

## A Structured Framework for Bayesian Sequential Monitoring in Clinical Trials

Evan Kwiatkowski<sup>1,\*</sup>, Eugenio Andraca-Carrera<sup>2</sup>, Mat Soukup<sup>2</sup>, and Matthew A. Psioda<sup>1</sup>

<sup>1</sup>Department of Biostatistics, University of North Carolina,

McGavran-Greenberg Hall, CB#7420,

Chapel Hill, North Carolina, USA

<sup>2</sup>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 Bayesian framework for sequential monitoring that makes use of external data that can be consistently applied for a wide range of clinical trial applications. The basis for this framework is the idea that specification of priors used for sequential monitoring and the stopping criteria can be semi-algorithmic byproducts of the trial hypotheses and relevant external data, decreasing the degree of subjectivity in prior elicitation. Monitoring priors are defined using the family of generalized normal distributions which provide a flexible approach for prior specification, naturally allowing one to construct a prior that is more peaked or flat about the most plausible parameter values compared to a normal prior. External data are incorporated into the monitoring process through mixing an a priori skeptical prior with an enthusiastic one using a weight that can be fixed or adaptively estimated. **Cut the following sentence. This is a good example of a sentence that really does not say anything specific and therefore really doesn't add much. These types of sentences almost always lead to reviewer questions regarding what is meant. 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 examples (1) a simple single-arm proof-of-concept trial and (2) a two-group randomized controlled trial. Both examples are modeled after completed pediatric trials and the designs incorporate information from adult trials to varying degrees. Preposterior analysis of each trial design is performed to illustrate that the proposed Bayesian approaches proposed provide good frequentist operating characteristics without having that explicit focus.

### KEY WORDS:

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

## **1. Introduction**

Enrolling the number of patients needed to adequately address objectives of a clinical trial can be difficult. This challenge is especially apparent in cases where the disease for which the investigational product (IP) is an intended treatment is rare or where the focus is on pediatric disease. There is increased interest in innovative ways to perform trials in these settings to facilitate obtaining substantial evidence of treatment benefit as efficiently as possible given the challenge of enrolling patients in a timely fashion. Indeed, this is exemplified by in the draft guidance from the United States Food and Drug Administration (FDA) which outlines procedures for interacting with FDA on complex innovative trial designs (FDA, 2019), abbreviated CID, and the FDA CID Pilot Program which was initiated to fulfill a performance goal set forth under the PDUFA IV legislation. In the draft guidance (see Section II) it is noted that CID are defined as “trial designs that have rarely or never been used to date to provide substantial evidence of effectiveness in new drug applications or biologics license applications”. The potential use of Bayesian methods are discussed throughout the draft guidance and, in particular, the use of Bayesian methods to extrapolate information from adult patients to pediatric settings is given as one type of CID for which there is interest (see Section II, Part C).

In settings where patients are difficult to enroll, and therefore meaningful numbers of patients will complete follow up prior to the trial reaching full enrollment, the concept of frequently monitoring interim data to determine whether a trial (or enrollment) can be stopped becomes appealing. In the author’s opinion, the use of Bayesian sequential methods, rooted in the likelihood principle and thus completely consistent with frequent or even continual data monitoring (provided it is logistically feasible), provide an ideal basis from which to develop CIDs for these settings. Moreover, in settings where pertinent preexisting

data are available, Bayesian methods inarguably provide a natural approach for incorporating that information via a prior distribution into the design and analysis of a future trial.

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

Skeptical and enthusiastic priors are developed using the generalized normal family of distributions. This flexible family includes the normal distribution and provides the capacity to construct monitoring priors that reflect nuanced prior opinion about the treatment effect (Section 2.2.1). A conditional-marginal prior factorization is proposed for settings where there are one or more nuisance parameters (Section 2.2.4) and we illustrate how prior information can be use in both marginal distribution for the treatment effect and conditional distribution for the nuisance parameters, if desired.

We perform preposterior analysis to examine a variety of operating characteristics for the proposed design and to understand how key operating characteristics are influenced by the frequency of monitoring. **General feedback: Compare this text against the previous version. There was a good deal of confusing/inaccurate language – e.g., coverage probability for the treatment effect or probability of early stopping at an interim or final analysis. We need all sentences to be 100% conceptually and mathematically precise.** Specifically, we estimate the probability of stopping early at an interim analysis due to their being substantial evidence of efficacy or for futility, the expected sample size and trial duration, the average posterior

mean, and the coverage probabilities for 95% credible intervals for the treatment effect. In most cases, patients will be ongoing in the trial at the time interim data are obtained that lead to ending enrollment as a result of their being substantial evidence of efficacy. It is our assumption that in many cases these patients will complete the study protocol and, accordingly, we also explore the degree that interim evidence attenuates on average, once final data are available.

Bayesian sequential designs are often restricted to have explicit frequentist properties (Ventz and Trippa, 2015; Zhu and Yu, 2015). Prior work has shown such restrictions can result in Bayesian and frequentist designs that have stopping rules which are nearly identical (Stallard et al., 2020; Kopp-Schneider et al., 2019; Zhu et al., 2019). While it is possible to calibrate a Bayesian design to have specific frequentist operating characteristics, we do not advocate for that strategy. Instead, we propose Bayesian framework that leverages what the authors argue is an intuitive criteria for stopping enrollment and/or data collection at any point (based on posterior inference using a consistent criteria for substantial evidence) without explicit focus on strict type I error control – something that is not achievable when prior information are incorporated into the analysis (Psioda and Ibrahim, 2019). **Please check the accuracy of the following claim.** As will be illustrated, when prior information is not incorporated into the analysis, the designs operating characteristics are “well-calibrated” as discussed by Grieve (2016).

**Please give a more granular breakdown of what is in the paper. In particular, please describe the content in Section 2 in more depth. You can go into what is in the subsections.** 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 given in Section 3, with Section 3.1 presenting an example

based on a single-arm trial and Section 3.2 presenting an example based on a parallel two-group randomized, control trial.

## 2. Methods

### 2.1 Preliminaries

*2.1.1 Bayesian Hypothesis Testing.* Consider a clinical trial application where the primary objective is to test a hypothesis about an unknown quantity of interest which we denote by  $\theta$  with possible values for  $\theta$  falling in the parameter space  $\Theta$ . 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  $\mathbf{D}$  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,  $\mathbf{D} = \{y_i, z_i : i = 1, \dots, n\}$  where  $y_i$  is an indicator of response for patient  $i$  and  $z_i$  is an indicator for whether patient  $i$  was assigned the 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  $\theta$  and potential nuisance parameters  $\eta$  which could be multi-dimensional. For the two-arm trial example,  $\eta$  might correspond to the response probability for patients receiving the control treatment or some transformation thereof. For ease of exposition, for the remainder of Section 2.1, we will focus on the case where  $\theta$  is the only unknown parameter and ignore  $\eta$ .

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

$\theta \in \Theta_i$  is given by

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

where  $p(\mathbf{D} | \theta)$  is also referred to as the likelihood and denoted by  $\mathcal{L}(\theta | \mathbf{D})$ . All of the appendices need to be moved to Supplemental Online Appendices in the format of the Biometrics paper I sent. If not, they are considered an appendix to the paper and also considered towards the length maximum. Hence, we have to move them to a separate PDF file. We will also refer to  $P(\theta \in \Theta_i | \mathbf{D})$  as the posterior probability of hypothesis  $H_i$ . See Appendix XX of the Supplementary Materials for a brief discussion of the appropriateness of referring to  $P(\theta \in \Theta_i | \mathbf{D})$  as the posterior probability of hypothesis  $H_i$ .

**2.1.2 Formalizing the Statistical Concept of Substantial Evidence.** Consider the one-sided hypotheses  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$  for fixed  $\theta_0$ . Often in Bayesian hypothesis testing, one rejects the null hypothesis when  $P(\theta > \theta_0 | \mathbf{D})$  exceeds a prespecified threshold. Let  $\epsilon$  represent the *residual uncertainty* regarding a claim and define  $1 - \epsilon$  to be the threshold such that posterior probabilities in favor of a claim (e.g., that  $\theta > \theta_0$ ) that exceed  $1 - \epsilon$  are viewed as providing *substantial evidence* that the claim is true. Leveraging common practice, we will use  $\epsilon = 0.025$  for the examples presented herein so that  $1 - \epsilon = 0.975$  is the threshold that determines when evidence of a claim is substantial. 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 and consistently.

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

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

where the subscript  $\pi$  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 on the premature termination of trials (Fayers et al., 1997), 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  $\theta$  relative to the hypotheses considered. 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 structured definitions for skeptical and enthusiastic perspectives that can be used to inform early stopping decisions in favor of efficacy and futility, respectively.

Consider again the hypotheses  $H_0 : \theta \leq \theta_0$  versus  $H_0 : \theta > \theta_0$  where  $\theta_0$  represents a

treatment effect of interest and let  $\theta_1 > \theta_0$  represent a plausible, clinically meaningful effect. Define an enthusiastic prior, denoted as  $\pi_E(\theta)$ , as a prior consistent with  $\theta_1$  being the most likely value of  $\theta$  (i.e., the prior mode) and that reflects the belief of an observer who is *all but convinced* that  $H_1$  is true a priori. Formally, this is defined as the prior  $\pi_E(\theta)$  satisfying (i)  $\arg\max_{\theta} \pi_E(\theta) = \theta_1$  and (ii)  $P_E(\theta > \theta_0) = 1 - \epsilon$ , where the subscript  $E$  indicates that the probability is based on  $\pi_E(\theta)$ . Similarly, define a skeptical prior, denoted as  $\pi_S(\theta)$ , as a prior consistent with  $\theta_0$  being the most likely value of  $\theta$  and that reflects the belief of an observer who is *all but convinced* that  $\theta < \theta_1$  is true a priori. Formally, this is defined as the prior  $\pi_S(\theta)$  satisfying (iii)  $\arg\max_{\theta} \pi_S(\theta) = \theta_0$  and (iv)  $P_S(\theta < \theta_1) = 1 - \epsilon$ . In what follows we refer to (i) and (iii) as *mode value constraints* and (ii) and (iv) as *tail-probability constraints*, respectively.

Note that the proposed development of the skeptical prior does not generally reflect skepticism regarding whether the alternative hypothesis is true. Indeed, assuming a symmetric skeptical prior is elicited (as we propose), the *induced* prior probabilities on the hypotheses satisfy  $p(H_0) = p(H_1)$ . See Appendix XX of the Supplemental Materials for more exposition on this point. **Please make sure the first appendix touches on this point.** Thus, the skeptical prior simply reflects skepticism regarding the possibility of large treatment effects but is otherwise consistent with clinical equipoise regarding the two hypotheses.

Unlike the frequentist approach, whether the totality of evidence in favor of a hypothesis is substantial is influenced by the prior distribution used for analysis. It is natural that one would stop a trial early in favor of efficacy or futility when the evidence in favor of the appropriate claim is compelling to 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 the alternative is true, then any less skeptical observer would also be convinced. Therefore, ceasing enrollment and possibly collection of additional



data in order to assess whether the treatment is beneficial would be a reasonable action from most any rational perspective. Similarly, if at any point data sufficiently convince an observer whose prior belief is in accordance with  $\pi_E(\theta)$  that the effect of interest is significantly less than what was originally believed, then any less enthusiastic observer would be similarly convinced. Therefore, ceasing the collection of data altogether would be a reasonable action from most any rational perspective.

*2.1.4 Maximum Sample Size and Formal Stoppage Criteria.* In this section we formalize stopping criteria for futility and efficacy and give general advice for specifying a maximum sample size for the trial. Although sequentially monitored trials in principle require no fixed sample size, in practice due to resource constraints it will almost always be the case that a maximum sample size exists. We recommend that (resources permitting) the maximum sample size, denoted by  $n_{\max}$ , should be chosen so that there is a high probability that the trial generates 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 substantial 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 reanalyze the data after each new patient's outcome is obtained due to scheduling or other constraints on the adjudication panel. In other scenarios, a patient's outcome may be based on a laboratory parameter's change after a fixed period of time and the rate limiting factor for sequential monitoring will be how quickly samples can be shipped, processed, and

entered into a database for analysis. The strategies presented herein for sequential monitoring are appropriate regardless of how frequently data can be monitored even if the motivation for sequential monitoring is scenarios where frequent monitoring is feasible.

Note that I am revising this section significantly. Predominately this is due to imprecise language and what seems to be unnecessary introduction of additional notation. For example,  $\text{eff}(\mathbf{D})$  is the same thing as  $P_S(\theta > \theta_0|\mathbf{D})$  which has already been defined so there is no need to relabel it. Also, it is not “efficacy criteria”, it is a posterior probability – this is what I mean by an imprecise statement. We need to be careful to avoid these.

Stopping criteria for efficacy are defined from the perspective of a skeptical observer. The skeptic becomes convinced that a treatment is effective if at some point the observed data suggest there is substantial evidence that the alternative hypothesis is true. Formally, the early stopping criteria are met based on data  $\mathbf{D}$  when  $P_S(\theta > \theta_0|\mathbf{D}) > 1 - \epsilon$ . Note that the evidence must *exceed* the threshold for what defines it 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  $P_E(\theta \leq \theta_0|\mathbf{D}) > 1 - \epsilon$ . However, notice that when  $\theta = \theta_0$ ,  $P_E(\theta \leq \theta_0|\mathbf{D})$  approaches 0.5 for large sample sizes. Therefore this potential futility criteria would not be satisfiable unless the observed data were consistent with values of  $\theta$  much less than  $\theta_0$ . For this reason, we consider a different approach. Recalling that  $\theta_1$  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  $P_E(\theta < \theta_m|\mathbf{D}) > 1 - \epsilon$ , the trial may be stopped due to there being substantial evidence that the treatment effect is much less than that hypothesized (i.e.,  $\theta_1$ ).

2.1.5 *Example Paths.* More build up is needed in this section. For example, what is  $\theta_0$  and what is  $\theta_1$ . There is currently no context. If this matches the single arm trial example you need to think about connecting this part of the paper to that example without having duplicate content. You can say a little bit here just to give the reader context and then refer to that section where more in depth explanation is given. Also, at this point the reader has not been introduced to the GNP and so the skeptical prior and enthusiastic prior look quite strange compared to one another. You need to at least point this out in a coherent way and note that more details on the precise specification of skeptical/enthusiastic priors subject to the constraints previously introduced is given in Section 2.2. Also, given that you have not mentioned mixture priors at all at this point, should they be included here? I think only if you come back to this figure (which I do not think you do).

Violin plots are used to illustrate the results of two hypothetical trials. For each trial, the monitoring priors (first set of distributions), the posterior distributions from three interim analyses (middle sets), and from the final analysis (last set) are shown. For these hypothetical trials, interim analyses are conducted after 10 additional outcomes were obtained.

Figure 1(a) shows the results of a trial where at the third interim analysis the early stopping criteria for efficacy is satisfied and enrollment is terminated. Note that the final data in this example no longer meet the criteria for substantial evidence of efficacy. A discussion of this type of evidence attenuation is given in Section 2.3.4. Figure 1(b) shows the results of a trial where at the third interim analysis the futility criteria is met and enrollment is terminated.

[Figure 1 about here.]

## 2.2 Specifying Monitoring Priors

2.2.1 *Generalized Normal Distribution.* The skeptical and enthusiastic monitoring priors defined in Section 2.1.3 have mode value and tail-probability constraints. However, these constraints alone do not uniquely determine the priors. There are infinitely many distributions

which satisfy these conditions. However, the mode and tail constraints do uniquely determine a pair of normal distributions which might serve as a default set of monitoring priors. Using a normal distribution for the monitoring prior is motivated by the Bayesian Central Limit Theorem which states that, under general conditions, that the posterior distribution for  $\theta$  approaches normality as the sample size increases, regardless of the initial choice of prior. Therefore, a normally distributed monitoring prior is consistent with belief derived from a sufficiently large (hypothetical) dataset with maximum likelihood estimate equal to the model value required by the prior.

I don't like the way this sentence reads. Saying "There is the possibility to change local behavior of the monitoring prior distribution around the mode value" makes is sound like a theoretical exercise. Contrast that with "It may be desirable to construct a monitoring prior with different local behavior about the mode than is afforded by the normal distribution. For example, one may wish to reflect greater skepticism regarding larger treatment effects than the normal distribution can accommodate." Always be careful to write the development as though there is a specific reason for doing everything we are doing. Generality for generality sake will result in questions/rejection of the paper. Please revise accordingly. The family of generalized normal distributions, which contains the normal distribution as a special case, is able to accommodate changes in local behavior around the mode while still satisfying the mode value and tail probability constraints. Question: Is it really accurate to use the word local? The changes affect the tail too, just not the area in the tail. I'm just thinking out loud here and questioning whether use of "local" is critical to this description (and even accurate)? I'd prefer not to get into a discussion of how the tail area shape changes. The density for a generalized normal distribution  $\mathcal{GN}(\mu, \alpha, \beta)$  is  $f(\theta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\{-\left(\frac{|\theta-\mu|}{\alpha}\right)^\beta\}$  where  $\mu$  is a location parameter,  $\alpha > 0$  is a scale parameter, and  $\beta > 0$  is a shape parameter (Nadarajah, 2005). Fixing the location parameter to be the mode value and changing the shape and scale

parameters in conjunction can maintain the tail probability constraint while also changing the behavior around the mode value. **Make sure to correct the reference to the appendix once the supplement has been created.** See Appendix 5.2 of the Supplemental Materials for more exposition on this point. For example, one can construct a flattened monitoring that is locally uniform around the mode value or a concentrated monitoring prior that is more peaked around the mode value (relative to the normal distribution). **The following sentence does not make sense to me. What is “local dispersion”? This is not a well-defined concept. Are you really talking about Kurtosis (i.e., peakedness about the mode)? For every term we introduce, we need to make sure it is 100% well-defined or you are guaranteed to get a question by a reviewer.** Flattened and concentrated priors have different levels of local dispersion around the mode 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 one wishes to reflect more uncertainty regarding the likelihood that  $\theta$  is near  $\theta_1$  (relative to what is permitted by the normal distribution), while maintaining the same residual uncertainty that  $\theta < \theta_0$ .** Similarly, choosing a concentrated distribution is appropriate **when one wishes to reflect** a higher degree of certainty that  $\theta$  is near  $\theta_1$ , while maintaining residual uncertainty that  $\theta < \theta_0$ .

[Figure 2 about here.]

**The concept of inflating or deflating a probability distribution is not well-defined. Please use precise terminology. This seems to be another way of saying increasing or decreasing. Also, the statement “will be parameterized by” is followed by a statement of action. That does make sense. Contrast that statement with “We parameterize generalized normal monitoring priors as a function of the probability mass in the interval ...” Please revise accordingly. The flattened and enthusiastic priors will be parameterized by inflating or deflating the**

probability mass in the interval between  $\theta_0$  and  $\frac{\theta_0+\theta_1}{2}$  as compared to a normal distribution.

Note: The area (not the change) is shaded gray and comparison across panels helps one to appreciate the increase or decrease in mass relative to the normal distribution. The following sentence should be cleaned up to be more precise. The inflation/deflation of the area in this interval is shaded gray in Figure 2. Panel B of Figure 2 presents a concentrated enthusiastic prior that satisfies the mode value and tail probability constraints given in Section 2.1.3, and that has 25% less mass in the interval  $[\theta_0, \frac{\theta_0+\theta_1}{2}]$  as compared to the normal distribution satisfying the same mode value and tail probability constraints. Reducing the mass in this interval translates to a more peaked distribution about the mode value compared to the default normal distribution (shown in Panel A). Panel C of Figure 2 presents a flattened enthusiastic prior that that satisfies the mode value and tail probability constraints given in Section 2.1.3, and that has 50% more mass in the interval  $[\theta_0, \frac{\theta_0+\theta_1}{2}]$ . This translates to a distribution that is significantly more flat about the mode value than the default normal distribution. Although we have focused on an enthusiastic prior in our discussion here, the same ideas apply to skeptical priors as well. The approach we have proposed (given a choice for the area in the interval  $[\theta_0, \frac{\theta_0+\theta_1}{2}]$ ), results in a unique flattened or concentrated prior (see technical definition in Section 5.2.2). While there is no *correct* choice for the percentage increase or decrease in the mass allocated to the interval  $[\theta_0, \frac{\theta_0+\theta_1}{2}]$  for either an enthusiastic or skeptical prior, the author's choices of 25% and 50% are reasonable perturbations from that afforded by the normal distribution and will be used henceforth to demonstrate the methodology proposed. Lastly, we note that the default normal, flattened, and concentrated priors can all be truncated while maintaining the mode and tail probability constraints. This will be necessary when the parameter of interest has bounded support (e.g.,  $\theta$  is a response probability).

Note that I am changing the notation to what I think will make it more understandable

but the graphs need to change too. I also struggle with how the definition given forces the distribution to be locally non-informative. It seems like the flattening criteria is what causes that. I recommend adding in the criteria  $\pi_{NI}(\theta_0) \approx \pi_{NI}(\theta_1)$  as part of the definition and revising this section accordingly. Another take on it would be to require  $\pi_{NI}(\theta)$  be approximately equal for all  $\theta \in [\theta_0, \theta_1]$  as a part of the definition. We define a *locally non-informative prior* as a prior that is relatively flat over a wide range of plausible values of  $\theta$  (in particular, flat over the interval  $[\theta_0, \theta_1]$ ). Let  $\delta_\theta = 2(\theta_1 - \theta_0)$  (i.e., twice the difference between  $\theta_1$  and  $\theta_0$ ). Specifically, we consider the locally non-informative prior, denoted by  $\pi_{NI}(\theta)$ , that suggests  $\theta_m = \frac{\theta_0 + \theta_1}{2}$  is the most likely value of  $\theta$  and that reflects the belief of an observer who is all but convinced that  $\theta > \theta_m - \delta_\theta$  a priori and all but convinced that  $\theta < \theta_m + \delta_\theta$ . Formally, this is defined as the prior  $\pi_{NI}(\theta)$  satisfying

$$\operatorname{argmax}_\theta \pi_{NI}(\theta) = \frac{\theta_0 + \theta_1}{2} \text{ and } P_{NI}(\theta > \theta_m - \delta_\theta) = P_{NI}(\theta < \theta_m + \delta_\theta) = 1 - \epsilon. \quad (3)$$

This proposed prior is flattened such that there is 50% more mass in the intervals  $[\theta_m - \delta_\theta, \theta_0]$  and  $[\theta_1, \theta_m + \delta_\theta]$  as compared to a normal distribution having the same mode value and tail probability constraints. 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.** This section needs some type of lead in to help the reader with the big picture. Feel free to revise my edits. As described previously, prior information may be directly used in the construction of the enthusiastic prior (e.g., to choose  $\theta_1$ ). It may also be desirable to incorporate prior information into the monitoring process when making a determination of when to stop enrollment early for efficacy. To facilitate this, we introduce a procedure for modifying the skeptical prior such that, if the enthusiastic prior is congruent with observed data, the degree of skepticism can be adaptively lessened. We propose incorporating prior information into the monitoring process for efficacy, when desirable, through constructing a mixture prior from the skeptical

and enthusiastic priors previously discussed using a mixing weight that is constructed from measures of compatibility between the observed data and the skeptical and enthusiastic priors. We define the *adaptive skeptical monitoring prior* as the mixture distribution

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

where  $\omega \in [0, 1]$  is an adaptively determined mixing weight.

The mixing weight  $\omega$  is determined by an assessment of prior-data conflict, proposed by Box (Box, 1980), derived using the prior predictive distribution of the data which is defined (in our case) using the skeptical and enthusiastic monitoring priors. The prior-predictive distribution for data  $\mathbf{D}$  (also called the marginal likelihood) reflects the probability of observing  $\mathbf{D}$  given the assumed prior distribution for  $\theta$  and is defined formally as

$$p(\mathbf{D}) = \int \mathcal{L}(\theta|\mathbf{D})\pi(\theta)d\theta. \quad (5)$$

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

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

which in the case of discrete data is equal to  $\psi(\mathbf{D}_{\text{obs}}) = \sum_{\mathbf{D}} p(\mathbf{D})1[p(\mathbf{D}) \leq p(\mathbf{D}_{\text{obs}})]$ . Box's p-value can be interpreted as the probability of observing data as or more extreme than  $\mathbf{D}_{\text{obs}}$ , given the predictive distribution. Small values of  $\psi(\mathbf{D}_{\text{obs}})$  indicate a lack of compatibility between the prior and the data. **Note, please fix all the references to use  $\mathbf{D}_{\text{obs}}$ .**

We propose using the skeptical and enthusiastic priors  $\pi_S(\theta)$  and  $\pi_E(\theta)$  to compute the quantities in (5) and (6) to create compatibility measurements  $\psi^{(S)}(\mathbf{D})$  and  $\psi^{(E)}(\mathbf{D})$  which are used to determine the mixing weight in (4). If  $\psi^{(E)}(\mathbf{D}) > \psi^{(S)}(\mathbf{D})$ , then the observed 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.

We propose a conservative choice of mixing weight which gives full weight to the skeptical



prior (i.e.  $\omega = 1$ ) whenever  $\psi^{(S)}(\mathbf{D}) \geq \psi^{(E)}(\mathbf{D})$  is

$$\omega = \begin{cases} 1 & \text{if } \psi^{(S)}(\mathbf{D}) \geq \psi^{(E)}(\mathbf{D}) \\ 1 - [\psi^{(E)}(\mathbf{D}) - \psi^{(S)}(\mathbf{D})] & \text{if } \psi^{(S)}(\mathbf{D}) < \psi^{(E)}(\mathbf{D}) \end{cases} \quad (7)$$

As can be seen, for the proposed approach the weight given to the enthusiastic component is the *excess compatibility* in favor of the enthusiastic prior.

**2.2.3 Inference Priors.** 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  $\theta$ . While analysis with these priors provides a rational perspective from which one can determine whether interim data are sufficient to cease enrolling patients, the belief of most stakeholders will likely fall somewhere between the two perspectives. Thus, when interpreting the final data once in hand, intermediate perspectives should be considered. To that end, we define an inference prior through mixing the two monitoring priors along with (potentially) a locally non-informative prior to provided added robustness, if desired.

Define the inference prior as

$$\pi(\theta) = \omega_S \cdot \pi_S(\theta) + \omega_E \cdot \pi_E(\theta) + \omega_{NI} \cdot \pi_{NI}(\theta) \quad (8)$$

where  $\omega_S + \omega_E + \omega_{NI} = 1$  and each weight is non-negative. One may take  $\omega_{NI} = 0$  and fixed value of  $\omega_S = \omega_E = 1/2$  to obtain a static, *agnostic* inference prior since it gives equal weight to the skeptical and enthusiastic prior. This type of inference prior is used in Section 3.1.

Alternatively, an adaptive inference prior is obtained as follows. One first computes the quantities  $\psi^{(E)}$ ,  $\psi^{(S)}$ , and  $\psi^{(NI)}$  using the priors  $\pi_S(\theta)$ ,  $\pi_E(\theta)$ , and  $\pi_{NI}(\theta)$ , respectively, based on Box's p-value as given in (6). Each of  $\psi^{(E)}$ ,  $\psi^{(S)}$ , and  $\psi^{(NI)}$  characterizes compatibility of the data with the respective prior. Our goal is to create a mixture prior which gives more weight to the skeptical (or enthusiastic) prior in situations where the data are highly

compatible with that prior and that favors the locally non-informative prior if data exhibit low compatibility with both the skeptical and enthusiastic priors. To that end, we propose transforming  $\psi^{(NI)}$  to obtain  $\tilde{\psi}^{(NI)}$  which is defined as

$$\tilde{\psi}^{(NI)} = \begin{cases} \psi^{(NI)} - \max(\psi^{(S)}, \psi^{(E)}) & \text{if } \psi^{(NI)} > \max(\psi^{(S)}, \psi^{(E)}) \\ 0 & \text{if } \psi^{(NI)} \leq \max(\psi^{(S)}, \psi^{(E)}) \end{cases}. \quad (9)$$

We 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 Prior Specification for Nuisance Parameters.** This section confuses me a good deal. In reference to the figure (and hypotheses to be tested), what are  $\eta_0$  and  $\eta_1$ . These things have not be defined. Second, it seems like one would elicit  $\pi(\theta)$  to satisfy tail probability constraints but it is not clear how or why they would do that for  $\pi(\eta|\theta)$  since this is a conditional distribution for nuisance parameters. Third, why would  $\pi(\eta|\theta)$  be a default skeptical prior as stated in the figure caption. That does not make sense as  $\eta$  has nothing to do with the hypotheses being tested. I think this entire section needs to be rewritten and perhaps simulations redone. It is hard for me to tell based on what is written here. In the case of nuisance parameters, marginal-conditional factorization of the joint prior  $\pi(\theta, \eta)$  will be used. Let  $\theta$  be the parameter of interest and  $\eta$  be the nuisance parameters. Consider the following representation of the joint prior for  $\theta$  and  $\eta$ :  $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ . Then priors for both  $\pi(\theta)$  and  $\pi(\eta|\theta)$  will be given that satisfy the mode value and tail probability constraints (??)-(??) (see details in Section 5.2.2). For example, suppose that  $\theta$  is the risk difference between response probabilities of the treatment group and the placebo group, and denote the probability of the placebo group by  $\eta$ . 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

All of the concepts in this Section are basic and need not take up space in the main paper. I would cut this entire section. Still, I give some comments to help with clarity for the future (still, cut this all). There is some useful information in this section but it can be sprinkled in much more briefly when results are presented with notation revised according to the changes I made before. Keep the content of Section 2.3.4. Perhaps that can be a brief subsection of its own. The rest can go keeping on minimal discussion where the associated results are first presented.

2.3.1 *Expected sample size and trial duration.* I do not understand what this paragraph is saying. Neither the expected sample size or expected trial duration are rigorously defined. In the absence of that, there is no point for this subsection. Suppose that the trial has interim analyses based on the number of completed outcomes. Let  $\mathbf{D}(n)$  denote the data after  $n$  completed outcomes are ascertained. Let  $n_{\text{initial}}$  be the first instance where either the efficacy criteria (??) or futility criteria (??) are satisfied and enrollment is henceforth terminated, or  $n_{\text{max}}$  if the efficacy criteria or futility criteria are not satisfied at any point. Let  $n_{\text{final}}$  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 (4) and the data likelihood  $\mathcal{L}(\theta|\mathbf{D})$  are used to evaluate the posterior distribution for  $\theta$  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  $\theta$  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 (??) or futility criteria (??) 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  $\text{eff}(\mathbf{D}(n_{\text{initial}}))$ . 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  $\text{eff}(\mathbf{D}(n_{\text{initial}})) > 1 - \epsilon$  and  $\text{eff}(\mathbf{D}(n_{\text{final}})) \leq 1 - \epsilon$ , 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 (??) and futility criteria (??), as well as any quantity involving the posterior distribution of  $\theta$  requires evaluating integrals of the dimension of  $\theta$  (or the dimension of  $(\theta, \eta)$  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).

**STARTING WITH THIS SECTION**

### 3. Examples

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

3.1.1 *Motivating Example.* Some suggestions: (1) can you say more than just “primary endpoint was response at week 8” – specifically, what type of response – clinical, objective, etc? There is commonly a qualifier. (2) What does 12% precision mean? Can you use more well-defined statistical language? For example does 12% precision mean 0.12 CI width or something like that? Consider the T72 pediatric trial “A Study of the Safety and Efficacy of Infliximab (REMICADE) in Pediatric Patients With Moderately to Severely Active Ulcerative Colitis” (NCT00336492) (Hyams et al., 2012) which was conducted between August 2006 and June 2010. In the trial, a 5mg/kg dose of infliximab was given to all patients at weeks 0, 2, and 6, and the primary endpoint was clinical response based on the Mayo score which was measured at week 8. Patients were enrolled over approximately 33.5 months (approximately 1 patient enrolled per 17 days). The sample size of 60 patients was chosen to ensure 12% precision in estimating the true response proportion based on a 95% confidence interval, assuming a response rate of 0.67 as was observed among adults in the ACT 1 and ACT 2 trials (Rutgeerts et al., 2005) at the same 5mg/kg dose ( $N = 242$ ). Obtaining a 95% confidence interval that excluded 0.40 was used as the criterion for classifying the results as clinically significant. 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 motivating example to demonstrate the proposed framework for sequential monitoring. The data  $\mathbf{D}$  are assumed to be comprised of independent Bernoulli random variables having common response probability  $\theta$ . As mentioned above, the primary hypotheses evaluated in the trial were  $H_0 : \theta \leq 0.4$  and  $H_1 : \theta > 0.4$ . For purposes of monitoring, we took  $\theta_1 = 0.67$  consistent with the ACT 1 and ACT 2 trial data which results in  $\theta_m = 0.535$ . The example presented in this section make use

of a concentrated skeptical prior and a default enthusiastic prior for monitoring. Make sure this is moved to supplemental appendix and appropriately referenced. See my recent paper in biometrics for examples for how to write references to the online supplementary materials. A comparison of design properties based on various combinations of the monitoring priors is given in Supplemental Appendix 5.4. In this example, we consider an inference prior defined as the mixture (4) using the fixed value  $\omega = 0.5$  (i.e., a non-adaptive inference prior).

This does not seem to fit with the subsection. Consider adding a different subsection label. Also, I think the first sentence can be cut and you can just start this subsection by saying, “We considered a maximum sample size  $n_{\max} = 112$  ...” With that said, I do not understand the following “based on a frequentist design to have 80% power to conclude efficacy with the skeptical monitoring prior” Frequentist analysis would not use a prior. What do you mean here? This needs to be 100% precise. A Bayesian design would not choose a maximum sample size based on anything frequentist. Also, the choice of 112 cannot be based on providing a “better setting to demonstrate how the flattening or concentration of the skeptical monitoring prior impacts the operating characteristics”, as this makes the example seem artificial and our point here is to provide a real example application. 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. For sequential monitoring, we consider analyzing the accumulating data after every 2 patients complete follow-up.

3.1.3 *Preposterior Analysis of Operating Characteristics.* I edited this sentence some but is the 10,000 per value of  $\theta$ ? If so, please correct. The estimated operating characteristics

presented in this section based on 10,000 simulated trials using the trial design as described in Section 3.1.2. **note that the panels need to be labeled (A) and (B) in the graph. They are not.** Figure 4(a) shows key operating characteristics for this particular trial design. When the true response probability is  $\theta_0 = 0.4$ , the probability of stopping the trial early for efficacy is equal to 0.039, higher than the nominal rate of 0.025 due to the high frequency data monitoring (see also Section 3.1.4). **You cannot refer to another section (Above) without giving some reason why. (Below) Should you say Bayesian design with one analysis using the skeptical prior as the analysis prior?** 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. **Operating characteristics represent average behaviors – you should not say “then decreases as data is shown...” as this suggests a single dataset phenomenon.** 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  $\theta > \theta_0$ . **These sentence suggests the only thing causing the expected sample size to be what it is is futility monitoring. This seems like an over statement. Consider rephrasing.** The impact of futility monitoring at  $\theta_0 = 0.4$  is an expected sample size of 79.2. **The large expected sample size under the null is going to be the death knell for the paper. I think the issue is a bad approach for futility monitoring. See the email I sent you. We need to try that and then discuss.** 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  $\theta = 0.4675$ . **What is meant by acceptable levels? Please be more precise. Sooner relates to time, not sample size. Please revise.** 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  $\theta$  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  $\theta$  with more than 95% probability when  $\theta$  is near  $\theta_0$  or  $\theta_1$ , and less than 95% of the time when  $\theta$  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  $\theta_0$  and  $\theta_1$  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 completion 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 (7). 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  $\mathbf{D}$  are assumed to be independent Bernoulli random variables with response probability  $\eta_0$  for the placebo group and  $\eta_1$  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 (7). A 3-part mixture inference prior of the form (8) will be used to estimate the posterior mean and coverage probabilities for  $\theta$ . 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 mode 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  $\theta$  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 (4) for efficacy monitoring with a fixed choice of  $\omega$  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  $\theta$ . Figure 7(B) shows the 3-part mixture inference prior weights  $\omega_S$ ,  $\omega_E$ ,  $\omega_{NI}$  using (8)-(9) 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 mode value,

a quantile condition, and an additional parameter which can concentrate or flatten the distribution around the mode 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 (??) and (??). In practice, the value of  $\epsilon$  could be different for these two purposes (e.g.  $\epsilon_{\text{eff}} = 0.025$ ,  $\epsilon_{\text{fut}} = 0.05$ ). The intermediate effect size  $\theta_m$  used in futility monitoring (??) was chosen to be  $\theta_m = (\theta_0 + \theta_1)/2$ , but another effect size between  $\theta_0$  and  $\theta_1$  could be considered. Different assessments of prior-data conflict could be used to create a dynamic prior (4). Other distributions than the generalized normal could be used to create monitoring priors with the desired mode 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.

#### ACKNOWLEDGEMENTS

#### 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.
- FDA (2019). Interacting with the fda on complex innovative trial designs for drugs and biological products.
- 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.
- Psioda, M. A. and Ibrahim, J. G. (2019). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics* **20**, 400–415.
- 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  $\Theta_0 \cup \Theta_1 = \Theta$  and  $\Theta_0 \cap \Theta_1 = \emptyset$ . 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  $(\theta, \eta)$  specified over the parameter space defined with respect to each of the hypotheses (e.g.  $\pi(\theta, \eta | H_i)$  for  $i = 0, 1$ ).

The posterior probability of hypothesis  $H_i$  is given by

$$p(H_i|\mathbf{D}) = \frac{p(\mathbf{D}|H_i) \cdot p(H_i)}{p(\mathbf{D}|H_0) \cdot p(H_0) + p(\mathbf{D}|H_1) \cdot p(H_1)}, \quad (10)$$

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  $\theta$  overall and compute the posterior distribution. The posterior probability that  $\theta \in \Theta_i$  is given by

$$P(\theta \in \Theta_i|\mathbf{D}) = \frac{\int_{\Theta_i} p(\mathbf{D}|\theta)\pi(\theta|\theta \in \Theta_i)d\theta \cdot P(\theta \in \Theta_i)}{\sum_{j=0,1} \int_{\Theta_j} p(\mathbf{D}|\theta)\pi(\theta|\theta \in \Theta_j)d\theta \cdot P(\theta \in \Theta_j)} \quad (11)$$

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  $\pi(\theta|H_i) = \pi(\theta|\theta \in \Theta_i)$  for  $i = 0, 1$ . If in fact  $\pi(\theta)$  does represent belief about  $\theta$ , 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.

## 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  $P(\theta \leq q) = p$ . The values for the mean and standard deviation are  $\mu = \tilde{\mu}$  and  $\sigma = \frac{q-\tilde{\mu}}{\Phi^{-1}(p)}$ , where  $\Phi$  denotes the cumulative distribution function for a standard normal distribution and  $\Phi^{-1}$  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  $(\mu, q, p)$  that satisfy  $\frac{q-\mu}{p-0.5} > 0$ . Since the normal distribution is completely specified by  $(\mu, \sigma)$ , quantities such as  $P(\theta \leq \tilde{q})$  are also specified for any  $\tilde{q} \in \mathbb{R}$ . In particular, if  $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$  then  $P(\theta \leq \frac{q+\mu}{2}) = \Phi(\frac{\Phi^{-1}(p)}{2})$ . 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)

$$P(\theta \leq q | \mu, \alpha, \beta) = \frac{1}{2} + \frac{\text{sign}(q - \mu)}{2} \int_0^{|q - \mu|^\beta} \frac{w^{1/\beta - 1}}{\alpha \Gamma(1/\beta)} \exp \left\{ - \left( \frac{1}{\alpha} \right)^\beta w \right\} dw \quad (12)$$

Define  $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$  as the generalized normal distribution  $\mathcal{GN}(\mu, \alpha, \beta)$  that satisfies  $\text{mode}(\theta) = \tilde{\mu}$ ,  $P(\theta \leq q) = p$ , 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  $\alpha$  and  $\beta$  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  $\text{mode}(\theta) = \tilde{\mu}$ ,  $P(\theta \leq q) = p$ , 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.

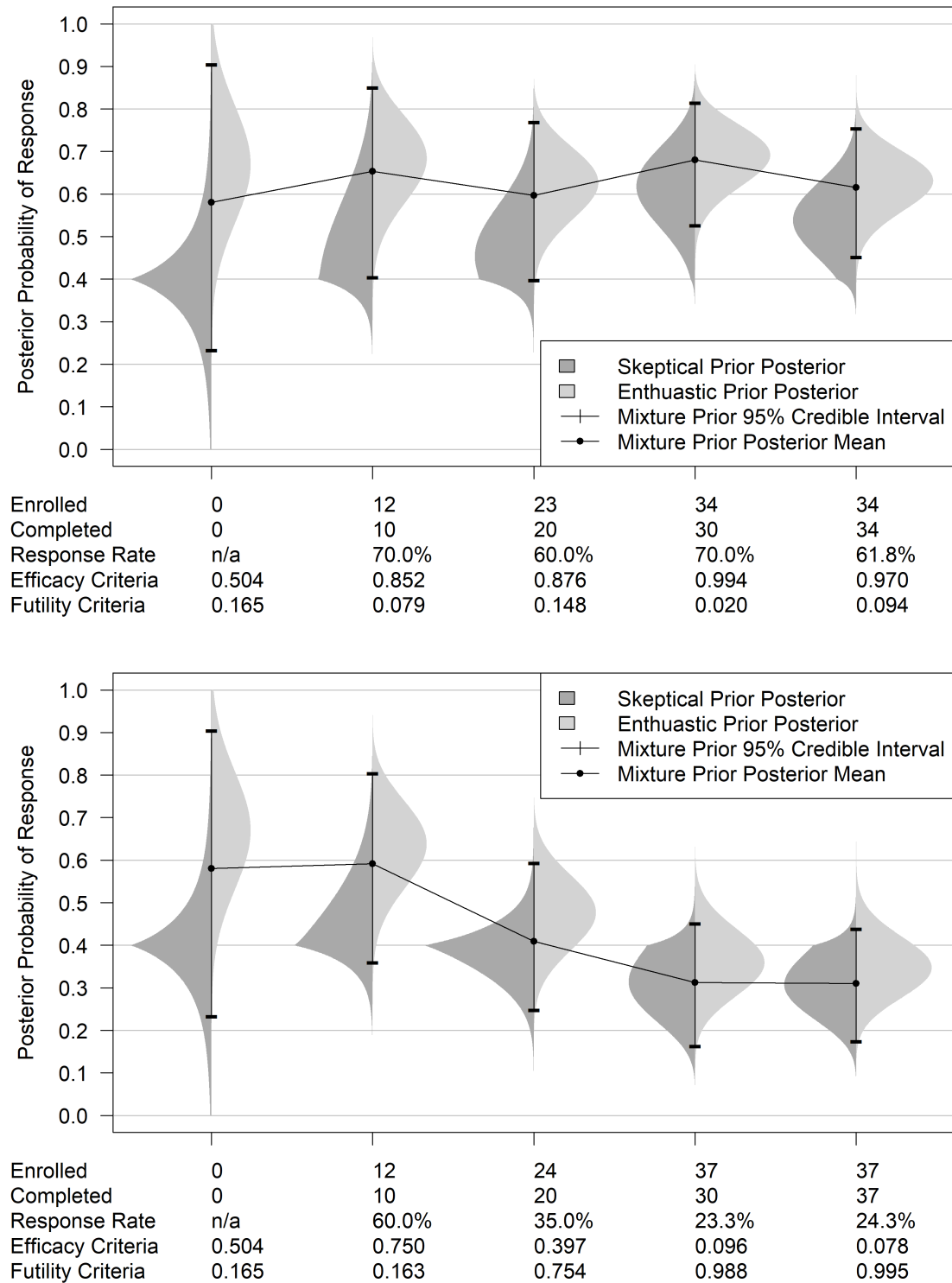
Figures 9-10 shows what happens when the enthusiastic prior shifts from default to flattened, with the skeptical prior remaining fixed. Note that in the region between  $\theta_0$  and  $\frac{\theta_0 + \theta_1}{2}$  as the enthusiastic prior shifts from default to flattened, (a) the probability of 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  $\theta$ . 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  $\theta_0$  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.

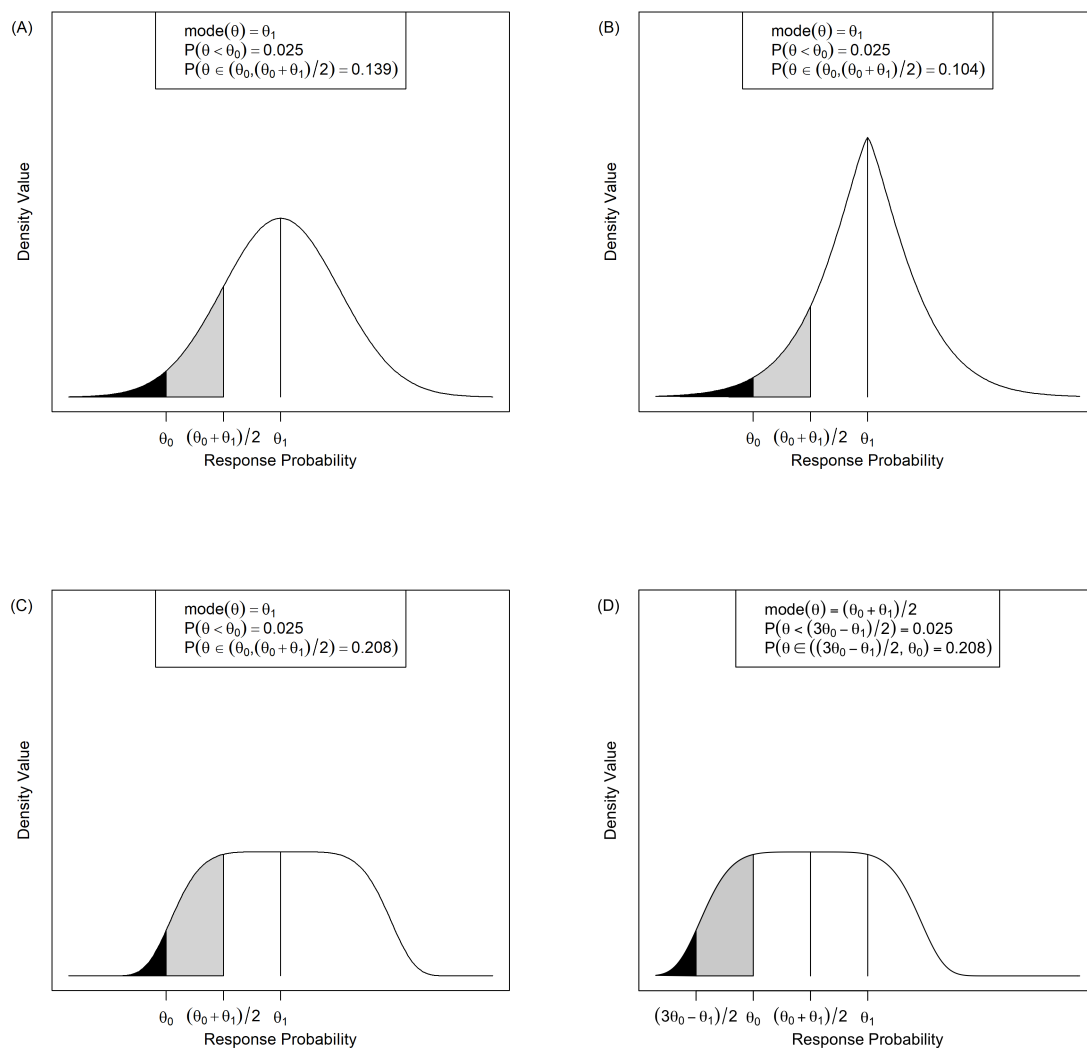
The choice of skeptical and enthusiastic prior affects the analysis, and their specification (e.g. default, skeptical, enthusiastic) should be made with these properties in mind.

[Figure 9 about here.]

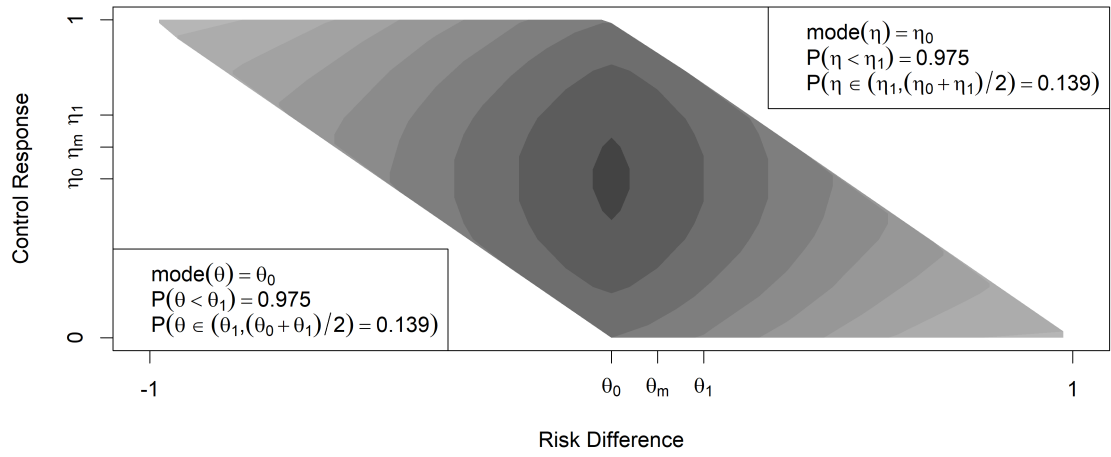
[Figure 10 about here.]



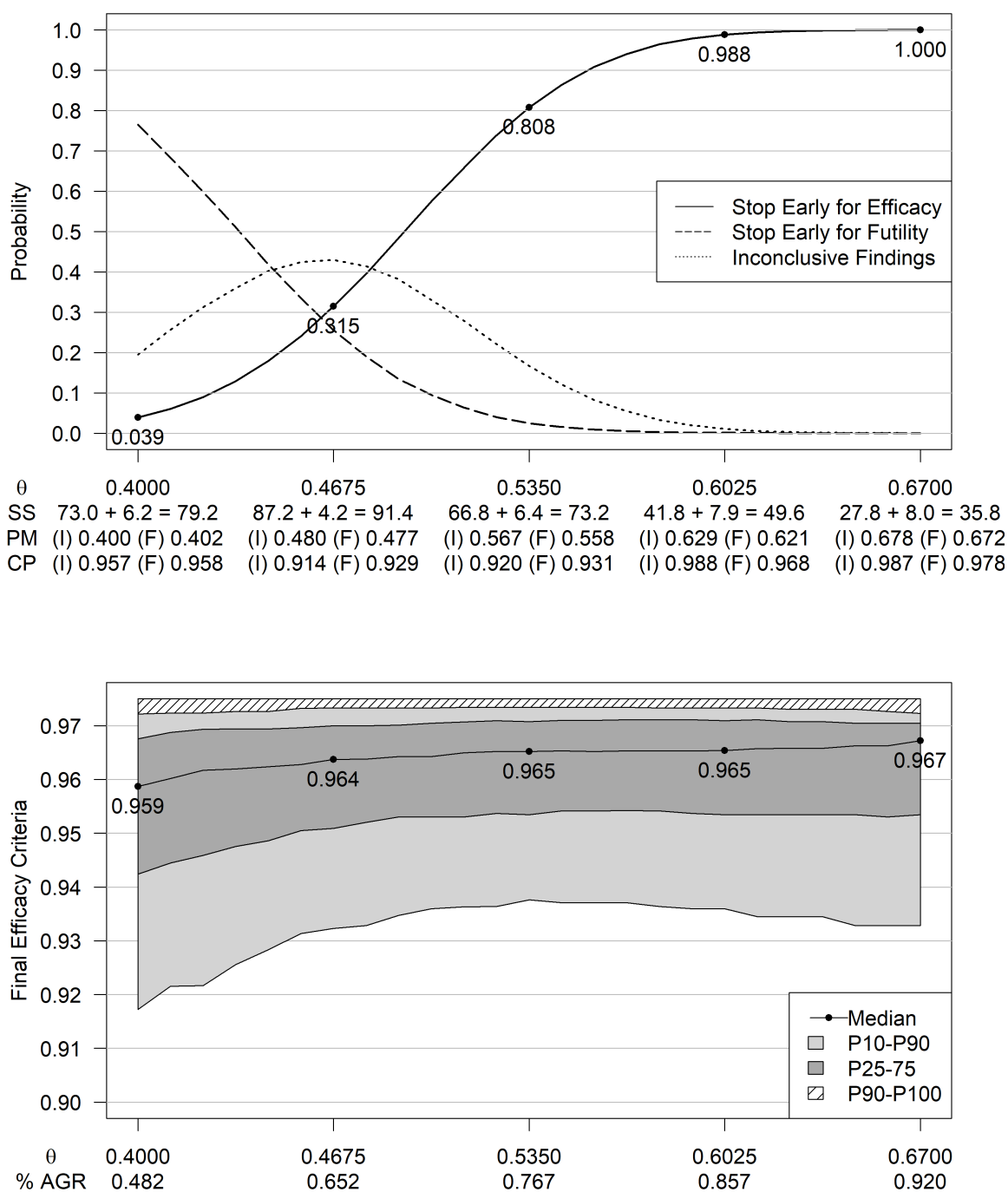
**Figure 1.** Example paths for the trial described in Section 3.1.2. A, Early stopping for efficacy. B, early stopping for futility.



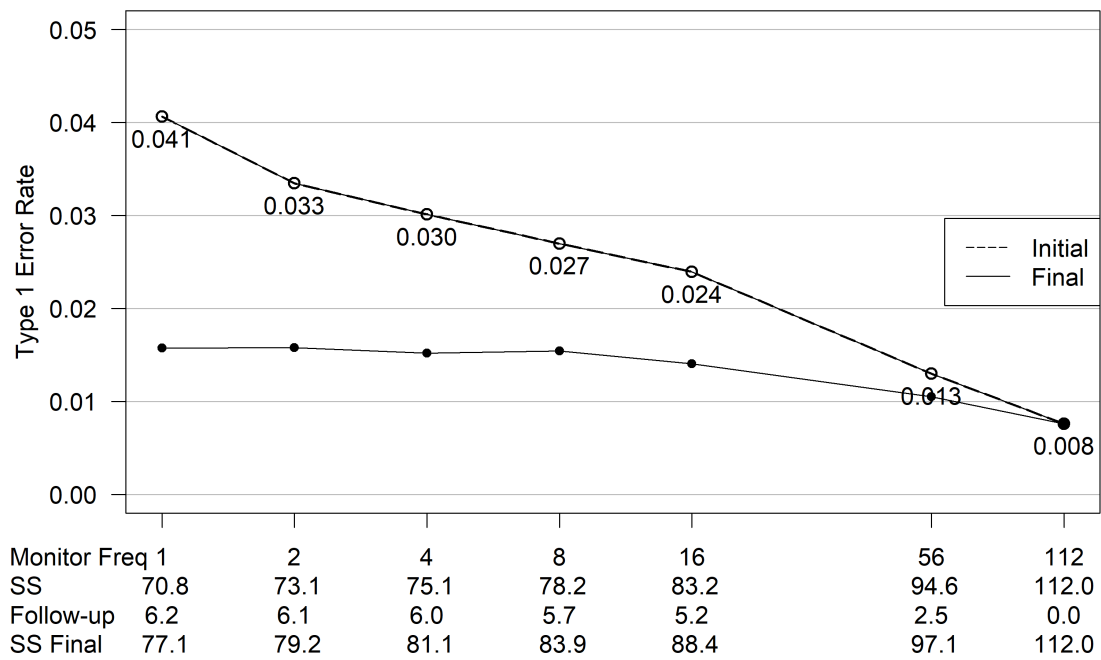
**Figure 2.** A, Default enthusiastic prior. B, Concentrated enthusiastic prior. C, Flattened enthusiastic prior. D, Locally non-informative prior



**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  $[\max(-\theta, 0), \min(1, 1 + \theta)]$ .

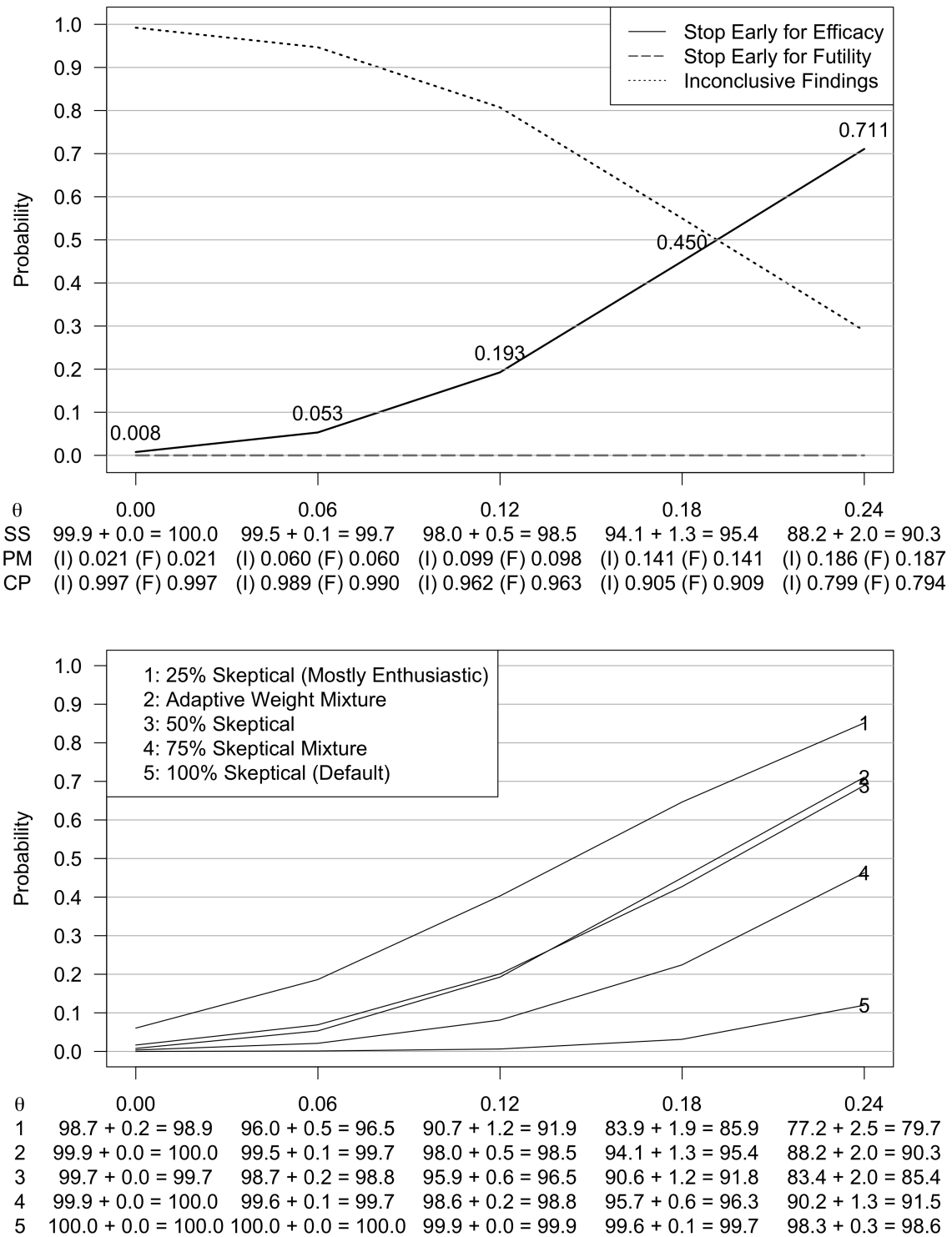


**Figure 4.** A, Sequential design properties. (SS; sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis). B, Distribution of final posterior probability given interim stoppage and evidence decrease (% AGR; Percent of agreement between final and interim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

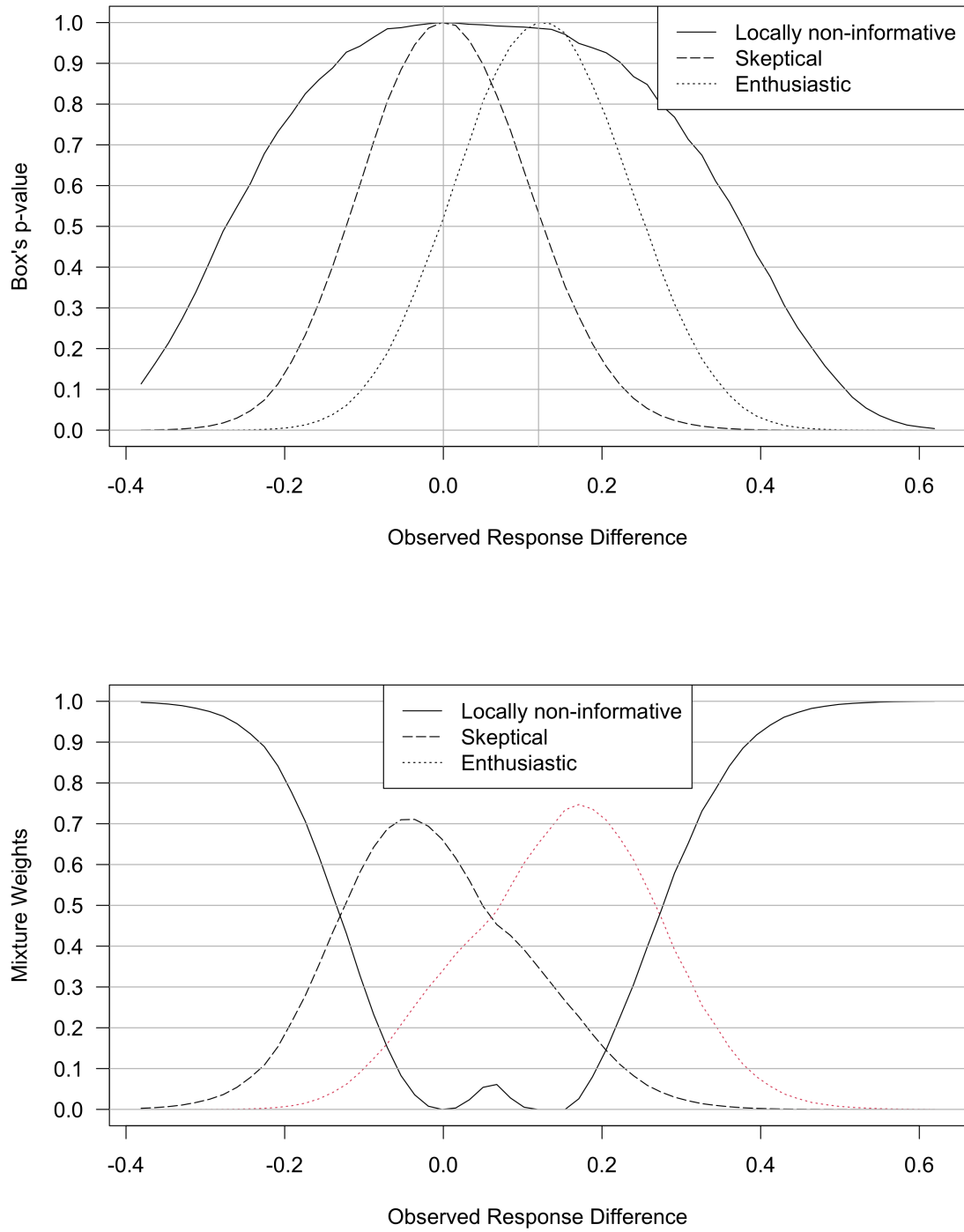


**Figure 5.** Probability of efficacy criteria being satisfied when  $\theta = \theta_0$ . SS; sample size. Monitor Freq; monitoring frequency.

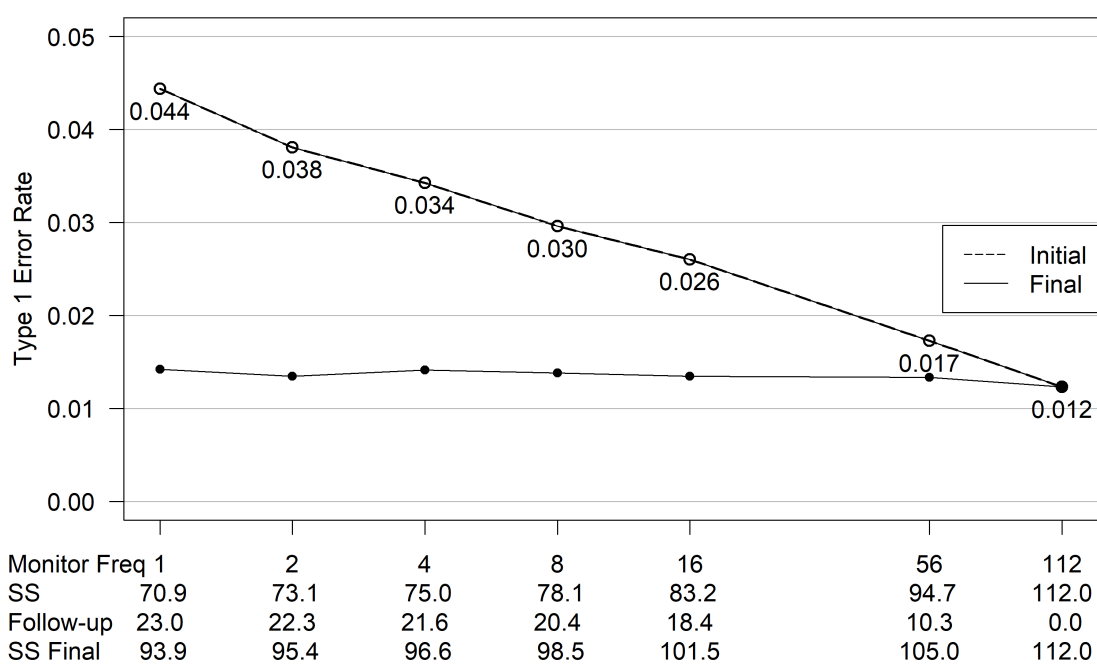




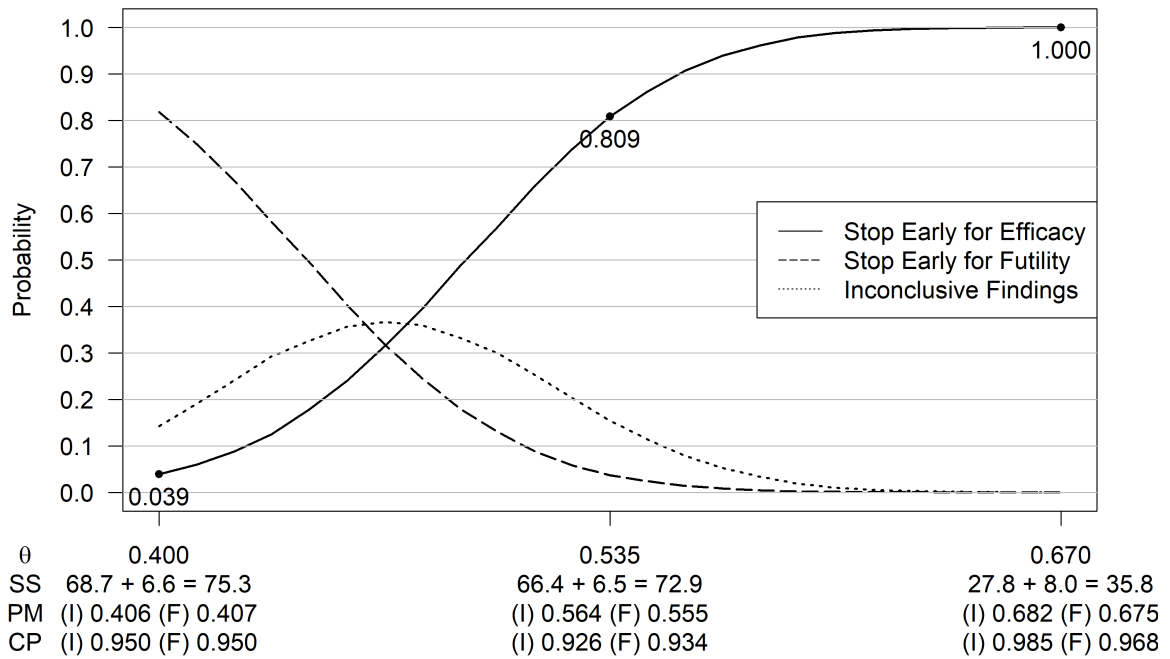
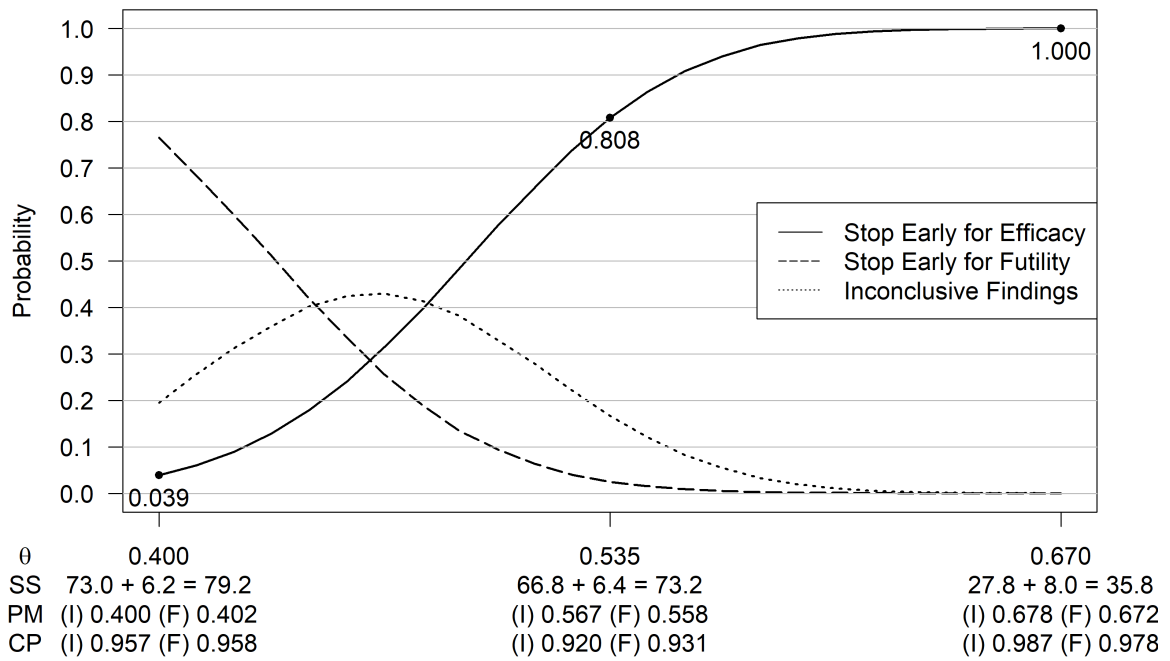
**Figure 6.** A, Sequential design properties using adaptive weight monitoring prior. (SS; sample size, PM; posterior mean, CP; coverage probability, (I); interim analysis, (F); final analysis). B, Probability of stopping for efficacy and associated sample sizes by true IP response rate  $\theta$  with different choices of efficacy monitoring prior.



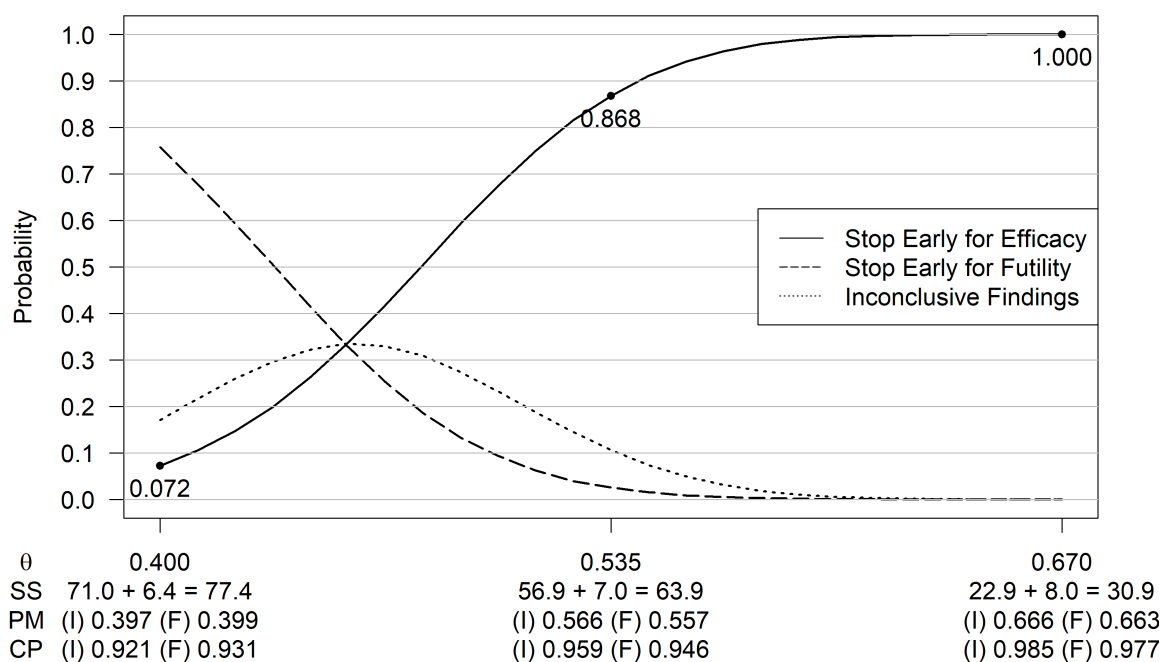
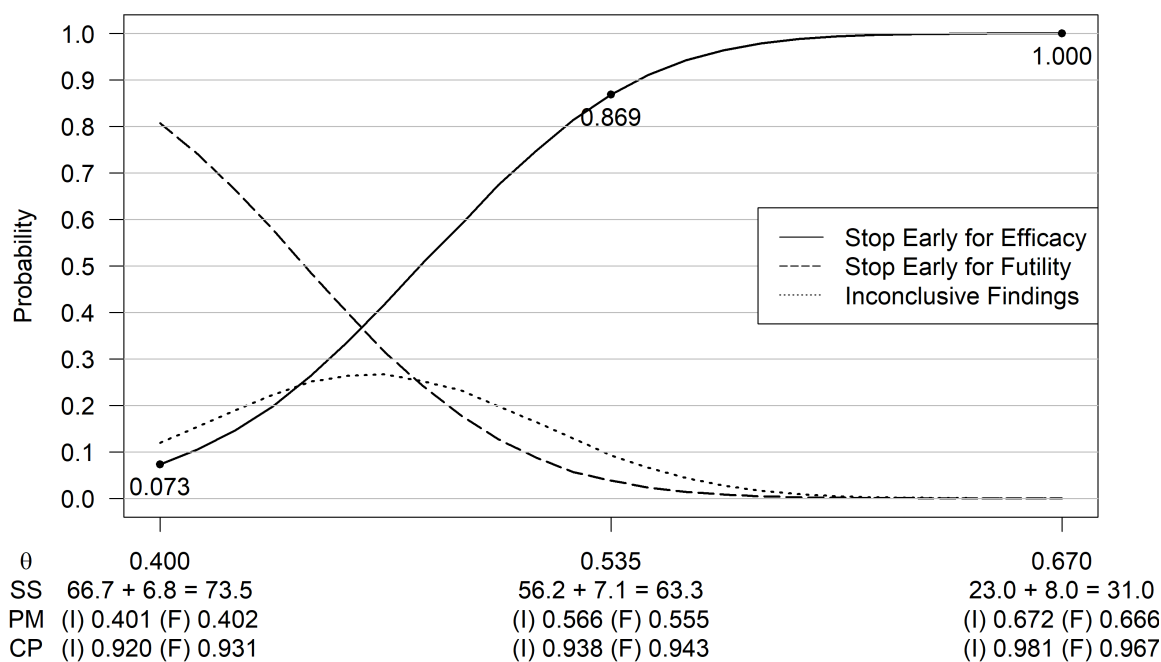
**Figure 7.** A; Prior-data compatibility assessments  $\psi^{(S)}$ ,  $\psi^{(E)}$ ,  $\psi^{(NI)}$  by observed risk difference. B; 3-part mixture inference prior weights  $\omega_S$ ,  $\omega_E$ ,  $\omega_{NI}$  by observed risk difference.



**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)).