A Structured Framework for Bayesian Sequential Monitoring in Clinical Trials

Evan Kwiatkowski[†], Eugenio Andraca-Carrera[‡],
Mat Soukup[‡], Matthew A. Psioda[†*]

† Department of Biostatistics, University of North Carolina,
McGavran-Greenberg Hall, CB#7420,
Chapel Hill, North Carolina, U.S.A.

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

April 2, 2020

Abstract

Conclusions from Bayesian clinicial trials are most coherent when specification of prior distributions are intuitively related to the research objectives. A compelling level of evidence is required to terminate enrollment for either efficacy or futility. This paper defines monitoring priors using generalized normal distributions parameterized via three specified quantities (mode and two tail probabilities), which reflect prior belief defined using a consistent definition of compelling level of evidence. These concepts are demonstrated through simulations based on a single-arm proof-of-activity trial in pediatric ulcerative colitis and a parallel two-group design in pediatric lupus.

Keywords: Generalized normal distribution, monitoring priors, real world evidence

*The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

A goal of the 21st Century Cures Act (U.S. Congress 2016) is to expedite the approval process for new drugs and devices through the incorporation of real-world evidence into clinical trial data summaries. Similarly, an objective of the Prescription Drug User Fee Act VI is to enhance the capacity to review "complex adaptive, Bayesian, and other novel clinical trial designs." (U.S. Food and Drug Administration n.d.) Bayesian designs are natural for incorporating external evidence through prior distributions. This is useful in areas where clinical trials would take a lot of time due to slow enrollment or long follow-up periods, and where relevant external data exists. For example, a treatment for a rare pediatric disease might have difficulties with enrollment, and there may be relevant data from adult trials which could augment the trial findings and lead to a conclusion of efficacy sooner. Group sequential designs also have the possible benefits of leading to conclusions earlier, saving time and resources, as well as reducing the exposure of patients to inferior treatments. Bayesian sequential designs which incorporate external evidence therefore have a much increased capacity to expedite the approval process for effective treatments, but they most be carefully planned. The effects of interim analyses and the informativeness of the priors must be well understood. This paper presents a structured or default way to determine prior distributions based on the trial design. Our major contribution is to present methods for the default or automatic selection of prior distributions in a way that is applicable to a wide array of clinical trial designs.

The likelihood principle asserts that any data with the same likelihood function should lead to the same conclusion (Berger & Wolpert 1988). In the context of sequential analysis of clinical trials, the stopping rule that led to the termination of data collection is irrelevant to the conclusion (Barnard 1947, Anscombe 1963, Cornfield 1966*a*,*b*). It is most succinctly as "It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience."(Edwards et al. 1963). Bayesian conclusions are not affected by frequent or even continual monitoring of the data, and such interim analyses should be encouraged to terminate data collection if appropriate (Berry 1989, 1993, Spiegelhalter et al. 1994).

Interpretation of conclusions from the Bayesian perspective is natural when specification of prior distributions are intuitively related to the research objectives (e.g. skeptical and enthusiastic priors) (Freedman & Spiegelhalter 1989, 1992, Spiegelhalter et al. 1993, Fayers et al. 1997) which is necessary for regulatory agencies (Parmar & Machin 1993).

The most common Bayesian metrics for assessing evidence at interim analyses are posterior probabilities and predictive probabilities, and the choice of metric depends on research objective. Posterior probabilities assess the level of evidence in favor of the null or alternative hypothesis, and predictive probabilities determine the capacity for the trial to show convincing evidence in favor of the alternative hypothesis

if more outcomes are ascertained. This paper uses posterior probabilities to highlight how the level of evidence at interim analyses is linked to monitoring prior specification.

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. For a two-arm trial with binary response primary 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 $\mathbf{D}$ represent the data collected in a trial at some point in time. For example, for the two-arm trial example above and assuming no covariates other than the treatment indicator, $\mathbf{D} = \{y_i, z_i : i = 1, ..., n\}$ where y_i is an indicator of response for patient i and z_i is an indicator for whether patient i was assigned the investigational treatment. We use the generic representation $p(\mathbf{D}|\theta, \eta)$ to reflect the density or mass function for the collective data $\mathbf{D}$ as a function of θ and potential nuisance parameters η which could be multi-dimensional. For example, for the two-arm trial example above η would correspond to the response probability for patients receiving the control treatment.

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 $i = 0, 1$). 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.

The posterior probability of hypothesis H_i is given by

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

where $p(\mathbf{D}|H_i) = \int_{\Theta_i} p(\mathbf{D}|\theta)\pi(\theta|H_i)d\theta$ is the marginal likelihood associated with hypothesis H_i. In practice, most Bayesian 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

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

where $P(\theta \in \Theta_i) = \int_{\Theta_i} \pi(\theta)d\theta$. We can readily see that the $P(\theta \in \Theta_i|\mathbf{D})$ is equal to $p(H_i|\mathbf{D})$ if one takes $p(H_i) = P(\theta \in \Theta_i)$ and $\pi\left(\theta\middle|H_i\right) = \pi\left(\theta\middle|\theta \in \Theta_i\right)$ 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 $P(\theta \in \Theta_i|\mathbf{D})$ as the probability that hypothesis H_i is true. For these reasons, in what follows we will refer to the quantity $P(\theta \in \Theta_i|\mathbf{D})$ as the posterior probability of H_i for ease of exposition.

2.1.2 Formalizing the Concept of Substantial Evidence

For testing one-sided hypotheses such as $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta > \theta_0$ using posterior probabilities, the null hypothesis is rejected when $P(\theta > 0|\mathbf{D})$ exceeds a predefined threshold. Leveraging this common practice, we define the posterior probability threshold $1-\epsilon = 0.975$ to be the threshold for what constitutes *substantial evidence* in favor of a claim (e.g., that $\theta > \theta_0$) and, correspondingly, we refer to ϵ as the threshold for insignificant *residual uncertainty*. 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

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

where the subscript π in (1) 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 & 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 to 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 subjects 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 definitions for an 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 $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta > \theta_0$ where θ 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 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 satisfying

$$P_E(\theta > \theta_0) = 1 - \epsilon, \tag{2}$$

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 satisfying

$$P_S(\theta < \theta_1) = 1 - \epsilon. \tag{3}$$

In what follows we refer to (2) and (3) as *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 (**??**) will be such that $p(H_0) \approx p(H_1)$. This prior simply reflects that $\theta \geq \theta_1$ 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. Through 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 $\mathbf{D}$ as

$$\text{eff}(\mathbf{D}) = P(\theta > \theta_0 | \mathbf{D}, \pi_S), \tag{4}$$

and the threshold for substantial evidence to be

$$\text{eff}(\mathbf{D}) > 1 - \epsilon \tag{5}$$

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. However, for this to be the case the observed data must suggest $\theta \ll \theta_0$ since when $\theta = \theta_0$, $P_E(\theta \leq \theta_0|\mathbf{D}) \to 0.5$ for large samples sizes. For this reason, we consider a different approach. Recalling that θ_1 represents a plausible, clinically meaningful treatment effect, we define $\theta_m = c \cdot \theta_0 + (1 - c) \cdot \theta_1$ for $c \in [0, 1]$ to be a less than desirable treatment effect such, that if it is the case that $P_E(\theta < \theta_m|\mathbf{D}) > 1 - \epsilon$, the trial may be stopped due to have a low probability of having a meaningful treatment effect. For example applications in this paper, we consider $c = 0.5$, so that $\theta_m = \frac{\theta_0+\theta_1}{2}$.

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

$$\text{fut}(\mathbf{D}) = P\left(\theta \leq \theta_m \middle| \mathbf{D}, \pi_E\right) \tag{6}$$

and the threshold for substantial evidence to be

$$\text{fut}(\mathbf{D}) > 1 - \epsilon \tag{7}$$

2.1.5 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. In this paper we propose using an inference prior that is a combination of the skeptical and enthusiastic priors that were used for monitoring.

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. Thus, when it 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.

Define the inference prior associated with mixing weight ω as

$$\pi_I(\theta) = \omega \cdot \pi_S(\theta) + (1 - \omega) \cdot \pi_E(\theta), \tag{8}$$

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

The specification of ω is done a priori, and the value $\omega = 1/2$ will be referred to as an agnostic inference prior since it gives equal weight to the skeptical and enthusiastic prior.

The distribution of θ using the inference prior, $p(\theta|\mathbf{D}, \pi_I)$, will be used to compute summaries of θ such as the posterior mean and quantiles. The posterior distribution for θ using (8) is

$$p(\theta|\mathbf{D}, \pi_I) = \hat{\omega} \cdot p(\theta|\mathbf{D}, \pi_S) + (1 - \hat{\omega}) \cdot p(\theta|\mathbf{D}, \pi_E)$$

where the updated mixing weight is

$$\hat{\omega} = \frac{\omega \cdot p(\mathbf{D}|\pi_S)}{\omega \cdot p(\mathbf{D}|\pi_S) + (1 - \omega) \cdot p(\mathbf{D}|\pi_E)}$$

where $p(\mathbf{D}|\pi_S) = \int p(\mathbf{D}|\theta)\pi_S(\theta)d\theta$ and $p(\mathbf{D}|\pi_E) = \int p(\mathbf{D}|\theta)\pi_S(\theta)d\theta$.

2.1.6 Incorporating Prior Information in the Monitoring Priors

The skeptical and enthusiastic monitoring priors can be viewed as a weighted combination such as in the inference prior (8) with weights $\omega = 1$ and $\omega = 0$ respectively. This weighted combination can be used as a replacement for the monitoring prior for any fixed weight ω. For example, a "not-as-skeptical" prior can be defined as $\tilde{\pi_S}(\theta) = 0.75 \cdot \pi_S(\theta) + 0.25 \cdot \pi_E(\theta)$.

Alternatively, ω can be informed by the data. Let $\hat{\theta} = \text{argmax}\{p(\mathbf{D}|\theta)\}$ be the maximum likelihood estimator of θ given the data.

Define

$$\omega = \omega_{min} + (1 - \omega_{min})\frac{\pi_S(\hat{\theta})}{\pi_S(\hat{\theta}) + \pi_E(\hat{\theta})} \tag{9}$$

as a mixing weight that dynamically gives preference to the prior that has the higher density evaluated at $\hat{\theta}$ with a minimum weight ω_{min} given to the skeptical component with $\omega_{min} \in [0, 1]$.

2.1.7 Operating Characteristics

Suppose that the trial has interim analyses based on the number of completed outcomes.

Let $n_{initial}$ be the first instance where either the efficacy criteria or futility criteria are satisfied, 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 subjects whose outcomes were in progress at the time of enrollment termination.

Let $\mathbf{D}(n)$ denote the data after n completed outcomes are ascertained. Type 1 error and power are computed based on eff($\mathbf{D}(n_{initial})$) and eff($\mathbf{D}(n_{final})$). The posterior mean and credible intervals for θ are also evaluated for both $\mathbf{D}(n_{initial})$ and $\mathbf{D}(n_{final})$.

There is the possibility for disagreement between $\text{eff}(\mathbf{D}(n_{initial}))$ and $\text{eff}(\mathbf{D}(n_{final}))$. Of particular interest are cases when $\text{eff}(\mathbf{D}(n_{initial})) > 1-\epsilon$ and $\text{eff}(\mathbf{D}(n_{final})) < 1-\epsilon$, that is, the threshold for substantial evidence is satisfied for an interim analysis but is no longer satisfied once outcomes from subjects in progress are ascertained.

2.2 Monitoring Prior Parameterization

The distributions defined in (2)-(3) each have a required modal value and tail-probability constraint, however, there are still many ways to parameterize such distributions. 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 enthusiastic and skeptical priors.

2.2.1 Normal Distribution $\mathcal{N}_p(\tilde{\mu}, q)$

Definition 1 *Define $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$ as the normal distribution $\mathcal{N}(\mu, \sigma)$ that satisfies mode$(\theta) = \tilde{\mu}$ and $P(\theta < q) = p$.*

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. The distribution $\mathcal{N}_p(\tilde{\mu}, q)$ 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 $P(\theta < \tilde{q})$ are also specified for any $\tilde{q}$. In particular, if $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$ then $P(\theta < \frac{q+\mu}{2}) = \Phi(\frac{\Phi^{-1}(p)}{2})$. Furthermore, $P(\theta \in (\mu, \frac{q+\mu}{2})) = |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.

2.2.2 Generalized Normal Distributions $\mathcal{GN}_p(\tilde{\mu}, q, \gamma)$

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 $\mu \in \mathbb{R}$ is a location parameter, $\alpha > 0$ is a scale parameter, and $\beta > 0$ is a shape parameter (Nadarajah 2005). Let $F_{\mu,\alpha,\beta}$ denote the cumulative distribution function of $\mathcal{GN}(\mu, \alpha, \beta)$. An expression for $F_{\mu,\alpha,\beta}$ is provided in the appendix.

Definition 2 *Define $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ as the generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ that satisfies mode$(\theta) = \tilde{\mu}$, $P(\theta < q) = p$, and $P(\theta \in (q, \frac{q+\mu}{2})) = \gamma \cdot |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.*

As with the normal distribution $\mu = \tilde{\mu}$, and α and β are determined to minimize the function $(F_{\mu,\alpha,\beta}(q) - (p))^2 + (F_{\mu,\alpha,\beta}(\frac{q+u}{2})) - (\gamma \cdot |p - \Phi(\frac{\Phi^{-1}(p)}{2})|)^2$.

The flexibility to modify γ can change the distribution to be flatter or more concentrated around the modal value. If $\gamma > 1$ then the distribution will be more concentrated around the modal value, and if $\gamma < 1$ then the distribution will be flatter around the modal value, and $\gamma = 1$ corresponds to the normal distribution.

An example for parameterizing an enthusiastic prior with a $\mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ is demonstrated in Figure 1.

2.2.3 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 $c = \frac{\beta}{2\alpha\Gamma(1/\beta)}(F_{\mu,\alpha,\beta}(\theta_{max}) - F_{\mu,\alpha,\beta}(\theta_{min}))^{-1}$.

Definition 3 *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 mode$(\theta) = \tilde{\mu}$, $P(\theta < q) = p$, and $P(\theta \in (q, \frac{q+\mu}{2})) = \gamma|p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.*

For example, consider creating skeptical and enthusiastic priors for a response probability on domain $[0, 1]$, as $\pi_S(\theta) = \mathcal{GN}_{p=1-\epsilon,\Theta=[0,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 1)$ and $\pi_E(\theta) = \mathcal{GN}_{p=\epsilon,\Theta=[0,1]}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$ respectively.

The generalized normal distribution can be used to parameterize skeptical and enthusiastic priors for trials with multiple unknown quantities of interest. 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: $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$. For example, suppose that θ is the risk difference between response probabilities of the treatment group and the placebo group η. The distribution for the risk difference is $\pi(\theta) \sim \mathcal{GN}_{p,\Theta=[-1,1]}$. The domain of η is $[0, 1-\theta)$ if $\theta \geq 0$ and $(-\theta, 1)$ if $\theta < 0$. Define the prior for η to be $\pi(\eta|\theta) \sim \mathcal{GN}_{p,H=[\max(-\theta,0),\min(1,1+\theta)]}$. This prior specification is demonstrated in Figure 2, and Section 3.2.2 uses this representation of the joint prior.

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 subjects 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.

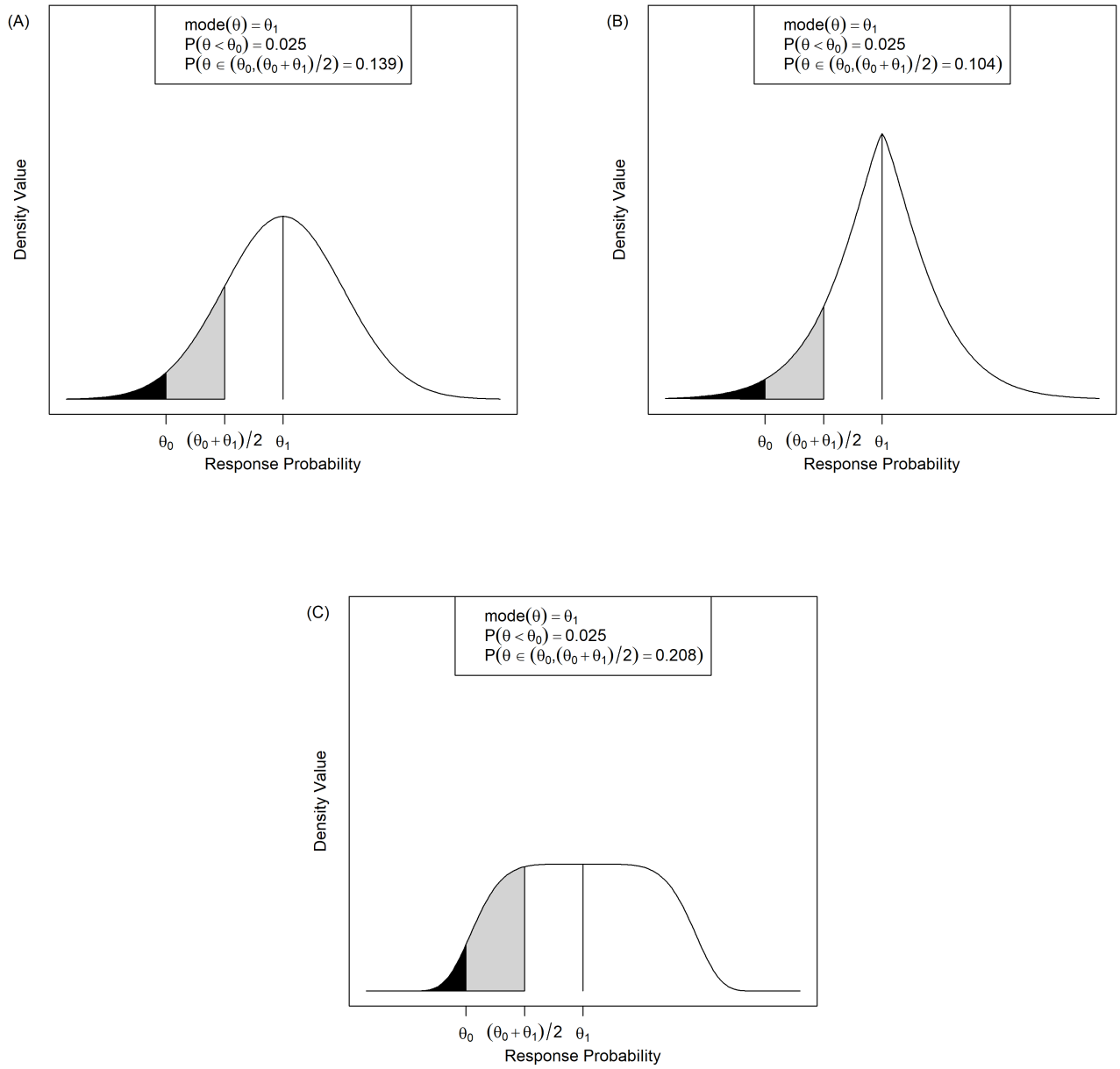


Figure 1: A, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$. B, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 0.75)$, C, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1.5)$

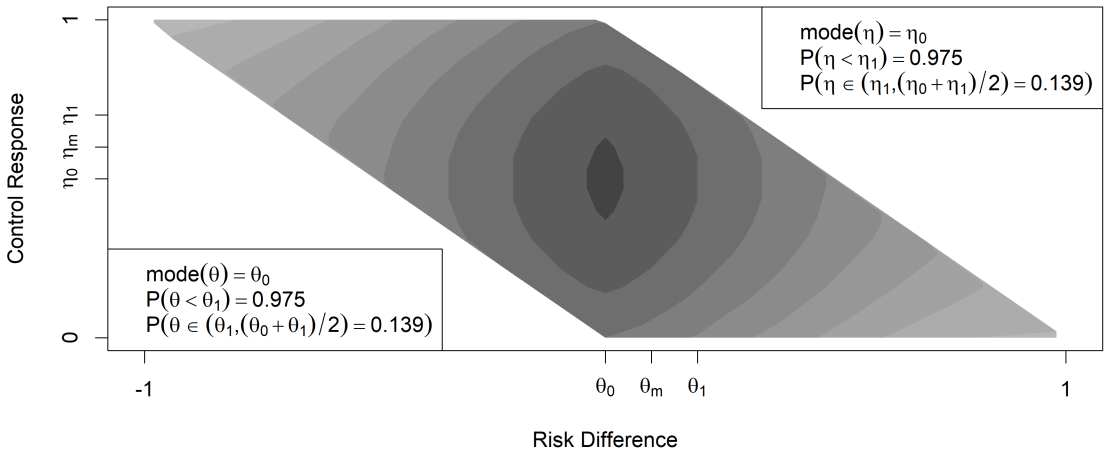


Figure 2: Joint distribution $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$, where $\pi(\theta) \sim \mathcal{GN}_{p=0.975,\Theta=[-1,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 1)$ and $\pi(\eta|\theta) \sim \mathcal{GN}_{p=0.975,H=[\max(-\theta,0),\min(1,1+\theta)]}(\tilde{\mu} = \eta_0, q = \eta_1, \gamma = 1)$.

The initial enrollment occurred on August 25, 2006, and the last assessment was June 24, 2010. If the patient showed improvement at week 8 then they continued treatment and would have their last assessment at week 54. The response rate at 8 weeks was 73.3% ($N = 60$), so it is likely that the last assessment was at week 54, and we can infer enrollment took place over approximately 33.5 months (approximately 1 enrollment per 17 days).

The sample size of 60 patients was chosen to ensure 12% precision in estimating the true response proportion with at 95% CI, assuming a response rate of 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.

3.1.2 Model formulation & prior elicitation

The data $\mathbf{D}$ are assumed to be independent Bernoulli random variables with common response probability θ. The null response value is $\theta_0 = 0.4$, a highly efficacious response probability is $\theta_1 = 0.67$. The hypothesis for this trial is $H_0 :\leq \theta_0$ vs $H_1 : \theta > \theta_0$. The intermediate response value is $\theta_m = 0.535$. The skeptical prior is $\pi_S(\theta) \sim \mathcal{GN}_{p=0.975,\Theta=[0,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 0.75)$. The scaling factor $\gamma < 1$ was chosen to concentrate the distribution around the modal value to provide additional Type 1 error control. The enthusiastic prior is $\pi_E(\theta) \sim \mathcal{GN}_{p=0.025,\Theta=[0,1]}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$. The scaling factor $\gamma = 1$ was chosen as a default value corresponding to the truncated normal distribution. A comparison of the various combinations of the monitoring priors is provided in Appendix 5.2. An inference prior is defined as the mixture (8) with

$\omega = 0.5$. A maximum sample size of $n_{max} = 112$ was chosen based on a frequentist design to have 80% power at a true response proportion of $\theta_m = 0.535$.

3.1.3 Example paths

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

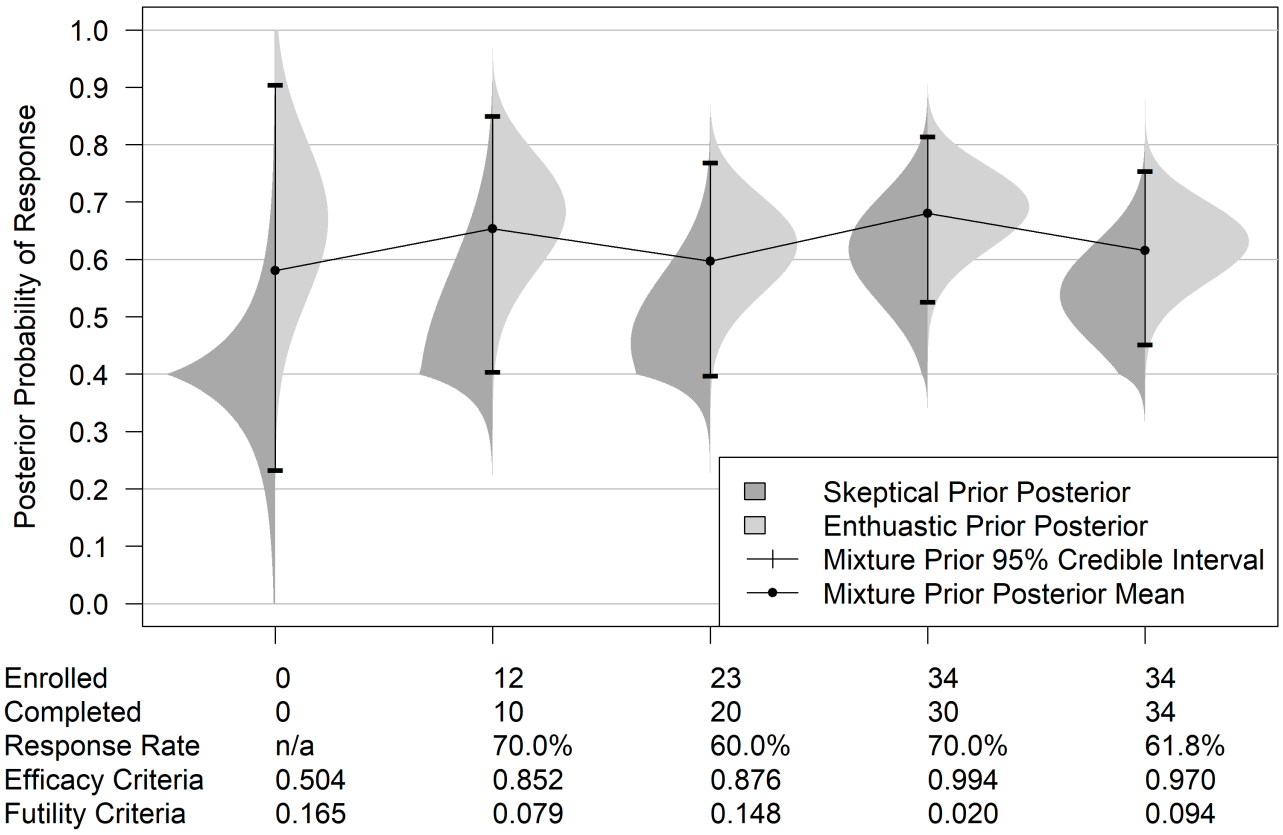
Figure 3(a) shows the results of a trial where at the third interim analysis the efficacy conditions satisfied and enrollment is terminated. Note that in the final analysis the efficacy condition is no longer at the $1 - \epsilon$ threshold. A discussion of this type of "evidence decrease" is given in Section 3.1.4. Figure 3(b) shows the results of a trial where at the third interim analysis the futility condition is satisfied and enrollment is terminated.

3.1.4 Preposterior Analysis of Operating Characteristics

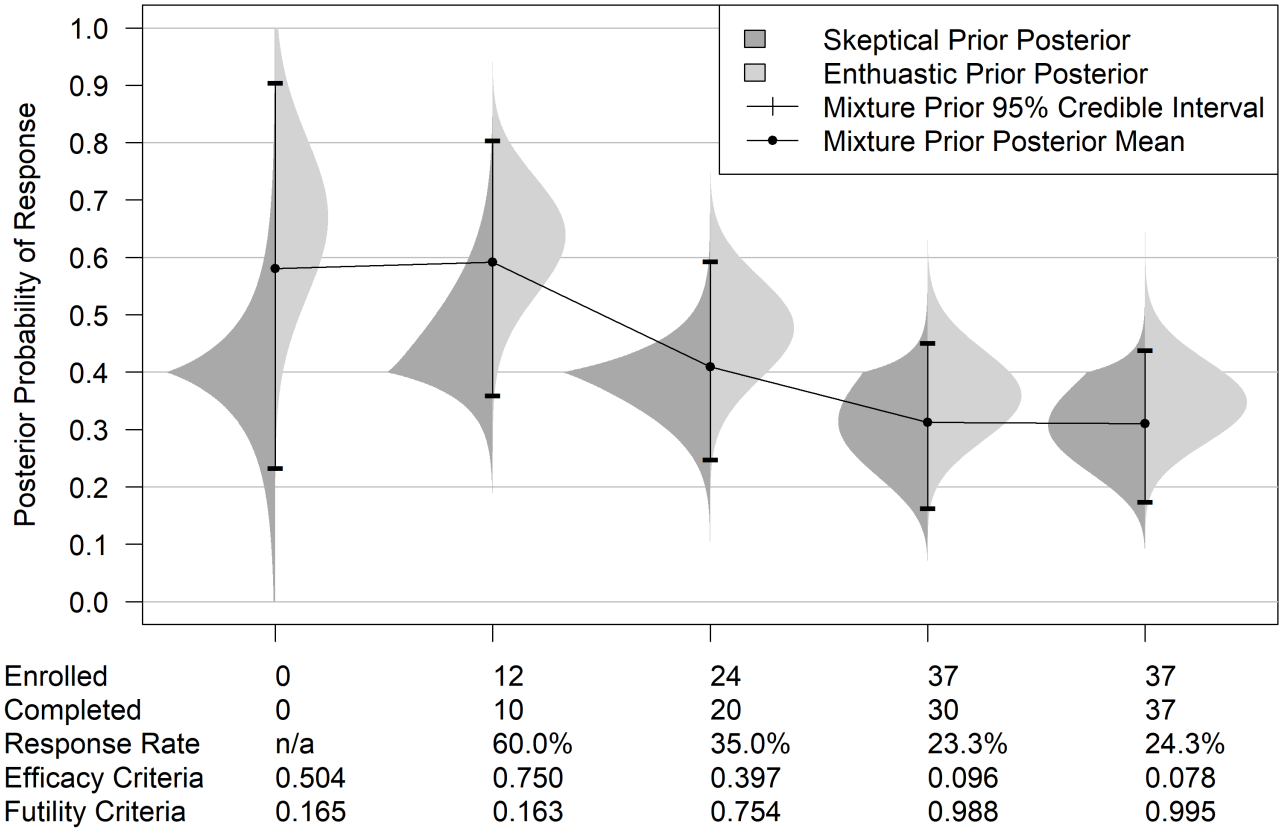
Consider more frequent monitoring with an interim analysis will be completed after every 2 subjects complete follow-up. The following are generated from $10,000$ simulations for the trial described in Section 3.1.2.

Figure 4(a) shows the operating characters of this particular trial design. Inconclusive findings refers to situations where the efficacy and futility criteria are not satisfied for any interim analysis. When the true response probability is $\theta = \theta_0$, there is a 3.9% probability of stopping the trial early for efficacy. The posterior mean shows bias towards the alternative hypothesis.

Figure 4(b) addresses the particular situation of when the efficacy criteria is satisfied at an interim analysis triggering termination of enrollment of additional subjects, but once the outcomes of subjects undergoing follow-up are ascertained the criteria is no longer satisfied. The distribution of the efficacy criteria given the final data for these cases are shown. The probability of these cases occurring is reflected by the percent agreement between interim and final results, and it is shown that as θ increases there is a higher probability of agreement and therefore a lower probability of evidence decrease. The median efficacy criteria is very close to $1 - \epsilon$ and in the vast majority of cases (greater than 10th percentile) the efficacy criteria is still very high. Therefore, for this trial design, the interim and final results are not highly discrepant with respect to the efficacy criteria in these cases.

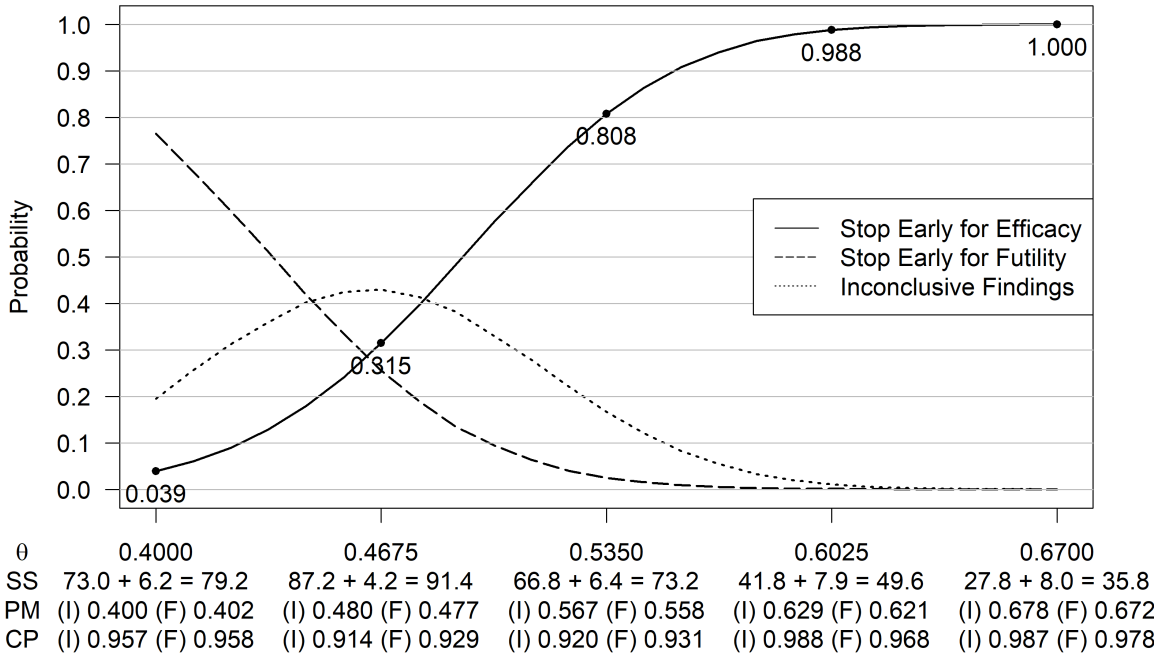


Enrolled	0	12	23	34	34
Completed	0	10	20	30	34
Response Rate	n/a	70.0%	60.0%	70.0%	61.8%
Efficacy Criteria	0.504	0.852	0.876	0.994	0.970
Futility Criteria	0.165	0.079	0.148	0.020	0.094



Enrolled	0	12	24	37	37
Completed	0	10	20	30	37
Response Rate	n/a	60.0%	35.0%	23.3%	24.3%
Efficacy Criteria	0.504	0.750	0.397	0.096	0.078
Futility Criteria	0.165	0.163	0.754	0.988	0.995

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



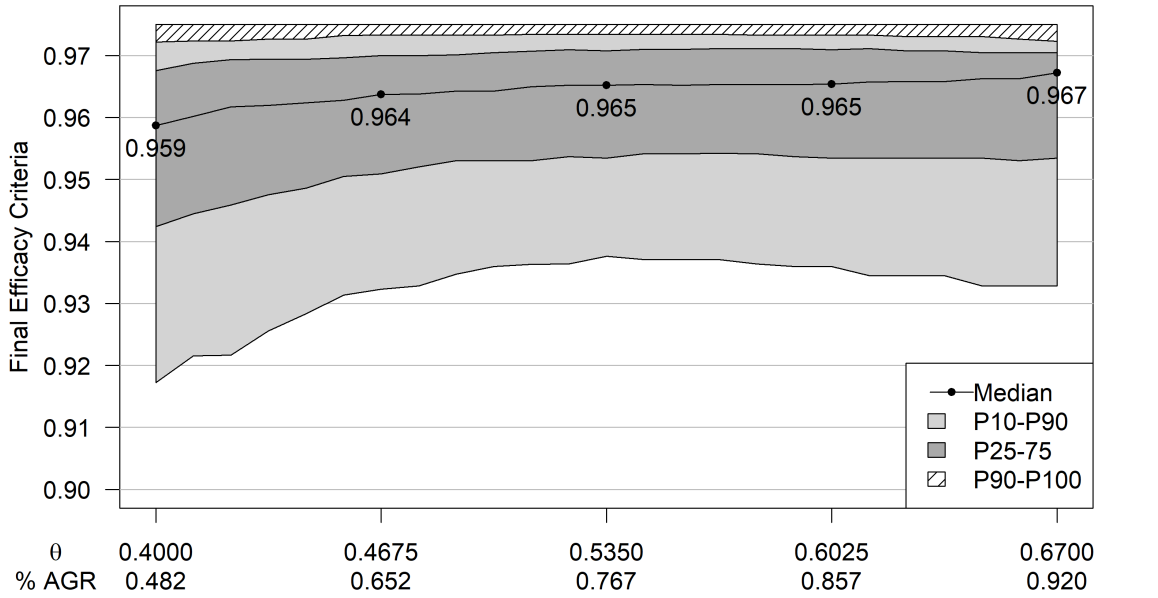


Figure 4: A, Sequential design properties. (SS; sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis) B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; Percent of agreement between final and interim posterior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

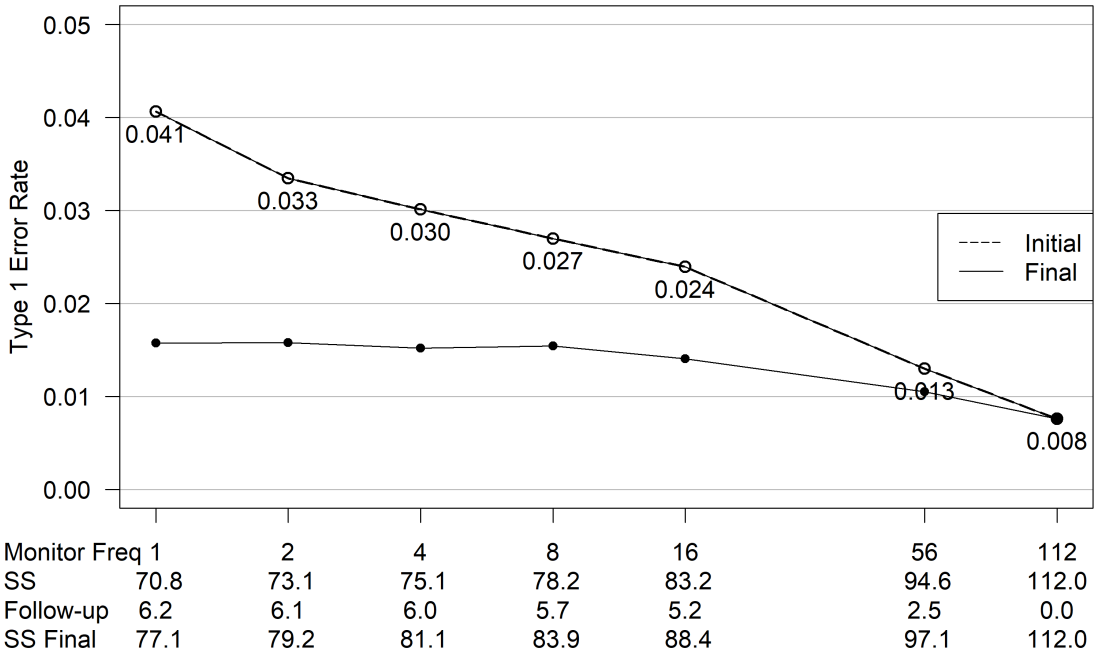


Figure 5: Probability of efficacy criteria being satisfied when $\theta = \theta_0$. SS; sample size. Monitor Freq; monitoring frequency.

3.1.5 Type 1 error rate by the frequency of data monitoring

Figure 5 shows the probability of the efficacy criteria being satisfied at an interim or 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.

The probability of the efficacy criteria being (errantly) satisfied is the highest for a fully sequential design, and decreases as the monitoring frequency decreases. However, the probability of the efficacy criteria being satisfied with the final data is very low for any type of sequential monitoring (under 0.02 in all cases).

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). Subjects were randomized to Belimumab 10mg/kg or placebo, and the primary endpoint was response at week 52. Response 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 17.2 days.

The study design included enrollment of 100 subjects, the first 24 subjects randomized in a 5:1 ratio (Belimumab:placebo) and the remaining 76 subjects would be randomized in a 1:1 allocation ratio. Therefore, 58 subjects 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 subjects enrolled.

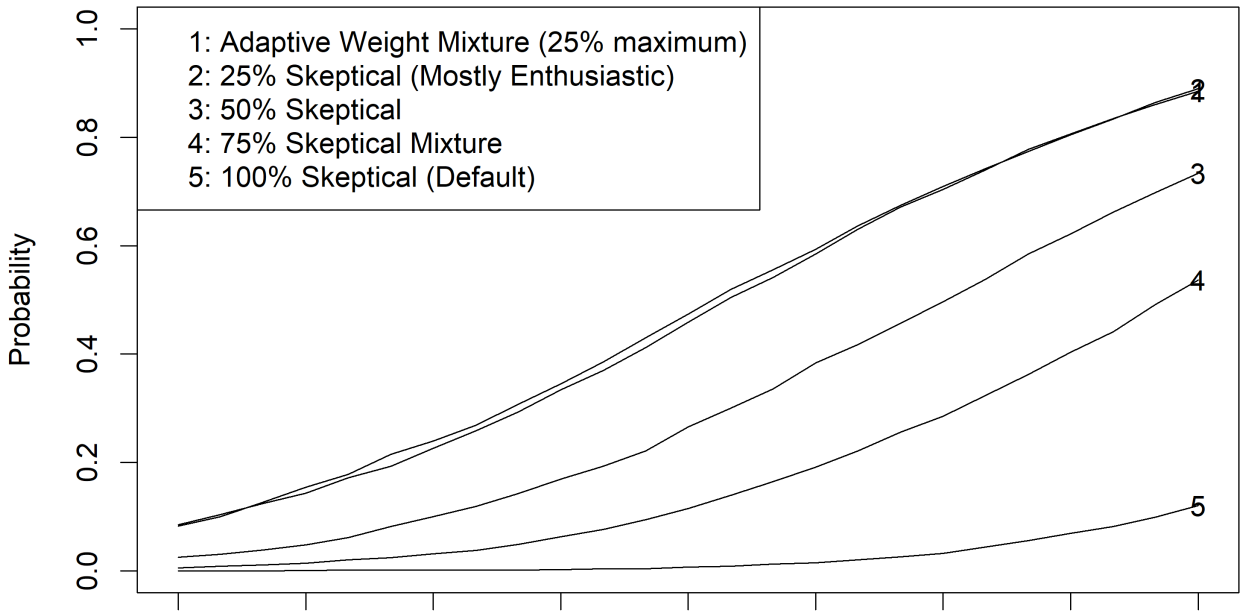
The results of this trial were inconclusive with the 93 subjects. A post-hoc Bayesian analysis that give 55% weight to the adult data was sufficient to provide evidence of positive treatment effect. Our method contrasts a this post-hoc analysis with the prospective use of a monitoring prior for efficacy which gives weight to the adult data.

3.2.2 Model formulation & prior elicitation

The data $\mathbf{D}$ are assumed to be independent Bernoulli random variables with response probability η_0 for the placebo group and η_1 for the treatment (IP for investigational product) group. The null response value for the difference $\theta_0 = 0$ and a highly efficacious difference is $\theta_1 = 0.12$. The hypothesis testing of IP superiority to control $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$. The intermediate response value is $\theta_m = 0.06$. An estimate for the treatment probability from adult trials is $\eta_1 = 0.51$. The skeptical prior is $\pi_S(\theta, \eta_0) = \pi(\theta) \times \pi(\eta_0|\theta)$, where $\pi(\theta) \sim \mathcal{GN}_{p=0.975,\Theta=[-1,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 0.75)$ and $\pi(\eta_0|\theta) \sim \mathcal{GN}_{p=0.975,H=[max(-\theta,0),min(1,1+\theta)]}(\tilde{\mu} = 0.39, q = 0.59, \gamma = 1.5)$. The enthusiastic prior is $\pi_E(\theta, \eta_0) = \pi(\theta) \times \pi(\eta_0|\theta)$, where $\pi(\theta) \sim \mathcal{GN}_{p=0.025,\Theta=[-1,1]}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$ and $\pi(\eta_0|\theta) \sim \mathcal{GN}_{p=0.975,H=[max(-\theta,0),min(1,1+\theta)]}(\tilde{\mu} = 0.39, q = 0.59, \gamma = 1.5)$. 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 subjects have completed outcomes.

3.2.3 Preposterior Analysis of Operating Characteristics

A mixture prior of the form (8) is used for efficacy monitoring where the choice of ω is chosen at the outset to be in the set $\{0.25, 0.5, 0.75, 1\}$. Note that $\omega = 1$ corresponds to the traditional skeptical prior and $\omega = 0.5$ gives equal weight to the skeptical and enthusiastic components. When $\omega = 0.25$ most of the weight to the enthusiastic component. Additionally, a mixture prior with dynamic weight of the form (9)



θ	0.39	0.42	0.45	0.48	0.51	0.54	0.57	0.60	0.63
1	98.4	97.2	95.4	93.1	90.3	87.2	84.0	81.0	78.2
2	98.4	97.0	95.1	92.9	89.9	86.9	83.8	81.0	78.2
3	99.5	99.1	98.3	96.9	95.0	92.5	89.9	86.9	83.9
4	99.9	99.8	99.5	99.0	98.2	96.8	95.0	92.6	89.7
5	100.0	100.0	100.0	100.0	99.9	99.9	99.7	99.3	98.6

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

is used with $\omega_{min} = 0.25$.

Figure 6 shows the probability of stopping early for efficacy based on the choice of ω used in the mixture prior for efficiency monitoring, and the associated sample sizes. Note that only in the case of $\omega = 0.25$ is the probability of stopping the trial early for efficacy near 50% when $\delta = \delta_1$. Also, only in the case of $\omega = 0.25$ is the expected sample size less than 90. Therefore, for this design, the monitoring prior for efficacy has to be mostly enthusiastic in order for reasonable efficacy conditions.

4 Discussion

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 subjects in progress at enrollment termination.

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.

Frequentist group sequential designs use spending functions associated with Type I and Type II error (Pocock 1977, O'Brien & Fleming 1979). While it is possible to calibrate a Bayesian design to have exact pre-specified frequentist properties (Kopp-Schneider et al. 2019), this re-introduces inflexibility (e.g. interim analyses must be pre-specified) that is unnecessary using a fully Bayesian method. Therefore, the objective is to present operating characteristics generally that are "well-calibrated" (Grieve 2016) but not motivated by strict adherence to pre-specified thresholds.

Examples included binomially distributed data and response probabilities as the parameter of interest. The generalized normal distribution with optional truncation can be used for any parameter of interest on an interval domain. Future work includes using the generalized normal distribution for priors in Bayesian clinical trials with survival outcomes.

5 Supplementary material

5.1 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 7 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 subjects in follow-up for monitoring frequencies of 8 or fewer, compared to approximately 6 subjects 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.

5.2 Robustness of parameterizations of monitoring priors

The analyses done in Section 3.1 used a concentrate 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 1.

Figures 8-9 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

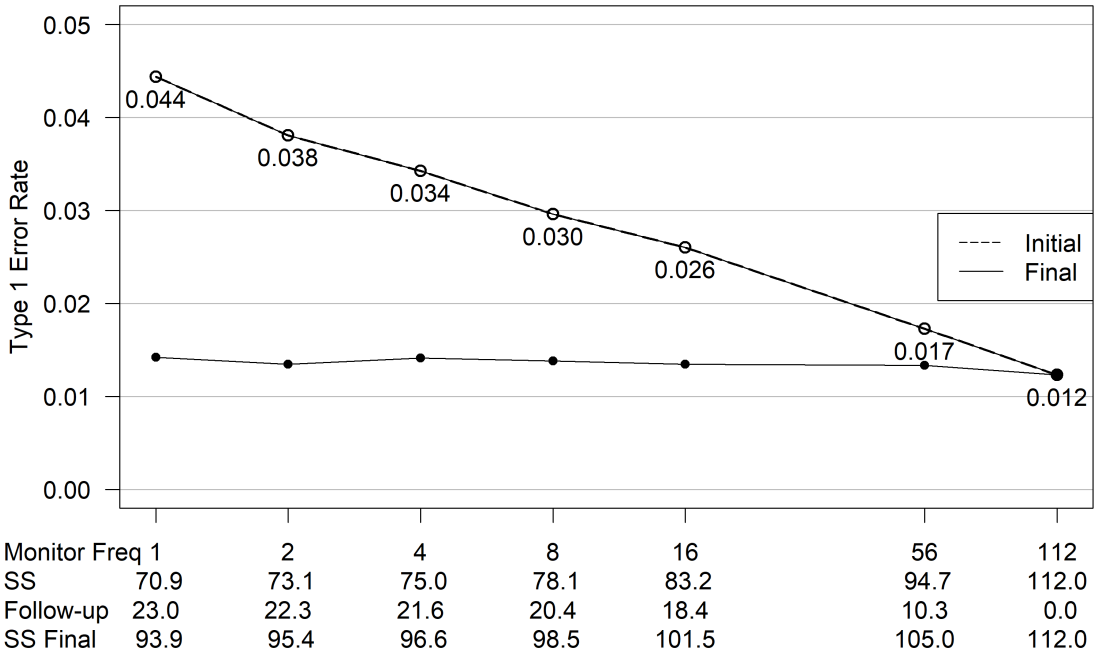
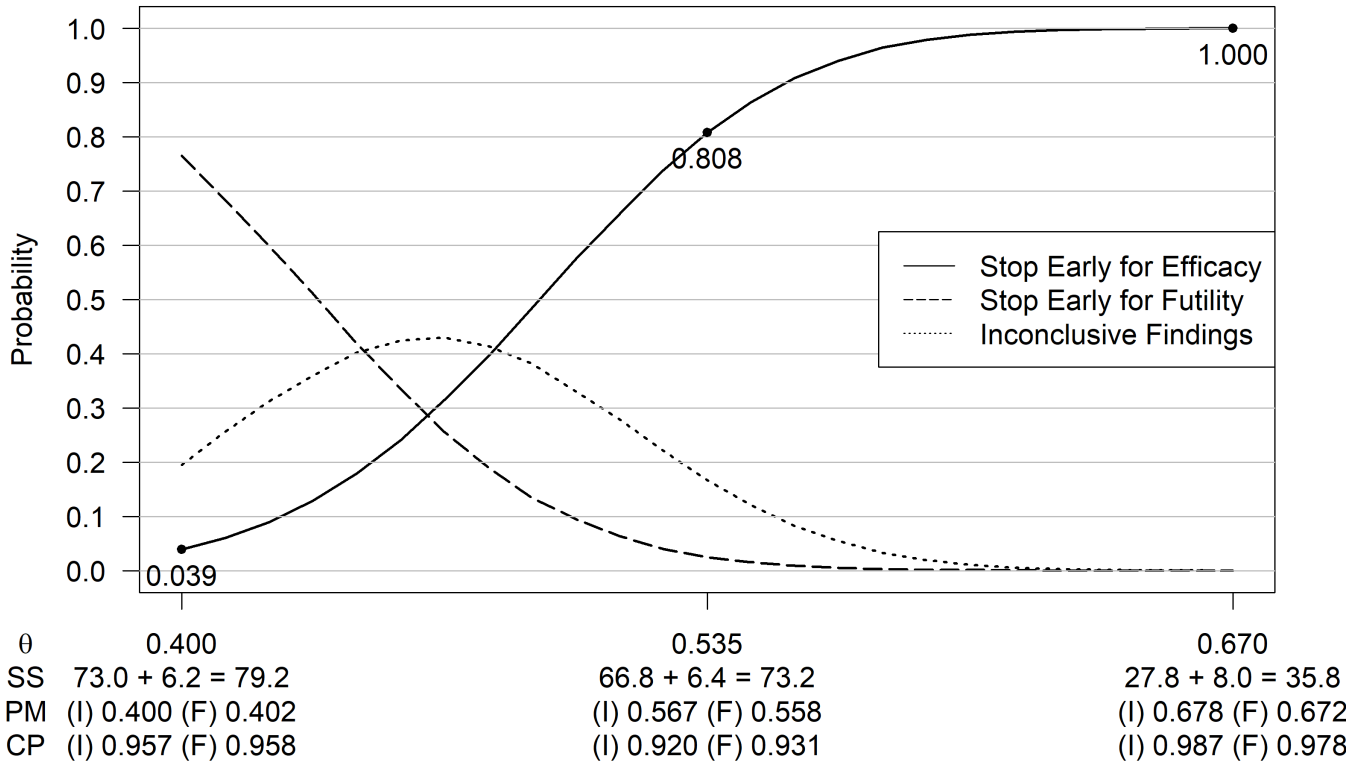


Figure 7: Single arm design from Example 3.1 with a longer follow-up period. Probability of efficacy criteria being satisfied when $\theta = \theta_0$. SS; sample size. Monitor Freq; monitoring frequency.

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 8 and 9, 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.



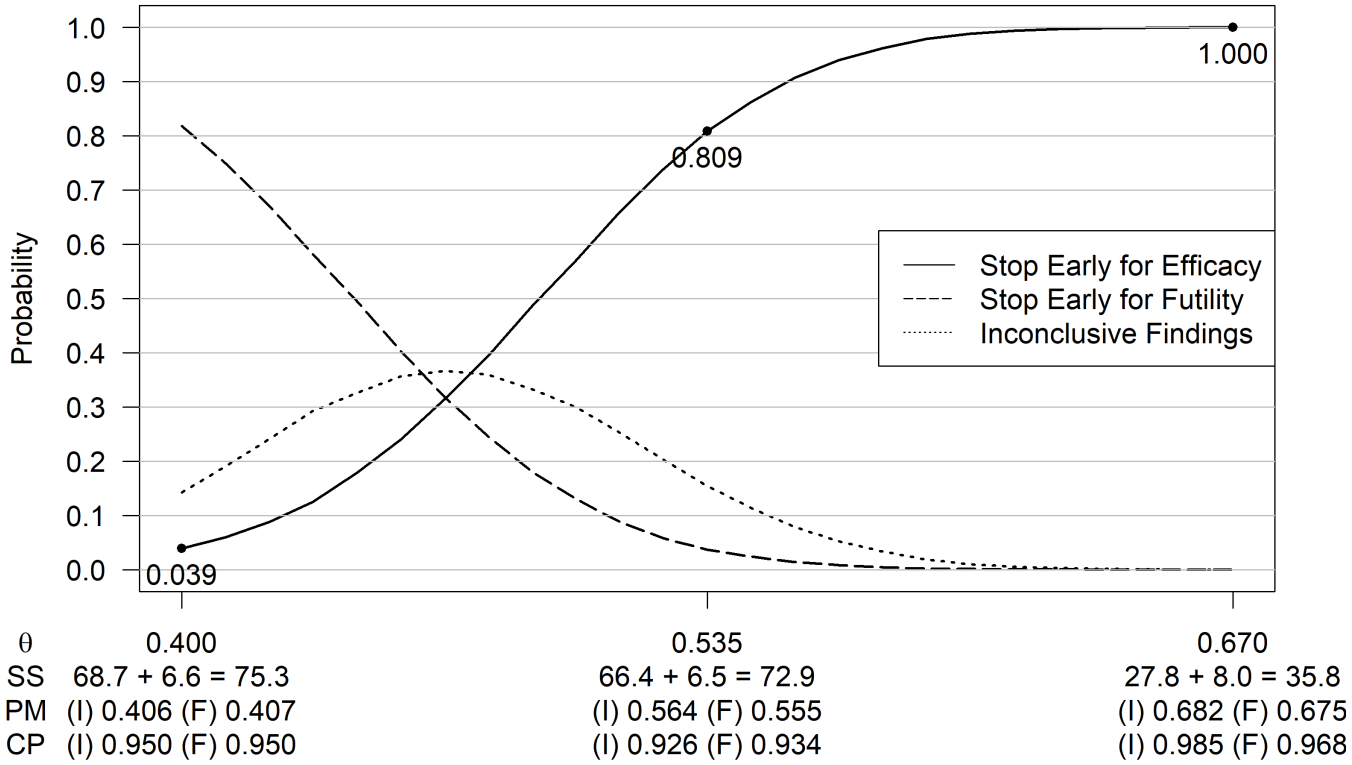
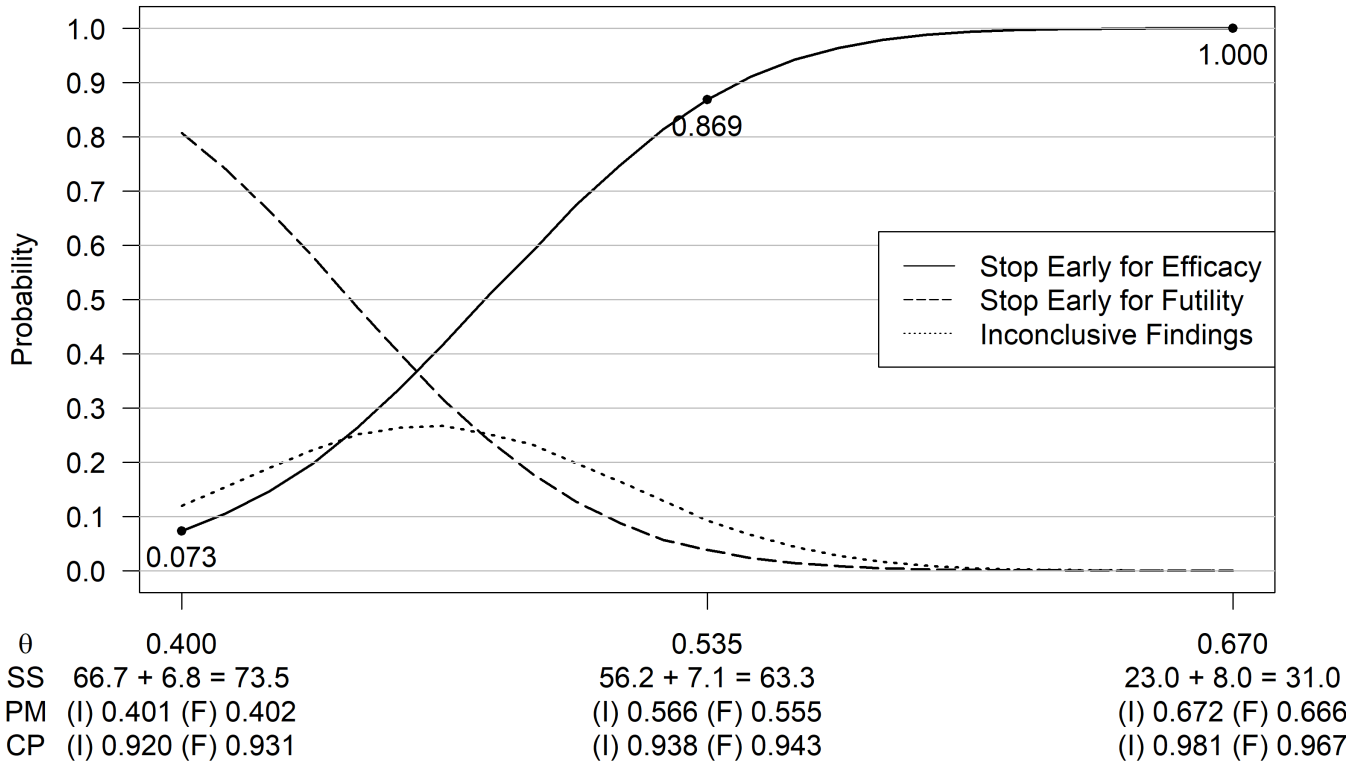


Figure 8: Modification of enthusiastic prior parameterization in Example 3.1. A, default enthusiastic prior (Figure 1(c)). B, flattened enthusiastic prior (Figure 1(d)). Both designs use concentrated skeptical prior (Figure 1(b)).



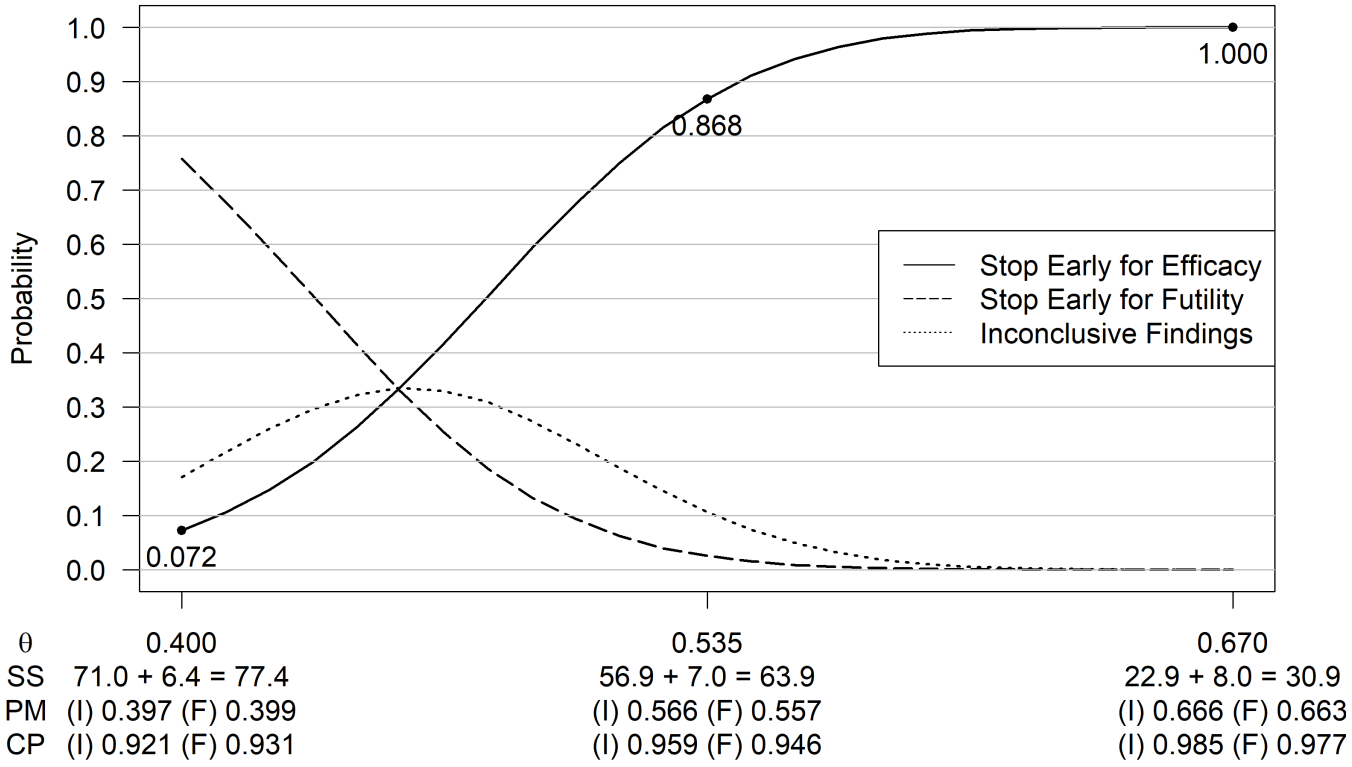


Figure 9: Modification of enthusiastic prior parameterization in Example 3.1. A, default enthusiastic prior (Figure 1(c)). B, flattened enthusiastic prior (Figure 1(d)). Both designs use default skeptical prior (Figure 1(a)).

References

Anscombe, F. J. (1963), 'Sequential medical trials', *Journal of the American Statistical Association* **58**(302), 365–383.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500851*

Barnard, G. A. (1947), 'A review of sequential analysis by abraham wald.', *Journal of the American Statistical Association* (42), 658–669.

Berger, J. O. & Wolpert, R. L. (1988), *The likelihood principle (2nd ed.).*, Hayward (CA): Institute of Mathematical Statistics.

Berry, D. A. (1989), 'Monitoring accumulating data in a clinical trial', *Biometrics* **45**(4), 1197.
URL: *https://www.jstor.org/stable/2531771?origin=crossref*

Berry, D. A. (1993), 'A case for bayesianism in clinical trials', *Statistics in Medicine* **12**(15-16), 1377–1404.

Carlin, B. P., Kadane, J. B. & Gelfand, A. E. (1998), 'Approaches for optimal sequential decision analysis in clinical trials', *Biometrics* **54**(3), 964–975.

Cornfield, J. (1966*a*), 'A bayesian test of some classical hypotheses, with applications to sequential clinical trials', *Journal of the American Statistical Association* **61**(315), 577.

Cornfield, J. (1966*b*), 'Sequential trials, sequential analysis and the likelihood principle', *The American Statistician* **20**(2), 18–23.

Edwards, W., Lindman, H. & Savage, L. J. (1963), 'Bayesian statistical inference for psychological research', *Psychological Review* **70**(3), 193–242.

Fayers, P. M., Ashby, D. & Parmar, M. K. B. (1997), 'Tutorial in biostatistics: Bayesian data monitoring in clinical trials', *Statistics in Medicine* **16**(12), 1413–1430.

Freedman, L. S. & Spiegelhalter, D. J. (1989), 'Comparison of bayesian with group sequential methods for monitoring clinical trials', *Controlled Clinical Trials* **10**(4), 357–367.

Freedman, L. S. & Spiegelhalter, D. J. (1992), 'Application of bayesian statistics to decision making during a clinical trial', *Statistics in Medicine* **11**(1), 23–35.
URL: *http://doi.wiley.com/10.1002/sim.4780110105*

Grieve, A. P. (2016), 'Idle thoughts of a well-calibrated bayesian in clinical drug development', **15**(2), 96–108.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.1736*

Hyams, J., Damaraju, L., Blank, M., Johanns, J., Guzzo, C., Winter, H. S., Kugathasan, S., Cohen, S., Markowitz, J., Escher, J. C., VeeremanWauters, G., Crandall, W., Baldassano, R. & Griffiths, A. (2012), 'Induction and maintenance therapy with infliximab for children with moderate to severe ulcerative colitis', *Clinical Gastroenterology and Hepatology* **10**(4), 391 – 399.e1.

Kopp-Schneider, A., Wiesenfarth, M., Witt, R., Edelmann, D., Witt, O. & Abel, U. (2019), 'Monitoring futility and efficacy in phase ii trials with bayesian posterior distributionsa calibration approach', **61**(3), 488–502.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201700209*

Nadarajah, S. (2005), 'A generalized normal distribution', *Journal of Applied Statistics* **32**(7), 685–694.
URL: *https://doi.org/10.1080/02664760500079464*

O'Brien, P. C. & Fleming, T. R. (1979), 'A multiple testing procedure for clinical trials', *Biometrics* **35**(3), 549–556.
URL: *www.jstor.org/stable/2530245*

Parmar, M. K. & Machin, D. (1993), 'Monitoring clinical trials: experience of, and proposals under consideration by, the cancer therapy committee of the british medical research council', *Statistics in Medicine* **12**(5-6), 497–504.

Pocock, S. J. (1977), 'Group sequential methods in the design and analysis of clinical trials', *Biometrika* **64**(2), 191–199.
URL: *www.jstor.org/stable/2335684*

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. & Colombel, J. F. (2005), 'Infliximab for induction and maintenance therapy for ulcerative colitis', *New England Journal of Medicine* **353**(23), 2462–2476.

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

Spiegelhalter, D. J., Freedman, L. S. & Parmar, M. K. B. (1994), 'Bayesian approaches to randomized trials', *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **157**(3), 357–416.

U.S. Congress (2016), '21st century cures act (pubic law 114-255, 130 stat 1033-1344)'.

U.S. Food and Drug Administration (n.d.), 'Pdufa reauthorization performance goals and procedures fiscal years 2018 through 2022'.