goal is to create a mixture prior which favors the skeptical or enthusiastic components in areas where high compatibility is demonstrated for those components, and favors the locally non-informative prior if both the skeptical and enthusiastic components show low compatibility. To this end, consider transforming $\psi^{(NI)}$ to

$$\tilde{\psi}^{(NI)} = \begin{cases} \psi^{(NI)} - \max(\psi^{(S)}, \psi^{(E)}) & \text{if } \psi^{(NI)} > \max(\psi^{(S)}, \psi^{(E)}) \\ 0 & \text{if } \psi^{(NI)} \leqslant \max(\psi^{(S)}, \psi^{(E)}) \end{cases} \tag{15}$$

Then let the mixing weights $\omega^{(E)}$, $\omega^{(S)}$, and $\omega^{(NI)}$ be the normalized values of $\psi^{(E)}$, $\psi^{(S)}$, and $\tilde{\psi}^{(NI)}$ so that their sum is unity (e.g. $\omega^{(E)} = \psi^{(E)}/(\psi^{(E)} + \psi^{(S)} + \tilde{\psi}^{(NI)})$). This type of inference prior is used in Section 3.2.

2.2.4 *Marginal-Conditional Specification.* In the case of nuisance parameters, marginal-conditional factorization of the joint prior $\pi(\theta, \eta)$ will be used. Let θ be the parameter of interest and η be the nuisance parameters. Consider the following representation of the joint prior for θ and η: $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$. Then priors for both $\pi(\theta)$ and $\pi(\eta|\theta)$ will be given that satisfy the modal value and tail probability constraints (3)-(4) (see details in Section

5.2.2). For example, suppose that θ is the risk difference between response probabilities of the treatment group and the placebo group, and denote the probability of the placebo group by η. This prior specification is demonstrated in Figure 3, and Section 3.2.2 uses this representation of the joint prior.

[Figure 3 about here.]

2.3 *Operating Characteristics*

2.3.1 *Expected sample size and trial duration.*   Suppose that the trial has interim analyses based on the number of completed outcomes. Let $\boldsymbol{D}(n)$ denote the data after n completed outcomes are ascertained. Let n_{initial} be the first instance where either the efficacy criteria (5) or futility criteria (7) are satisfied and enrollment is henceforth terminated, or n_{max} if the efficacy criteria or futility criteria are not satisfied at any point. Let n_{final} be the final sample size which includes patients whose outcomes were in progress at the time of enrollment termination. The expected trial duration is determined based the projected time needed for subject enrollment and outcome ascertainment for a number of subjects equal to the expected sample size.

2.3.2 *Posterior mean and coverage probability.*   Any prior of the form (10) and the data likelihood $\mathcal{L}(\theta|\mathbf{D})$ are used to evaluate the posterior distribution for θ denoted $\pi(\theta|\mathbf{D})$. The posterior mean at the final sample size is $E(\pi(\theta|\mathbf{D}(n_{\text{final}}))$. The coverage probability is the probability of the generating value of θ to be in an equal-tailed $100(1-\alpha)\%$ credible interval, which is the interval $[\theta^{(\alpha/2)}, \theta^{(1-\alpha/2)}]$ based on the quantiles of the posterior distribution $\pi(\theta|\mathbf{D}(n_{\text{final}}))$.

2.3.3 *Probability of efficacy, futility, and inconclusive findings.*   The probability of stopping early for efficacy and futility are the probability that the efficacy criteria (5) or futility criteria (7) are satisfied at an interim analysis with data $\mathbf{D}(n_{\text{initial}})$. The final determination

of efficacy and futility refer to the corresponding criteria begin satisfied with data $\mathbf{D}(n_{\text{final}})$. To evaluate the frequentist properties of Bayesian designs, type I error and power are based on $\text{eff}(\boldsymbol{D}(n_{\text{initial}}))$. Inconclusive findings refers to situations where neither the efficacy or futility criteria are satisfied.

2.3.4 *Distribution of final posterior probability of efficacy given interim stoppage.* Given that the efficacy criteria was satisfied as an interim analysis, it is of interest to compare the associated value once subjects in follow-up have completed outcomes. It is of particular interest when $\text{eff}(\boldsymbol{D}(n_{\text{initial}})) > 1 - \epsilon$ and $\text{eff}(\boldsymbol{D}(n_{\text{final}})) \leqslant 1 - \epsilon$, that is, the threshold for substantial evidence is satisfied for an interim analysis but is no longer satisfied once outcomes from patients in progress are ascertained. It is of interest to determine the probability of such an occurrence and the difference between the efficacy criteria evaluated at the different time points.

2.4 *Computation*

The efficacy criteria (5) and futility criteria (7), 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). In the examples in Section 3, these quantities are $1-$ and $2-$dimensional integrals which are evaluated using numerical integration in R (R Core Team, 2017), in particular using the pracma package (Borchers, 2019).

3. Examples

3.1 *Single-Arm Proof-of-Activity Trial with Binary Endpoint*

3.1.1 *Motivating Example.* 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). Infliximab was given to all patients at the 5mg/kg dose at weeks 0, 2, and 6, and the primary endpoint was response

at week 8. Response was measured by improvement in disease severity scores. The trial was conducted between August 2006 and June 2010. Patients were enrolled over approximately 33.5 months (approximately 1 patient enrolled per 17 days). The sample size of 60 patients was chosen to ensure 12% precision in estimating the true response proportion with 95% confidence interval, assuming a response rate of 0.67 as was observed among adults from the ACT 1 and ACT 2 trials (Rutgeerts et al., 2005) at the same 5mg/kg dose (N = 242). A 95% confidence interval that excluded 0.40 was determined to be a clinically significant result. At week 8, clinical response was observed in 44 out of 60 (73.3%) patients.

3.1.2 *Model Formulation & Prior Elicitation.*   We use this trial as a template to demonstrate our framework for sequential monitoring. The data $\boldsymbol{D}$ are assumed to be independent Bernoulli random variables with common response probability θ. This trial has a superiority hypothesis with null response value is $\theta_0 = 0.4$. For purposes of monitoring, a highly efficacious response probability is $\theta_1 = 0.67$ and an intermediate response value is $\theta_m = 0.535$. The monitoring priors used are a concentrated skeptical prior and a default enthusiastic prior. A comparison of the various combinations of the monitoring priors is provided in Appendix 5.4.

An inference prior is defined as the mixture (10) with $\omega = 0.5$. The maximum sample size was increased from 60 to increase the probability of stopping the trial early due to efficacy at response proportion values between the null $\theta_0 = 0.4$ and the adult data $\theta_1 = 0.67$. The sample size of $n_{\max} = 112$ based on a frequentist design to have 80% power to conclude efficacy with the skeptical monitoring prior at a true response proportion of $\theta_m = 0.535$. This increased power provides a better setting to demonstrate how the flattening or concentration of the skeptical monitoring prior impacts the operating characteristics. Consider an interim analysis occurring after every 2 patients complete follow-up.

3.1.3 *Preposterior Analysis of Operating Characteristics.* The following are generated from $10,000$ simulations for the trial described in Section 3.1.2. Figure 4(a) shows the operating characters of this particular trial design.

When the true response probability is $\theta_0 = 0.4$, there is a 3.9% probability of stopping the trial early for efficacy. This is higher than the nominal rate of 0.025 due to the frequent data monitoring after every 2 completed outcomes (see also Section 3.1.4). When the true response probability is $\theta_m = 0.535$, there is a 80.8% probability of stopping the trial early for efficacy, which is comparable to a frequentist design with one analysis at the maximum sample size of 112. The probability of stopping the trial for efficacy approaches 100% for response values at $\theta_1 = 0.67$, and accordingly the expected number of 35.8 patients is the lowest. The probability of stopping early for futility is the highest at the null response rate of $\theta_0 = 0.4$, and then decreases as data is shown to be compatible with the alternative hypothesis that $\theta > \theta_0$. The impact of futility monitoring at $\theta_0 = 0.4$ is an expected sample size of 79.2. The probability of inconclusive findings is the greatest at $\theta = 0.4675$, where there is often modest evidence that $\theta > \theta_0$ but not enough to satisfy the efficacy criteria at by the chosen maximum sample size. Accordingly, the expected sample size of 91.4 patients is the highest at this value.

For all the generating true response probabilities considered, the expected sample size is much less than the maximum of 112. The trial reaches the maximum sample size it at most 42% of cases at the response rate of $\theta = 0.4675$. Therefore this design benefits from reaching conclusions of efficacy at acceptable levels while reaching them sooner than the maximum sample size.

A mixture prior that equally weighs the skeptical and enthusiastic monitoring priors was used to derive the posterior mean and coverage probabilities. For the generating values of θ considered, the posterior mean shows bias towards the alternative hypothesis. This is

because the inference prior is still informative towards to alternative hypothesis. The coverage probability of an equal-tailed 95% credible interval is shown to cover the true generating value of θ with more than 95% probability when θ is near θ_0 or θ_1, and less than 95% of the time when θ is an intermediate value (e.g. final coverage probability of 93.1% when $\theta = 0.535$. This is because the mixture prior places high prior probability near θ_0 and θ_1 and less prior probability at intermediate values.

Figure 4(b) addresses the "evidence decrease" described in Section 2.3.4. The probability of these cases occurring is reflected by the percent agreement between interim and final results. For example, in the situations where the trial is stopped early for efficacy at the generating value of $\theta_0 = 0.4$, the final efficacy criteria is satisfied 48.2% of the time. This means that the outcomes in progress are likely to show evidence towards the null enough to contradict the conclusion of efficacy. When the trial is stopped early for efficacy and the generating value of $\theta_1 = 0.67$, the final efficacy criteria is satisfied in 92.0% of cases. The distribution of the efficacy criteria given the final data for these cases are shown. This shows that even in the cases where there is evidence decrease, the final efficacy criteria is still similar to the $1 - \epsilon = 0.975$ threshold. For example, when the generating value of $\theta_1 = 0.67$, and in the 8% of situations where there is evidence decrease, over 90% of these cases have a final efficacy criteria of over 0.93.

[Figure 4 about here.]

3.1.4 *Type 1 Error Rate by the Frequency of Data Monitoring.*   Figure 5 shows the probability of the efficacy criteria being satisfied at an interim and final analysis when $\theta = \theta_0$. The monitoring frequency is 1 when an interim analysis is made after every completed outcome (i.e. fully sequential), and is 112 (the maximum sample size) if the only analysis is done at the maximum sample size. When there is only a single analysis completed at the maximum sample size, the probability of determining efficacy is only 0.8%. This is because the determination

of efficacy is made with an informative skeptical prior. If the determination of efficacy was made with a non-informative prior, then the probability of the efficacy criteria being satisfied with a single analysis completed at the maximum sample size would be 2.5%. For more frequent data monitoring, the probability of concluding efficacy is increased from 0.8% to levels even higher than the nominal level of 2.5%. This is to be expected; there are more opportunities for a conclusion of efficacy to be made with more frequent interim analyses and it is important that this probability is examined and viewed as an acceptable trade-off for advantages such as the corresponding decrease in expected sample sizes.

[Figure 5 about here.]

3.2 *Parallel Two-Group Design with Binary Endpoint*

3.2.1 *Motivating Example.* Consider the trial "The Pediatric Lupus Trial of Belimumab Plus Background Standard Therapy (PLUTO)" (NCT01649765). Patients were randomized to belimumab 10mg/kg or placebo, and the primary endpoint was response at week 52. A binary response variable was measured by improvement in disease severity scores. The goal was to test for superiority of belimumab to placebo. The study start date was September 7, 2012, and the primary completion date was January 24, 2018. Since the follow-up period is 52 weeks the last enrollment is estimated to be a year prior to the primary competition date yielding an average enrollment rate of one enrollment per 17 days. The study design included enrollment of 100 patients, the first 24 patients randomized in a 5:1 ratio (belimumab:placebo) and the remaining 76 patients would be randomized in a 1:1 allocation ratio. Therefore, 58 patients would be randomized to belimumab and 42 to placebo. The sample size was based on feasibility constraints rather than a power calculation, and the study was terminated after 93 patients enrolled.

The results of this trial were inconclusive with the 93 patients. A post-hoc Bayesian analysis that gave 55% weight to adult data with response rate 0.51 was sufficient to provide evidence

of positive treatment effect (Travis et al., 2019). 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.

3.2.2 *Model Formulation & Prior Elicitation.*   We use this trial as a template to demonstrate

our framework, in particular the performance of our response-adaptive monitoring prior

(13). The adaptive prior is useful in this setting because the trial is underpowered and

the prospective use of an informative monitoring prior for efficacy provides a different

perspective than a post-hoc analysis with an informative prior. The data $\boldsymbol{D}$ are assumed to be

independent Bernoulli random variables with response probability η_0 for the placebo group

and η_1 for the treatment (IP for investigational product) group. This trial has a superiority

hypothesis of IP to control with null treatment difference value of $\theta_0 = 0$. For purposes of

monitoring, a highly efficacious difference probability is $\theta_1 = 0.12$ (Travis et al., 2019) and

an intermediate response value is $\theta_m = 0.06$ An estimate for the pediatric response rate is

$\eta_0 = 0.39$, which is 0.12 less than the adult data response rate.

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 (13). A 3-part mixture inference

prior of the form (14) will be used to estimate the posterior mean and coverage probabilities

for θ. The locally non-informative prior is $\pi_{NI}(\theta, \eta_0) = \pi_{NI}(\theta) \times \pi(\eta_0|\theta)$. For the skeptical,

enthusiastic, and locally non-informative priors, $\pi(\eta_0|\theta)$ is a flattened prior with modal value

0.39 and tail probability condition $P(\eta_0 > 0.59|\theta) = 0.025$.

A maximum sample size of $n_{\max} = 100$ was chosen based on the trial protocol. A minimum

sample size of $n_{min} = 70$ 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 competed after every 2 patients have completed outcomes beginning at n_{min}.

3.2.3 *Preposterior Analysis of Operating Characteristics.* Figure 6(A) shows the operating characteristics of this design using the adaptive weight monitoring prior and the 3-part mixture inference prior. The generating response probability in the placebo group is 0.39, and the generating response probability in the treatment group is based on risk differences θ in $\{0, 0.06, 0.12, 0.18, 0.24\}$. When $\theta = 0$, there is a 0.008 probability of concluding efficacy at an interim analysis. This probability increases to 0.193 at $\theta_1 = 0.12$. At the effect size 0.24 the probability of concluding efficacy increases to 0.711, and the expected final sample size is 90.3 patients.

Figure 6(B) shows the probability of stopping early for efficacy using a fixed weight mixture prior of the form (10) for efficacy monitoring with a fixed choice of ω chosen at the outset to be in the set $\{0.25, 0.5, 0.75, 1\}$, and the associated sample sizes. Note that $\omega = 1$ corresponds to the traditional skeptical prior, $\omega = 0.5$ gives equal weight to the skeptical and enthusiastic components, and $\omega = 0.25$ most of the weight is applied to the enthusiastic component. The adaptive weight mixture behaves similar to using a fixed weight prior that equally weighs the skeptical and enthusiastic prior.

[Figure 6 about here.]

Figure 7(A) shows the prior-data compatibility assessments $\psi^{(S)}$, $\psi^{(E)}$, $\psi^{(NI)}$ 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 7(B) shows the 3-part mixture inference prior weights ω_S, ω_E, ω_{NI} using (14)-(15) by observed risk difference. Recall our goal is to create a mixture prior which favors the skeptical or enthusiastic components in areas where high compatibility is demonstrated for those components, and favors the

locally non-informative prior if both the skeptical and enthusiastic components show low compatibility. To this end, the skeptical and enthusiastic components have the highest weight when the observed data is aligned with the prior mode, and the locally non-informative prior has highest weight towards extreme values of the observed response difference.

[Figure 7 about here.]

4. Discussion

Bayesian methods are well suited for sequentially monitored clinical trials because of their natural interpretations and ability to incorporate external evidence through prior distributions. Monitoring priors used for efficacy and futility stopping fundamentally determine the operating characteristics of the trial, in addition to factors such as the frequency of data monitoring and number of patients in progress at enrollment termination.

This paper presents a structured framework for designing a Bayesian sequentially monitored clinical trial. The generalized normal distribution gives a flexible and intuitive way to create monitoring priors. It is required that the practitioner specify the modal value, a quantile condition, and an additional parameter which can concentrate or flatten the distribution around the modal value (with the normal distribution as the default case). This paper demonstrates how the operating characteristics are affected by the choice of monitoring priors.

Presentation of these concepts was simplified by generic choices for design parameters. The same quantity $1 - \epsilon$ was used as the threshold for substantial evidence in efficacy and futility monitoring (6) and (8). In practice, the value of ϵ could be different for these two purposes (e.g. $\epsilon_{\text{eff}} = 0.025$, $\epsilon_{\text{fut}} = 0.05$). The intermediate effect size θ_m used in futility monitoring (7) was chosen to be $\theta_m = (\theta_0 + \theta_1)/2$, but another effect size between θ_0 and θ_1 could be considered. Different assessments of prior-data conflict could be used to create a

dynamic prior (10). Other distributions than the generalized normal could be used to create monitoring priors with the desired modal value and tail probability constraint (Section 2.2.1). Each of these quantities could be modified to better suit the needs of the practitioner.

Although the examples provided are superiority trials with binary data and response probabilities as the parameter of interest, the framework applies to any type of data likelihood and parameter of interest on an interval domain. Future work will involve demonstrating the framework in Bayesian clinical trials with survival outcomes.

References

Borchers, H. W. (2019). pracma: Practical numerical math functions.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A (General)* **143,** 383–430.

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

Fayers, P. M., Ashby, D., and Parmar, M. K. B. (1997). Tutorial in biostatistics: Bayesian data monitoring in clinical trials. *Statistics in Medicine* **16,** 1413–1430.

Freedman, L. S. and Spiegelhalter, D. J. (1989). Comparison of bayesian with group sequential methods for monitoring clinical trials. *Controlled Clinical Trials* **10,** 357–367.

Grieve, A. P. (2016). Idle thoughts of a 'well-calibrated' bayesian in clinical drug development. *Pharmaceutical Statistics* **15,** 96–108.

Griffin, M. (2018). Working with the exponential power distribution using gnorm.

Hyams, J., Damaraju, L., Blank, M., Johanns, J., Guzzo, C., Winter, H. S., Kugathasan, S., Cohen, S., Markowitz, J., Escher, J. C., Veereman–Wauters, G., Crandall, W., Baldassano, R., and Griffiths, A. (2012). Induction and maintenance therapy

with infliximab for children with moderate to severe ulcerative colitis. *Clinical Gastroenterology and Hepatology* **10,** 391 – 399.e1.

Kopp-Schneider, A., Wiesenfarth, M., Witt, R., Edelmann, D., Witt, O., and Abel, U. (2019). Monitoring futility and efficacy in phase ii trials with bayesian posterior distributions—a calibration approach. *Biometrical Journal* **61,** 488–502.

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

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., Travers, S., Rachmilewitz, D., Hanauer, S. B., Lichtenstein, G. R., de Villiers, W. J., Present, D., Sands, B. E., and Colombel, J. F. (2005). Infliximab for induction and maintenance therapy for ulcerative colitis. *New England Journal of Medicine* **353,** 2462–2476.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994). Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **157,** 357–416.

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

Travis, J., Neuner, R., Rothwell, R., Levin, G., Nie, L., Niu, J., Marathe, A., and Nikolov, N. (2019). Application of bayesian statistics to support approval of intravenous belimumab in children with systemic lupus erythematosus in the united states. In *2019 ACR/ARP Annual Meeting.*

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

Zhu, H. and Yu, Q. (2015). A bayesian sequential design using alpha spending function to

control type i error. *Statistical Methods in Medical Research* **26,** 2184–2196.

Zhu, L., Yu, Q., and Mercante, D. E. (2019). A bayesian sequential design for clinical trials with time-to-event outcomes. *Statistics in biopharmaceutical research* **11,** 387–397. 32226580[pmid] PMC7100880[pmcid].

5. Supplementary Material

5.1 *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, η) specified over the parameter space defined with respect to each of the hypotheses (e.g. $\pi\left(\theta, \eta \middle| H_i\right)$ for $i = 0, 1$).

The posterior probability of hypothesis H_i is given by

$$p(H_i|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|H_i) \cdot p(H_i)}{p(\boldsymbol{D}|H_0) \cdot p(H_0) + p(\boldsymbol{D}|H_1) \cdot p(H_1)}, \tag{16}$$

where $p(\boldsymbol{D}|H_i) = \int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta|H_i)d\theta$ is the marginal likelihood associated with hypothesis H_i. In practice, most Bayesian methods for clinical trials perform hypothesis testing 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 overall and compute the posterior distribution. The posterior probability that $\theta \in \Theta_i$ is given by

$$P(\theta \in \Theta_i|\boldsymbol{D}) = \frac{\int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_i)d\theta \cdot P(\theta \in \Theta_i)}{\sum_{j=0,1} \int_{\Theta_j} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_j)d\theta \cdot P(\theta \in \Theta_j)} \tag{17}$$

where $P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta)d\theta$. We can readily see that the $P(\theta \in \Theta_i|\boldsymbol{D})$ is equal to $p(H_i|\boldsymbol{D})$ if one takes $p(H_i) = P(\theta \in \Theta_i)$ and $\pi\left(\theta \middle| H_i\right) = \pi\left(\theta \middle| \theta \in \Theta_i\right)$ 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 $P(\theta \in \Theta_i|\boldsymbol{D})$ as the probability that hypothesis H_i is true.

5.2 *Monitoring Prior Parameterization*

5.2.1 *Normal Distribution $\mathcal{N}_p(\tilde{\mu}, q)$.*   Suppose $\theta \sim \mathcal{N}(\mu, \sigma)$ is a normal random variable that satisfies $\text{mode}(\theta) = \tilde{\mu}$ and $P(\theta \leqslant q) = p$. The values for the mean and standard deviation are $\mu = \tilde{\mu}$ and $\sigma = \frac{q-\mu}{\Phi^{-1}(p)}$, where Φ denotes the cumulative distribution function for a standard normal distribution and Φ^{-1} denotes its quantile function. Therefore we can denote the distribution with the desired mode and tail probability constraint as $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$, which is well-defined for values (μ, q, p) that satisfy $\frac{q-\mu}{p-0.5} > 0$. Since the normal distribution is completely specified by (μ, σ), quantities such as $P(\theta \leqslant \tilde{q})$ are also specified for any $\tilde{q} \in \mathbb{R}$. In particular, if $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$ then $P(\theta \leqslant \frac{q+\mu}{2}) = \Phi(\frac{\Phi^{-1}(p)}{2})$. Furthermore, $P(\theta \in (\mu, \frac{q+\mu}{2})) = |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.

5.2.2 *Generalized Normal Distributions $\mathcal{GN}_p(\tilde{\mu}, q, \gamma)$.*   Let $F_{\mu, \alpha, \beta}$ denote the cumulative distribution function of $\mathcal{GN}(\mu, \alpha, \beta)$. The CDF of a generalized normal random variable $\theta \sim \mathcal{GN}(\mu, \alpha, \beta)$ can be expressed as (Griffin, 2018)

$$P(\theta \leqslant q|\mu, \alpha, \beta) = \frac{1}{2} + \frac{\text{sign}(q - \mu)}{2} \int_0^{|q-\mu|^\beta} \frac{w^{1/\beta - 1}}{\alpha \Gamma(1/\beta)} \exp\left\{ -\left(\frac{1}{\alpha}\right)^\beta w \right\} dw \qquad (18)$$

Define $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ as the generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ that satisfies $\text{mode}(\theta) = \tilde{\mu}$, $P(\theta \leqslant q) = p$, and $P(\theta \in (q, \frac{q+\mu}{2})) = \gamma \cdot |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$. The mode is equal to $\mu = \tilde{\mu}$, and α and β are determined to minimize the function $(F_{\mu, \alpha, \beta}(q) - p)^2 + (F_{\mu, \alpha, \beta}(\frac{q+\mu}{2}) - \gamma \cdot |p - \Phi(\frac{\Phi^{-1}(p)}{2})|)^2$ with box-constrained optimization (Byrd et al., 1995).

Using this notation to define the distributions in Figure 2 yields the default enthusiastic prior as $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$, the concentrated enthusiastic prior as $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 0.75)$, the flattened enthusiastic prior as $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1.5)$, and the locally non-informative prior as $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \frac{\theta_0 + \theta_1}{2}, q = \frac{3\theta_0 - \theta_1}{2}, \gamma = 1.5)$.

5.2.3 *Truncated Generalized Normal Distribution $\mathcal{GN}_{p,\Theta}(\tilde{\mu}, q, \gamma)$.*   The density for a generalized normal distribution truncated to the interval domain $\Theta = (\theta_{min}, \theta_{max})$, denoted by $\mathcal{GN}_\Theta(\mu, \alpha, \beta)$,

is $f(\theta) = c \cdot \exp\left\{-\frac{|\theta-\mu|^{\beta}}{\alpha}\right\} I(\theta \in \Theta)$ where $c = \frac{\beta}{2\alpha\Gamma(1/\beta)}(F_{\mu,\alpha,\beta}(\theta_{max}) - F_{\mu,\alpha,\beta}(\theta_{min}))^{-1}$. Define $\theta \sim \mathcal{GN}_{p,\Theta}(\tilde{\mu}, q, \gamma)$ as the truncated generalized normal distribution $\mathcal{GN}_{\Theta}(\mu, \alpha, \beta)$ that satisfies $\mathrm{mode}(\theta) = \tilde{\mu}$, $P(\theta \leqslant q) = p$, and $P(\theta \in (q, \frac{q+\mu}{2})) = \gamma|p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.

Using this notation to define the distribution in Figure 3 yields the joint distribution $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$, as $\pi(\theta) \sim \mathcal{GN}_{p=0.975,\Theta=[-1,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 1)$ and $\pi(\eta|\theta) \sim \mathcal{GN}_{p=0.975,\Theta=[\max(-\theta,0),\min(1,1+\theta)]}(\tilde{\mu} = \eta_0, q = \eta_1, \gamma = 1)$.

5.3 *Type 1 Error Rate Depending on Enrollment Schemes*

Recall Figure 5 from Section 3.1 which showed Type 1 error properties for the single-arm design. Figure 8 shows results from a design that has a longer follow-up period. The interim sample sizes are the same for each monitoring frequency, however, the final sample sizes under the longer follow-up designs are much larger (over 20 patients in follow-up for monitoring frequencies of 8 or fewer, compared to approximately 6 patients in the shorter follow-up designs). The final probability of efficacy criteria being satisfied is generally slightly lower in the longer follow-up design, which is what we would expect since the larger final sample size contains more data consistent with a null result.

[Figure 8 about here.]

5.4 *Robustness of Parameterizations of Monitoring Priors*

The analyses done in Section 3.1 used a concentrated skeptical prior and default enthusiastic prior. In this section we show the four possible designs using the combinations of skeptical and enthusiastic prior given in Figure 2.

Figures 9-10 shows what happens when the enthusiastic prior shifts from default to flattened, with the skeptical prior remaining fixed. Note that in the region between θ_0 and $\frac{\theta_0+\theta_1}{2}$ as the enthusiastic prior shifts from default to flattened, (a) the probability of stopping early for futility increases (b) the probability of inconclusive findings decreases and (c) the

intermediate and final sample sizes decrease. This is because the enthusiastic prior gives more mass in for this region of θ. The flattened enthusiastic prior was used in Section 3.1 to enhance the ability of futility monitoring to reduce the sample size.

Contrasting 9 and 10, we see that the probability of stopping early for efficacy is much higher at θ_0 when the default skeptical prior is used rather than the concentrated skeptical prior. This is because the default skeptical prior has less mass around $\theta = \theta_0$, therefore it is easier to convince the skeptic that $\theta > \theta_0$ under the null result $\theta = \theta_0$. The concentrated skeptical prior was used in Section 3.1 to limit this probability and provide better Type 1 error control.

The choice of skeptical and enthusiastic prior affects the analysis, and their specification (e.g. default, skeptical, enthusiastic) should be made with these properties in mind.

[Figure 9 about here.]

[Figure 10 about here.]
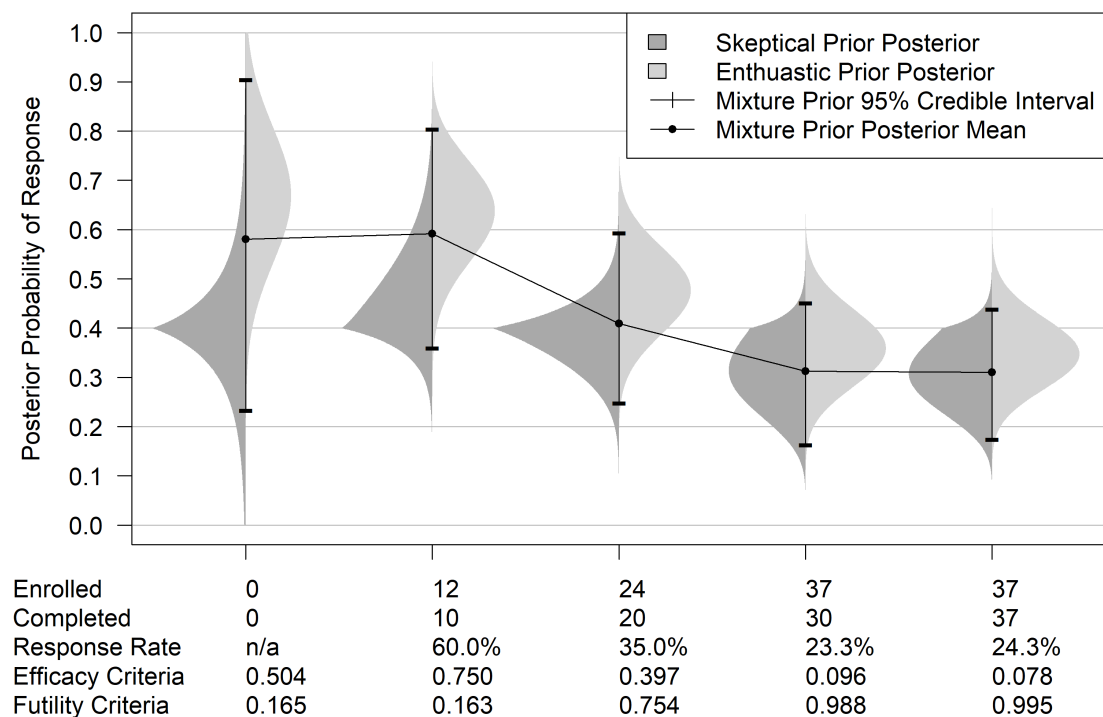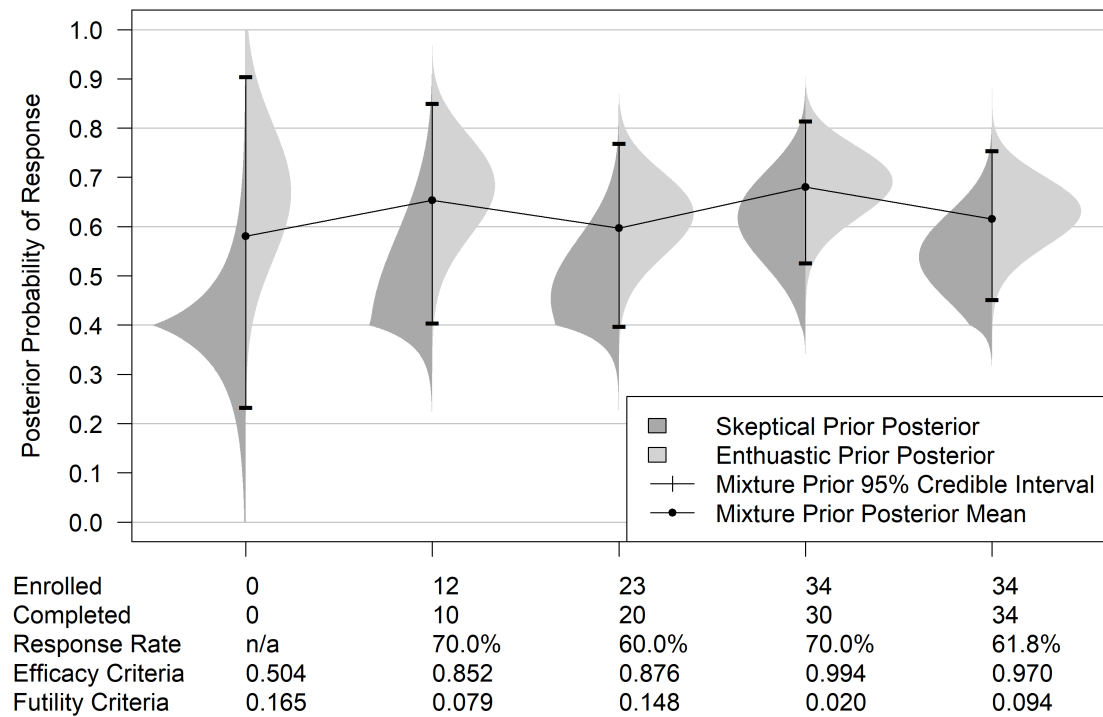
Figure 1. Example paths for the trial described in Section 3.1.2. A, Early stoppage for efficacy. B, early stoppage for futility.
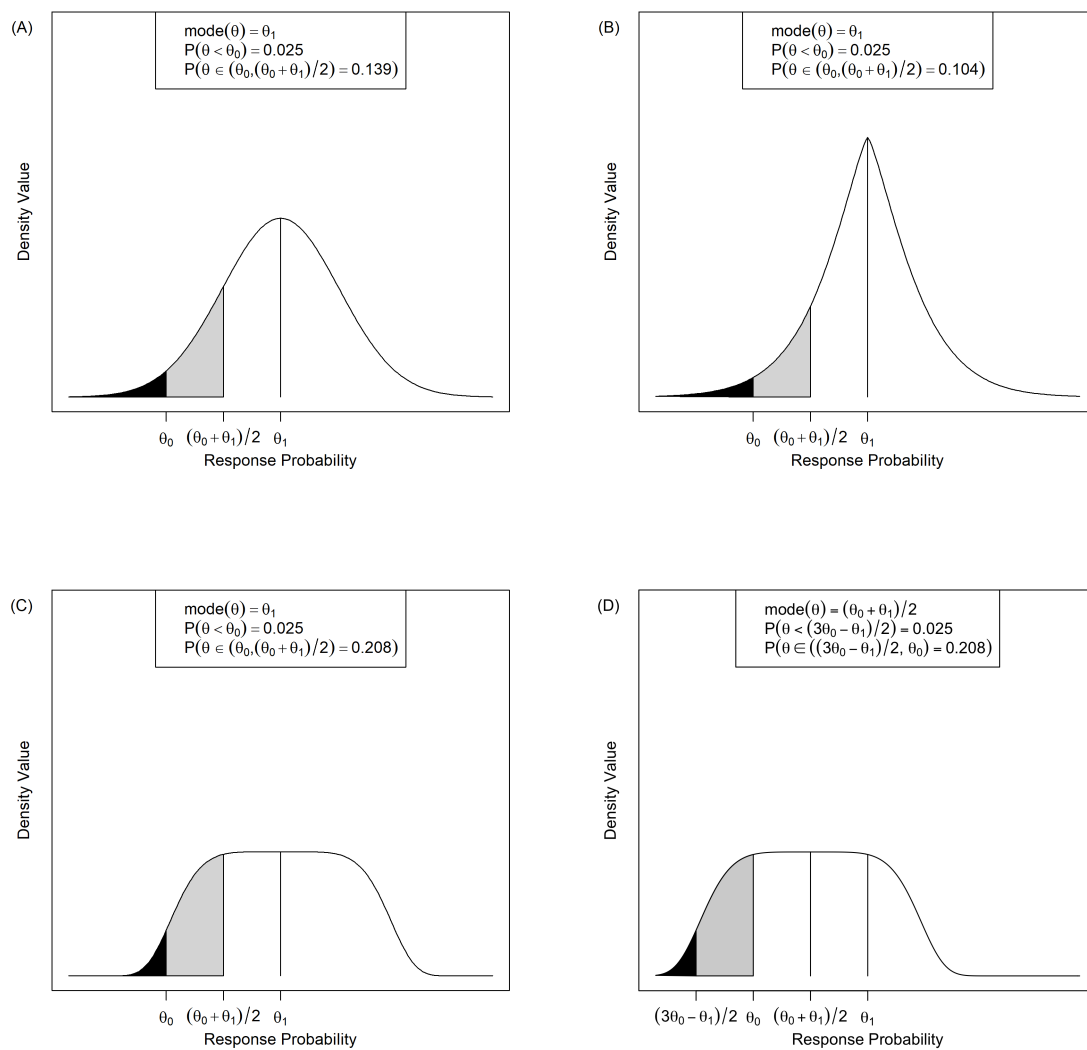
Figure 2.   A, Default enthusiastic prior. B, Concentrated enthusiastic prior. C, Flattened enthusiastic prior. D, Locally non-informative prior
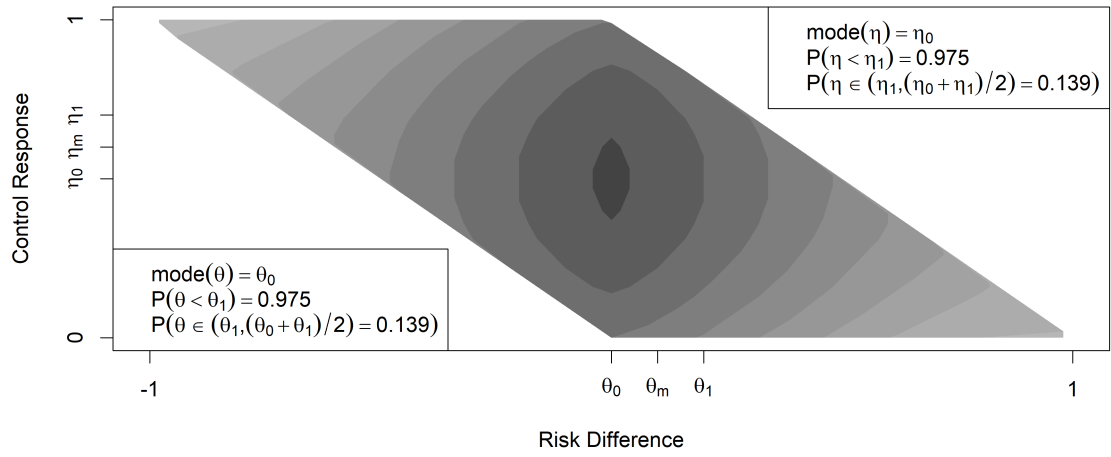
Figure 3. Joint distribution $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$, where $\pi(\theta)$ is a default skeptical prior truncated to $[-1, 1]$, and $\pi(\eta|\theta)$ is a default skeptical prior truncated to $[\max(-\theta, 0), \min(1, 1 + \theta)]$.
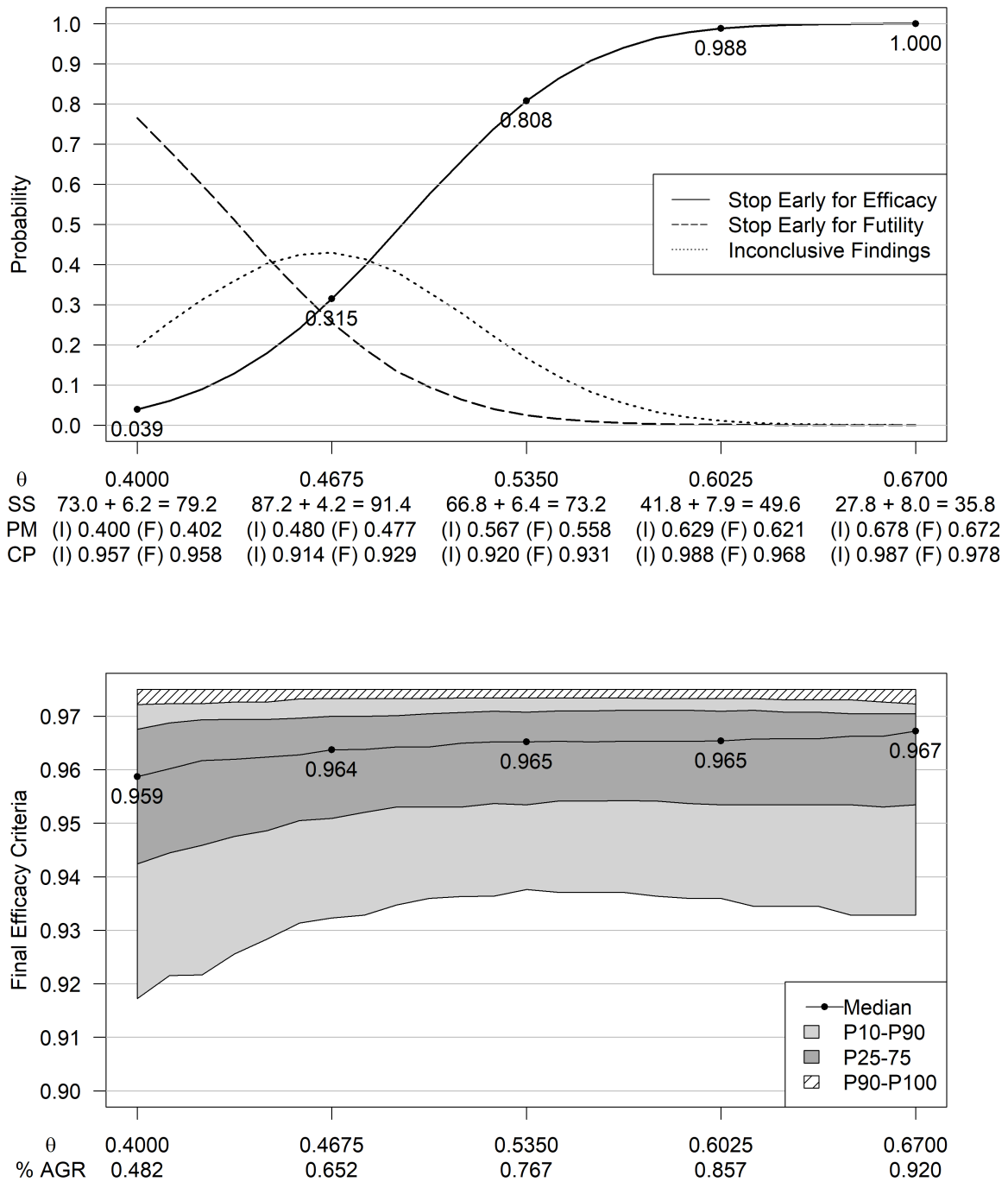
Figure 4. A, Sequential design properties. (SS; sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis). B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; Percent of agreement between final and interim posterior probabilities relative to $1 - \epsilon$ threshold)
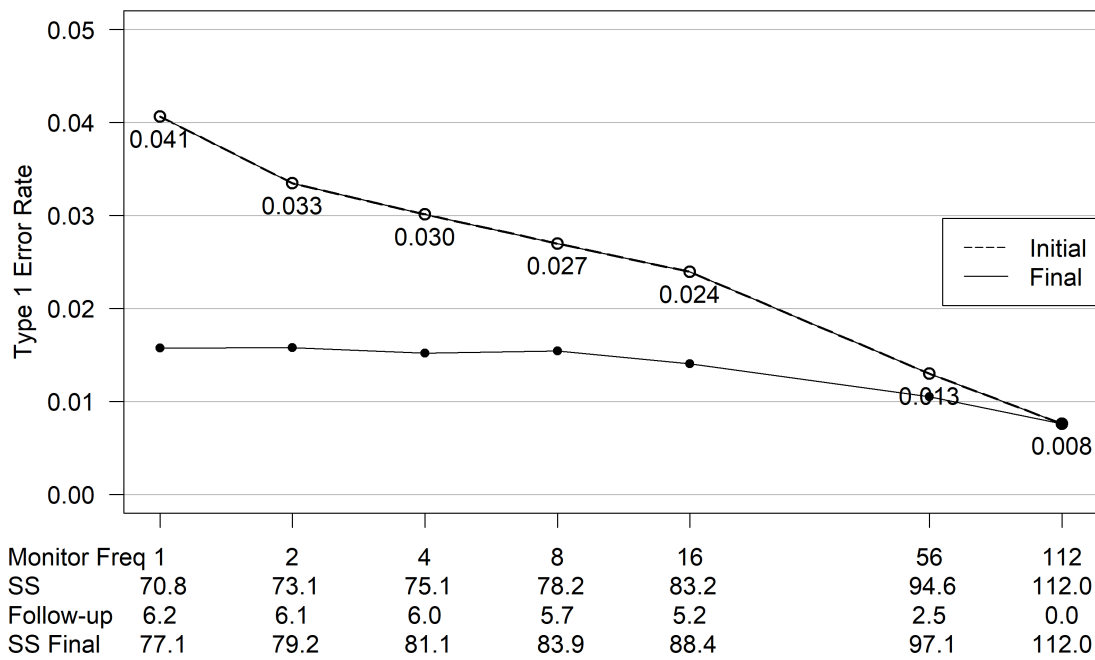
Figure 5. Probability of efficacy criteria being satisfied when $\theta = \theta_0$. SS; sample size. Monitor Freq; monitoring frequency.
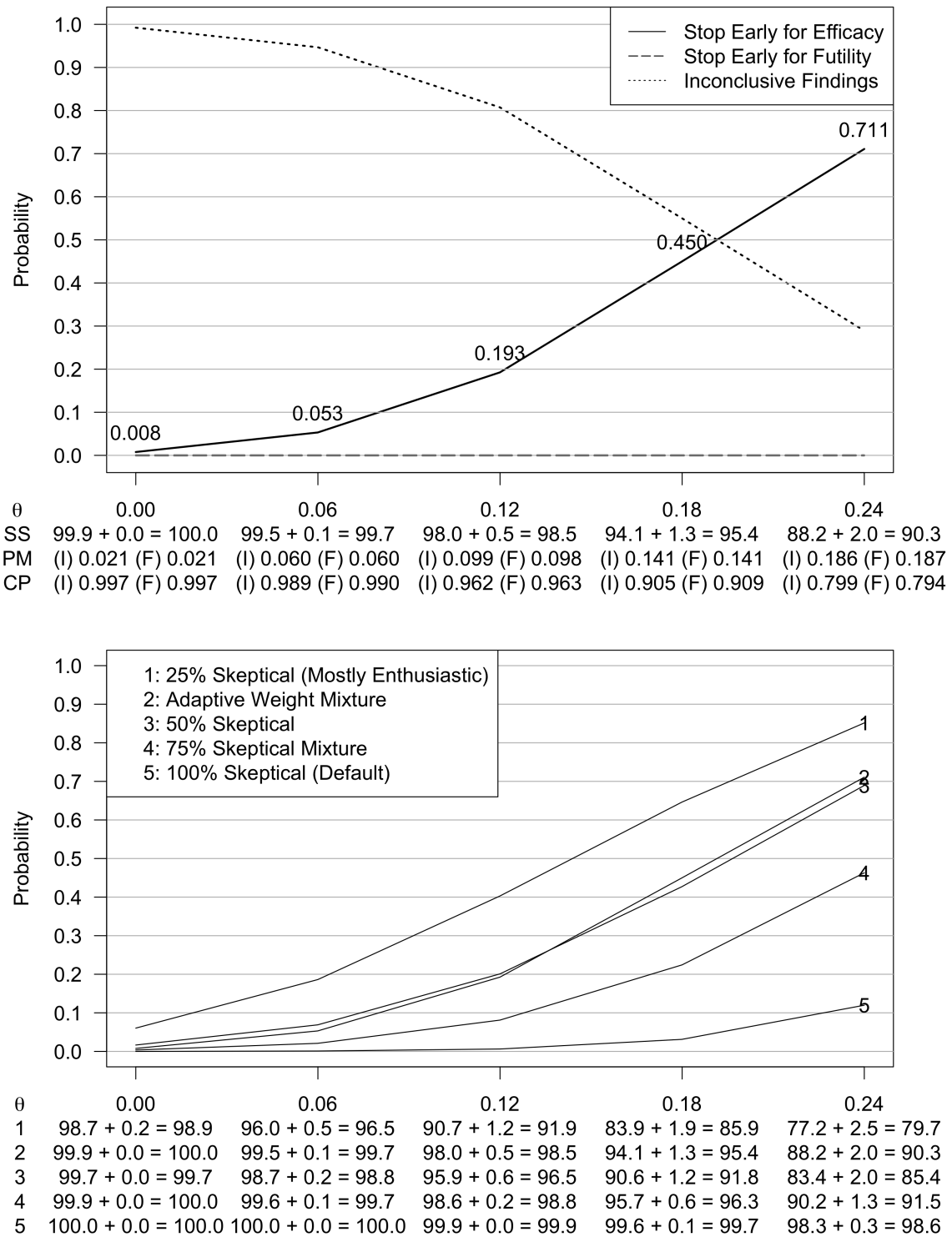
Figure 6. A, Sequential design properties using adaptive weight monitoring prior. (SS; sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis). B, Probability of stopping for efficacy and associated sample sizes by true IP response rate θ with different choices of efficacy monitoring prior.
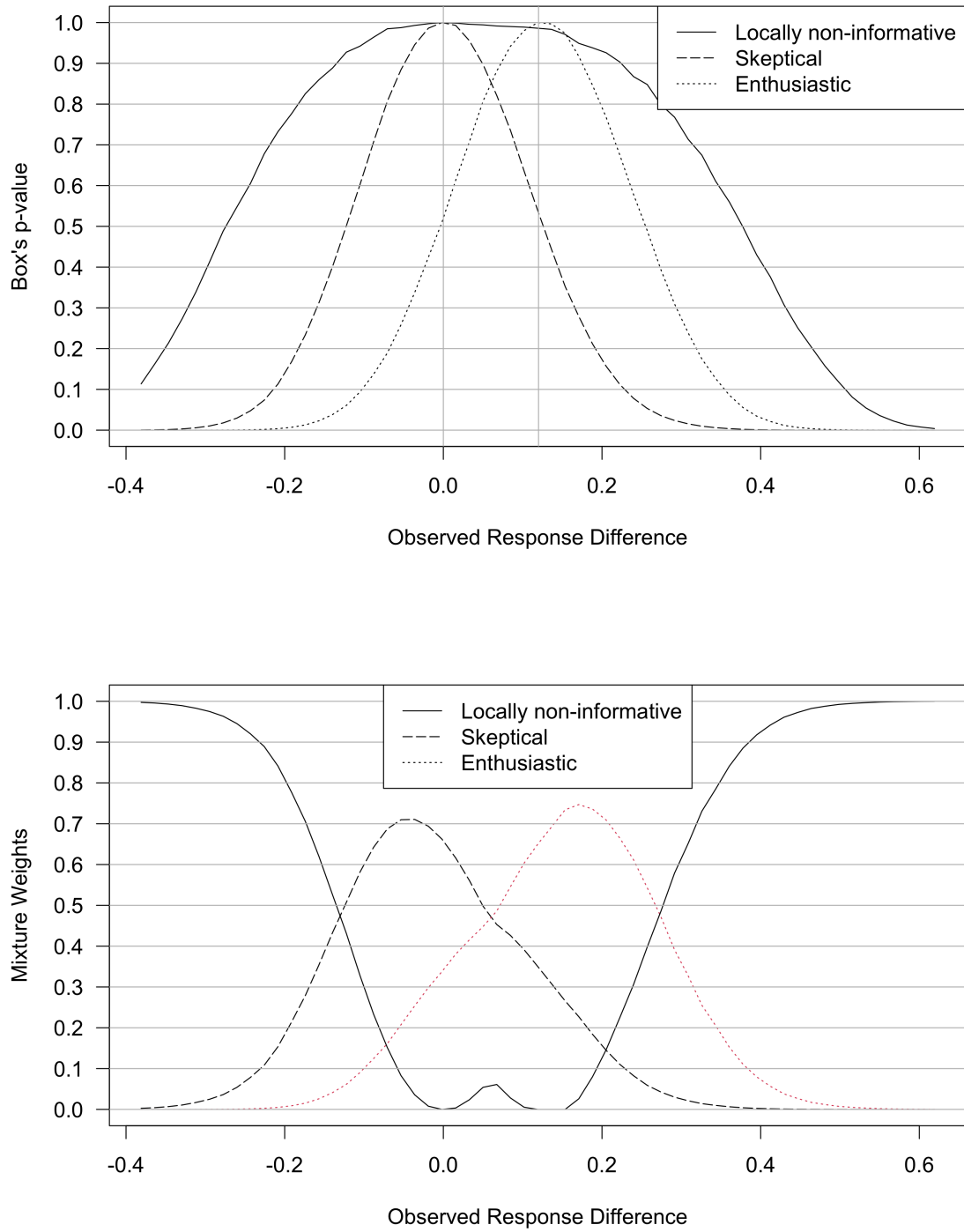
Figure 7. A; Prior-data compatibility assessments $\psi^{(S)}$, $\psi^{(E)}$, $\psi^{(NI)}$ by observed risk difference. B; 3-part mixture inference prior weights ω_S, ω_E, ω_{NI} by observed risk difference.
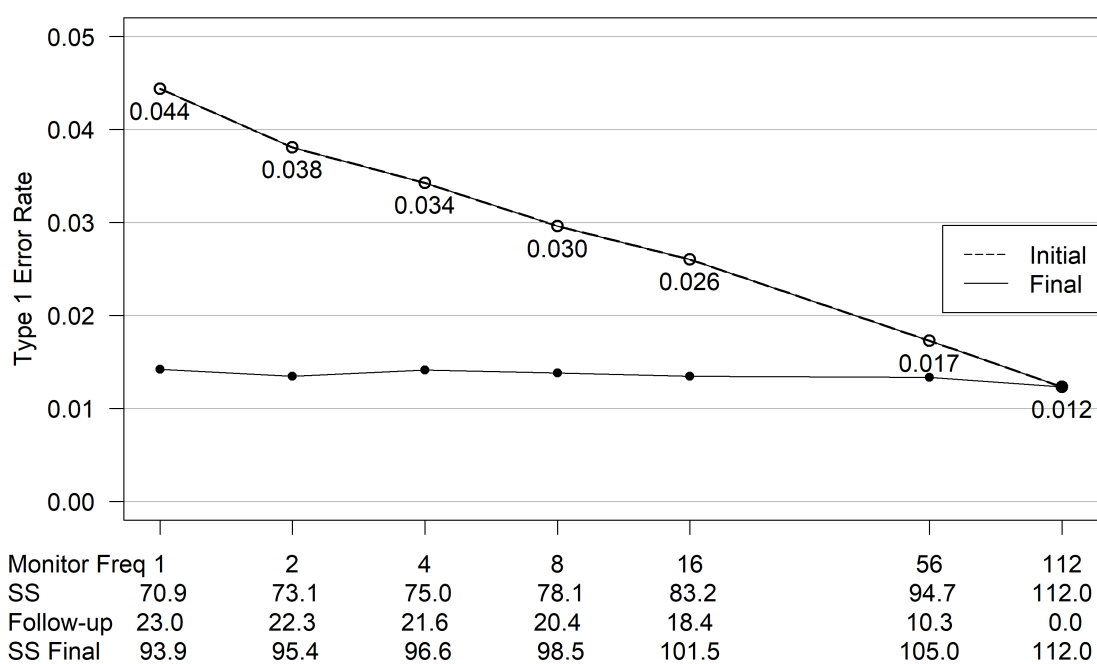
Figure 8. Single arm design from Example 3.1 with a longer follow-up period. Probability of efficacy criteria being satisfied when $\theta = \theta_0$. SS; sample size. Monitor Freq; monitoring frequency.
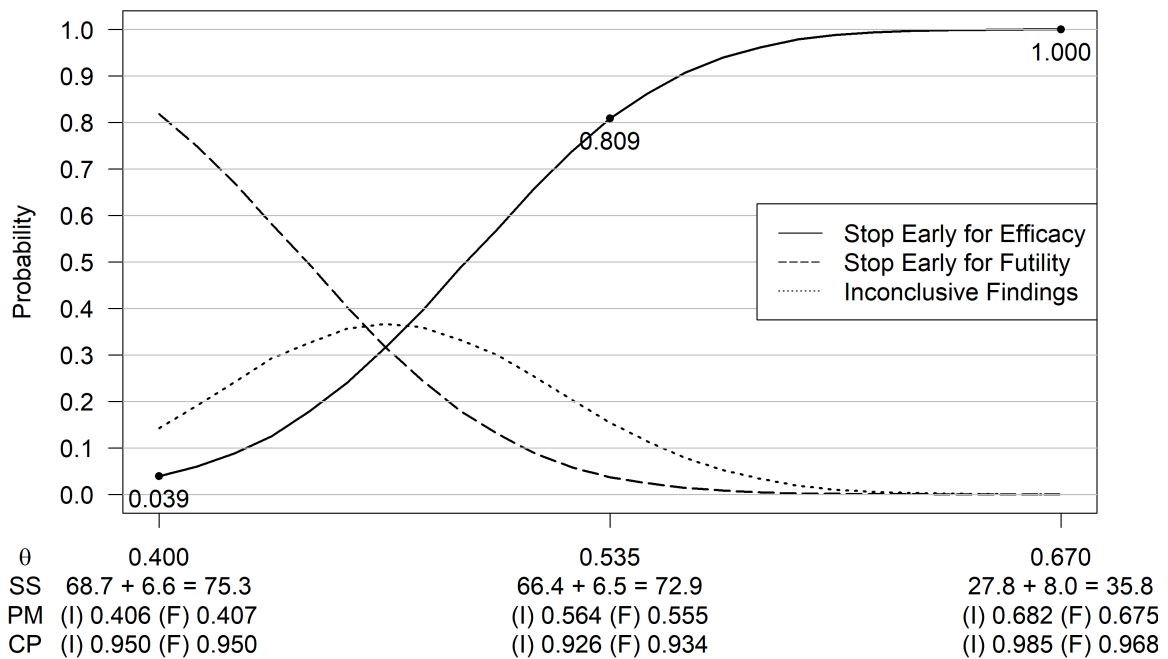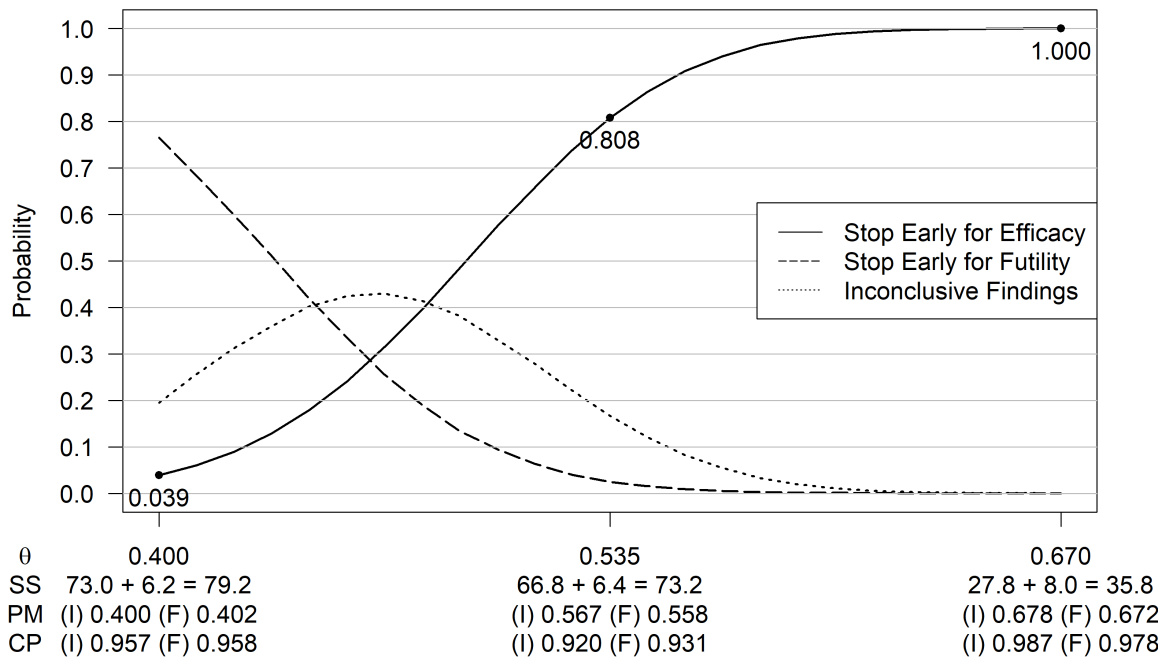
Figure 9. Modification of enthusiastic prior parameterization in Example 3.1. A, default enthusiastic prior (Figure 2(c)). B, flattened enthusiastic prior (Figure 2(d)). Both designs use concentrated skeptical prior (Figure 2(b)).
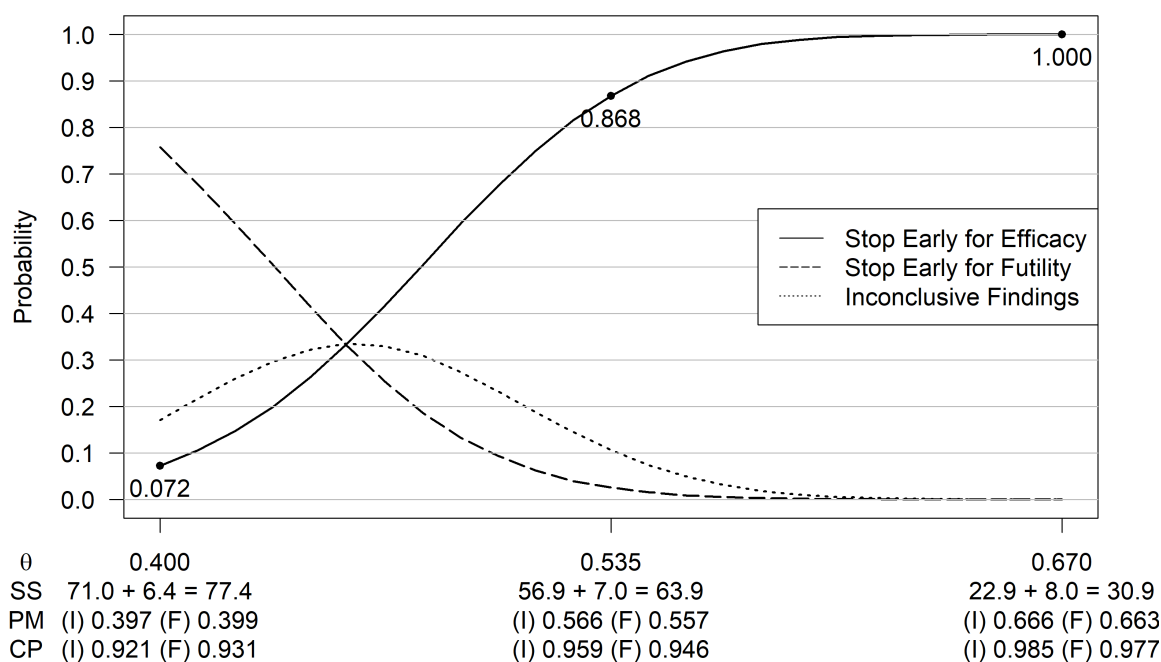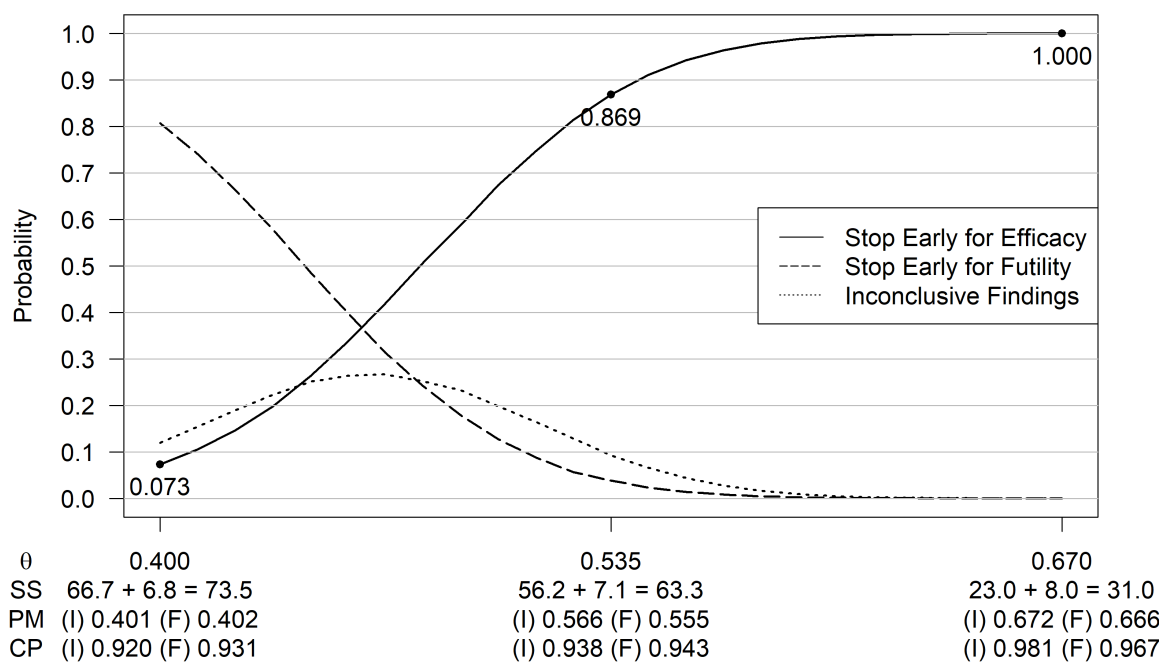
Figure 10. Modification of enthusiastic prior parameterization in Example 3.1. A, default enthusiastic prior (Figure 2(c)). B, flattened enthusiastic prior (Figure 2(d)). Both designs use default skeptical prior (Figure 2(a)).