# A Structured Framework for Bayesian Sequential Monitoring in Clinical Trials

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

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

McGavran-Greenberg Hall, CB#7420,

Chapel Hill, North Carolina, USA

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

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

Silver Spring, Maryland, USA

*email: ekwiatkowski@unc.edu

SUMMARY:

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

KEY WORDS:

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

## 1. Introduction

In the clinical development of a therapeutic product, it is often a challenge to obtain sufficient outcomes for the primary objective (e.g. difficulty in enrolling, limited patient populations). Bayesian sequential methods are ideal to address this uncertainty in outcome ascertainment by allowing the possibility for conclusions to be made earlier without inherent restrictions for frequent or even continual data monitoring. The advantages of sequential monitoring are especially apparent when there are additional difficulties in obtaining outcomes (e.g. long latency in observing the outcome, substantial financial cost associated with each patient). In settings where pertinent existing data are available, Bayesian designs offer a natural approach for incorporating that information into the design and analysis of future trials. An example where sequential Bayesian methods are valuable is in the area of pediatric clinical trials, where enrollment is often difficult and where relevant data from adult trials are often available. Utilizing information from the adult trials may result in pediatric trials that are more efficient (e.g., increased power, fewer subjects required and/or shorter trial durations).

We present a framework for sequential monitoring that makes structured use of external data and can be consistently applied to a wide range of clinical trial applications. Such standardization of fundamental concepts relating to the trial design provides for coherent interpretation that is appealing to regulators. The basis for this framework is the requirement that parameterization of the monitoring priors and specification of stopping criteria are algorithmic byproducts of the trial hypothesis and relevant external data summaries. This framework defines monitoring priors with a unifying distribution applicable to a wide array of data types (i.e. generalized normal distribution with optional truncation). A procedure for conditional-marginal specification is presented as the default option in multi-parameter settings. This paper uses posterior probabilities to highlight how the level of evidence at interim analyses is linked to monitoring prior specification.

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

This paper 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 $\theta$ 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 $\boldsymbol{D}$ represent the data collected in a trial at some point

in time. For example, for the two-arm trial example above and assuming no covariates other than the treatment indicator, $\boldsymbol{D} = \{y_i, z_i : i = 1, ..., n\}$ where $y_i$ is an indicator of response for patient $i$ and $z_i$ is an indicator for whether patient $i$ was assigned the investigational treatment. We use the generic representation $p(\boldsymbol{D}|\theta, \eta)$ to reflect the density or mass function for the collective data $\boldsymbol{D}$ as a function of $\theta$ and potential nuisance parameters $\eta$ which could be multi-dimensional. For example, for the two-arm trial example above $\eta$ 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 $(\theta, \eta)$ specified over the parameter space defined with respect to each of the hypotheses (e.g. $\pi\left(\theta, \eta \big| H_i\right)$ for $i = 0, 1$). For ease of exposition, for the remainder of Section 2.1, we will focus on the case where $\theta$ is the only unknown parameter and ignore $\eta$.

The posterior probability of hypothesis $H_i$ is given by

$$p(H_i|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|H_i) \cdot p(H_i)}{p(\boldsymbol{D}|H_0) \cdot p(H_0) + p(\boldsymbol{D}|H_1) \cdot p(H_1)}, \tag{1}$$

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

$$P(\theta \in \Theta_i|\boldsymbol{D}) = \frac{\int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta)d\theta}{\int_{\Theta} p(\boldsymbol{D}|\theta)\pi(\theta)d\theta} = \frac{\int_{\Theta_i} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_i)d\theta \cdot P(\theta \in \Theta_i)}{\sum_{j=0,1} \int_{\Theta_j} p(\boldsymbol{D}|\theta)\pi(\theta|\theta \in \Theta_j)d\theta \cdot P(\theta \in \Theta_j)} \tag{2}$$

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

For these reasons, in what follows we will refer to the quantity $P(\theta \in \Theta_i|\boldsymbol{D})$ as the posterior probability of $H_i$ for ease of exposition.

2.1.2 *Formalizing the Statistical Concept of Substantial Evidence.* For testing one-sided hypotheses such as $H_0 : \theta \leqslant \theta_0$ versus $H_1 : \theta > \theta_0$ using posterior probabilities, the null hypothesis is rejected when $P(\theta > \theta_0|\boldsymbol{D})$ exceeds a predefined threshold. 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 $\epsilon$ 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

$$P_\pi(\theta \in \Theta_i) = 1 - \epsilon, \tag{3}$$

where the subscript $\pi$ in (3) is simply to indicate that the probability is calculated based on $\pi(\theta)$ which could be either a prior or posterior distribution.

2.1.3 *Skeptical and Enthusiastic Monitoring Priors.* Having formalized concepts for *substantial evidence* and being *all but convinced* of a claim, we can now develop a structured framework for constructing skeptical and enthusiastic monitoring priors which will be used to determine early stopping rules for efficacy and futility, respectively. The use of monitoring based on changing the opinion of skeptical and enthusiastic observers has been described as overcoming a handicap (Freedman and Spiegelhalter, 1989) and providing a brake (Fayers et al., 1997)

on the premature termination of trials, and as constructing "an adversary who will need to be disillusioned by the data to stop further experimentation" (Spiegelhalter et al., 1994). Skeptical and enthusiastic monitoring priors represent two extreme but plausible beliefs about the quantity of interest $\theta$ relative to the hypotheses to be tested. The purpose of monitoring priors is to help answer the question "Is the evidence compelling enough to stop enrollment for the trial or possibly end it altogether?"

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

For ease of exposition, consider the hypotheses $H_0 : \theta \leqslant \theta_0$ versus $H_0 : \theta > \theta_0$ where $\theta$ 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 $\theta_1$ is the most likely value of $\theta$ 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

$$P_E(\theta > \theta_0) = 1 - \epsilon, \tag{4}$$

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 $\theta_0$ is 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 satisfying

$$P_S(\theta < \theta_1) = 1 - \epsilon. \qquad (5)$$

In what follows we refer to (4) and (5) 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 (1) will be such that $p(H_0) \approx p(H_1)$. The skeptical prior simply reflects that $\theta \geqslant \theta_1$ is exceedingly unlikely and is therefore consistent with their being clinical equipoise about the two hypotheses.

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

2.1.4 *Maximum Sample Size and Formal Stoppage Criteria.* In this section we formalize a stopping criteria for futility and efficacy and give general advice for specifying a maximum sample size for the trial. Although sequentially monitored trials in principle require no fixed sample size, in practice due to resource constraints it will almost always be the case that a maximum sample size exists. Resources permitting, the maximum sample size, denoted by

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

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

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

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

$$\text{eff}(\boldsymbol{D}) = P_S(\theta > \theta_0 | \boldsymbol{D}), \tag{6}$$

and the threshold for substantial evidence to be

$$\text{eff}(\boldsymbol{D}) > 1 - \epsilon \tag{7}$$

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 $\theta = \theta_0$, $P_E(\theta \leqslant \theta_0|\boldsymbol{D}) \to 0.5$ for large samples sizes. For this reason, we consider a different approach. Recalling that $\theta_1$ 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 $P_E (\theta < \theta_m|\boldsymbol{D}) > 1 - \epsilon$, the trial may be stopped due to having a low probability of reaching a meaningful treatment effect. For example applications in this paper, we consider $c = 0.5$, so that $\theta_m = (\theta_0 + \theta_1)/2$.

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

$$\text{fut}(\boldsymbol{D}) = P_E(\theta \leqslant \theta_m|\boldsymbol{D}) \tag{8}$$

and the threshold for substantial evidence to be

$$\text{fut}(\boldsymbol{D}) > 1 - \epsilon \tag{9}$$

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 $\theta$. While analysis with these priors provides a rational perspective from which one can determine whether interim data are sufficient to cease enrolling patients, the a priori belief of most stakeholders will likely fall somewhere in between the two perspectives. Thus, when interpreting the final data once in hand, intermediate perspectives should be

considered. To that end, we define an inference prior as a mixture prior constructed by mixing the monitoring priors.

Define the inference prior associated with mixing weight $\omega$ as

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

where $\omega \in [0, 1]$. The specification of $\omega$ 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 $\theta$ using the inference prior, $p(\theta|\boldsymbol{D}, \pi_I)$, will be used to compute summaries of $\theta$ such as the posterior mean and quantiles. The posterior distribution for $\theta$ using (10) is

$$p(\theta|\boldsymbol{D}, \pi_I) = \hat{\omega} \cdot p(\theta|\boldsymbol{D}, \pi_S) + (1 - \hat{\omega}) \cdot p(\theta|\boldsymbol{D}, \pi_E) \tag{11}$$

where the updated mixing weight is

$$\hat{\omega} = \frac{\omega \cdot p(\boldsymbol{D}|\pi_S)}{\omega \cdot p(\boldsymbol{D}|\pi_S) + (1 - \omega) \cdot p(\boldsymbol{D}|\pi_E)} \tag{12}$$

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

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 (10) 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 $\omega$. 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, $\omega$ can be informed by the data. Let $\hat{\theta} = \text{argmax}\{p(\boldsymbol{D}|\theta)\}$ be the maximum likelihood estimator of $\theta$ given the data. Define

$$\omega = \omega_{min} + (1 - \omega_{min})\frac{\pi_S(\hat{\theta})}{\pi_S(\hat{\theta}) + \pi_E(\hat{\theta})} \tag{13}$$

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

Prior information can also be incorporated into the monitoring priors by using the predictive distribution of the data under the skeptical and enthusiastic priors. Let $\mathcal{L}(\theta, \eta | \mathbf{D})$ be the likelihood for the new trial data at some point in time.

Compute the two predictive distributions as follows

$$p^{(h)}(\mathbf{D}) = \int \int \mathcal{L}(\theta, \eta | \mathbf{D}) \pi^{(h)}(\theta, \eta) d\theta d\eta \tag{14}$$

with $h \in \{S, E\}$ for the skeptical and enthusiastic prior respectively. Consider $\mathbf{D}_{obs}$ to be the observed data at some point in time. For each predictive density, we compute the following:

$$\psi^{(h)}(\mathbf{D}) = \sum_{\mathbf{D}} p^{(h)}(\mathbf{D}) 1[p^{(h)}(\mathbf{D}) \leqslant p^{(h)}(\mathbf{D}_{obs})] \tag{15}$$

This can be interpreted as the probability of observing data as or more extreme given the predictive distribution (either skeptical or enthusiastic predictive distribution).

Based on this, we can set the mixing weight as

$$w^{(E)}(\mathbf{D}) = \max(\psi^{(E)}(\mathbf{D}) - \psi^{(S)}(\mathbf{D}), 0) \tag{16}$$

with $w^{(S)}(\mathbf{D}) = 1 - w^{(E)}(\mathbf{D})$.

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 patients whose outcomes were in progress at the time of enrollment termination. Let $\boldsymbol{D}(n)$ denote the data after $n$ completed outcomes are ascertained. Type 1 error and power are computed based on $\text{eff}(\boldsymbol{D}(n_{initial}))$ and $\text{eff}(\boldsymbol{D}(n_{final}))$, respectively. The posterior mean and credible intervals for $\theta$ are also evaluated for both $\boldsymbol{D}(n_{initial})$ and $\boldsymbol{D}(n_{final})$.

There is the possibility for disagreement between $\text{eff}(\boldsymbol{D}(n_{initial}))$ and $\text{eff}(\boldsymbol{D}(n_{final}))$. Of

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

2.1.8 *Computation.* The efficacy criteria (6) and futility criteria (8), 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). MCMC would likely be necessary if the dimension of $(\theta, \eta)$ is greater than 2.

2.2 *Monitoring Prior Parameterization*

Skeptical and enthusiastic monitoring priors defined in Section 2.1.3 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 skeptical and enthusiastic priors.

2.2.1 *Normal Distribution $\mathcal{N}_p(\tilde{\mu}, q)$.* Define $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$ as the normal distribution $\mathcal{N}(\mu, \sigma)$ that satisfies $\text{mode}(\theta) = \tilde{\mu}$ and $P(\theta \leqslant q) = p$. The values for the mean and standard deviation are $\mu = \tilde{\mu}$ and $\sigma = \frac{q-\mu}{\Phi^{-1}(p)}$, where $\Phi$ denotes the cumulative distribution function for a standard normal distribution, and $\Phi^{-1}$ denotes its quantile function. The distribution $\mathcal{N}_p(\tilde{\mu}, q)$ 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 $P(\theta \leqslant \tilde{q})$ are also specified for any $\tilde{q}$. In particular, if $\theta \sim \mathcal{N}_p(\tilde{\mu}, q)$ then $P(\theta \leqslant \frac{q+\mu}{2}) = \Phi(\frac{\Phi^{-1}(p)}{2})$. Furthermore, $P(\theta \in (\mu, \frac{q+\mu}{2})) = |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$.

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 $\mu$ 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, and computations involving $F_{\mu,\alpha,\beta}$ are done using the gnorm package (Griffin, 2018).

Define $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ as the generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ that satisfies mode$(\theta) = \tilde{\mu}$, $P(\theta \leqslant q) = p$, and $P(\theta \in (q, \frac{q+\mu}{2})) = \gamma \cdot |p - \Phi(\frac{\Phi^{-1}(p)}{2})|$. The mode is equal to $\mu = \tilde{\mu}$, and $\alpha$ 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). The restrictions on well-defined choices of $(p, \tilde{\mu}, q, \gamma)$ for the $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ distribution are described in the appendix.

The flexibility to modify $\gamma$ can change the distribution to be flatter or more concentrated around the modal value. Choosing $\gamma = 1$ corresponds to the normal distribution which is viewed as a default choice. If $\gamma > 1$ then the probability that $\theta \in (q, \frac{q+\mu}{2})$ increases and the concentration of $\theta$ around the modal value $\mu$ accordingly decreases, resulting in a flatter distribution around the modal value. If $\gamma < 1$ then the probability that $\theta \in (q, \frac{q+\mu}{2})$ decreases and the concentration of $\theta$ around the modal value $\mu$ accordingly increases, resulting in a more concentrated distribution around the modal value. Note that the tail probability $P(\theta \leqslant q) = p$ is maintained regardless of the choice of $\gamma$.

Flattening or concentrating the distribution around the modal value is useful for specifying monitoring priors that reflect varying opinions about the dispersion of $\theta$. An example of parameterizing an enthusiastic prior with a $\mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ distribution is demonstrated in Figure 1. Choosing a concentrated distribution is appropriate when prior belief reflects a higher degree of certainty that $\theta$ is in a narrow range around $\theta_1$, while maintaining residual

uncertainty that $\theta < \theta_0$. This decreases the variance of the distribution, and can be seen to reflect a conservative opinion about the values of $\theta$. Choosing a flattened distribution is appropriate when $\theta$ is more likely to be in a wide range of values around $\theta_1$, while maintaining the same residual uncertainty that $\theta < \theta_0$. This increases the variance of the distribution, and can be seen to reflect a liberal opinion about the values of $\theta$. The specification of $\gamma = 0.75$ will be used henceforth as a pre-determined value for creating a concentrated distribution, and $\gamma = 1.5$ will be used for creating a flattened distribution. These values of $\gamma$ are chosen based on the intuitive graphical properties of the distribution (e.g. noticeably peaked around $\theta_1$, relatively flat around $\theta_1$). Alternatively, the choice of $\gamma$ could be specified to identify an additional probability constraint exactly (e.g. $P(\theta \in (\theta_0, \frac{\theta_0 + \theta_1}{2})) = 0.1$). The impact of flattening or concentrating the distribution of a monitoring prior on the operating characteristics of a trial is shown in Section 3.

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

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 $\text{mode}(\theta) = \tilde{\mu}$, $P(\theta \leqslant q) = p$, 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.

[Figure 1 about here.]

The generalized normal distribution can be used to parameterize skeptical and enthusiastic priors for trials with multiple unknown quantities of interest. Let $\theta$ be the parameter of interest and $\eta$ be the nuisance parameters. Consider the following representation of the joint prior for $\theta$ and $\eta$: $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$. For example, suppose that $\theta$ is the risk

difference between response probabilities of the treatment group and the placebo group $\eta$. The distribution for the risk difference is $\pi(\theta) \sim \mathcal{GN}_{p,\Theta=[-1,1]}()$. The domain of $\eta$ is $[0, 1-\theta)$ if $\theta \geqslant 0$ and $(-\theta, 1)$ if $\theta < 0$. Define the prior for $\eta$ 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.

[Figure 2 about here.]

## 3. Examples

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

3.1.1 *Motivating Example.* Consider the T72 pediatric trial "A Study of the Safety and Efficacy of Infliximab (REMICADE) in Pediatric Patients With Moderately to Severely Active Ulcerative Colitis" (NCT00336492) (Hyams et al., 2012). Infliximab was given to all patients at the 5mg/kg dose at weeks 0, 2, and 6, and the primary endpoint was response at week 8. Response was measured by improvement in disease severity scores.

The trial was conducted between August 2006 and June 2010. Patients were enrolled over approximately 33.5 months (1 patient enrolled per 17 days)

The sample size of 60 patients was chosen to ensure 12% precision in estimating the true response proportion with 95% confidence interval, assuming a response rate of 0.67 as was observed among adults from the ACT 1 and ACT 2 trials (Rutgeerts et al., 2005) at the same 5mg/kg dose (N = 242). A 95% confidence interval that excluded 0.40 was determined to be a clinically significant result. At week 8, clinical response was observed in 44 out of 60 (73.3%) patients.

3.1.2 *Model Formulation & Prior Elicitation.* The data $\boldsymbol{D}$ are assumed to be independent Bernoulli random variables with common response probability $\theta$. 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 $H_0 : \theta \leqslant \theta_0$ 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 4. An inference prior is defined as the mixture (10) with $\omega = 0.5$. The maximum sample size was increased from 60 to increase the probability of stopping the trial early due to efficacy at response proportion values between the null $\theta_0 = 0.4$ and the adult data $\theta_1 = 0.67$. The sample size of $n_{max} = 112$ based on a frequentist design to have 80% power to conclude efficacy with the skeptical monitoring prior at a true response proportion of $\theta_m = 0.535$. This increased power provides a better setting to demonstrate how the flattening or concentration of the skeptical monitoring prior impacts the operating characteristics.

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

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

[Figure 3 about here.]

3.1.4 *Preposterior Analysis of Operating Characteristics.*   Consider more frequent monitoring with an interim analysis occuring after every 2 patients 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 when the efficacy criteria is satisfied at an interim analysis triggering termination of enrollment of additional patients, but once the outcomes of patients 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 $\theta$ 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 the case of evidence decrease.

[Figure 4 about here.]

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 and final analysis when $\theta = \theta_0$. The initial sample size is the sample size when the efficacy criteria is satisfied and enrollment is henceforth terminated. The final sample size includes outcomes from those patients in progress at time of trial termination. If the efficacy criteria is not satisfied at any of the

interim analyses, then the initial and final sample sizes are equal to the maximum sample size.

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).

[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. 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 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.*   The data $\boldsymbol{D}$ 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. 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 $H_0 : \theta \leqslant \theta_0$ v.s. $H_1 : \theta > \theta_0$. The 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 $\pi_S(\theta) \sim \mathcal{GN}_{p=0.975, \Theta=[-1,1]}(\tilde{\mu} = \theta_0, q = \theta_1, \gamma = 0.75)$. The value $\gamma = 0.75$ reflects a concentrated distribution which was chosen to reflect a conservative opinion for added Type 1 error control. The enthusiastic monitoring prior is $\pi_E(\theta, \eta_0) = \pi_E(\theta) \times \pi(\eta_0|\theta)$, where $\pi_E(\theta) \sim \mathcal{GN}_{p=0.025, \Theta=[-1,1]}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$. The value $\gamma = 1$ was chosen as the default value. For both the skeptical and enthusiastic monitoring prior, $\pi(\eta_0|\theta) \sim \mathcal{GN}_{p=0.975, H=[max(-\theta,0),min(1,1+\theta)]}(\tilde{\mu} = 0.39, q = 0.59, \gamma = 1.5)$. The value $\gamma = 1.5$ was chosen to reflect a liberal opinion about the distribution of $\eta_0$. In general, it is recommended to use flattened priors for nuisance parameters.

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.* A mixture prior of the form (10) is used for efficacy monitoring where the choice of $\omega$ 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 is applied to the enthusiastic component. Additionally, a mixture prior with dynamic weight of the form (13) is used with $\omega_{min} = 0.25$.

Figure 6 shows the probability of stopping early for efficacy based on the choice of $\omega$ 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 $\theta = \theta_1$. 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.

[Figure 6 about here.]

## 4. Discussion

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

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

demonstrates how the operating characteristics are affected by the choice of monitoring priors.

Examples included binomially distributed data with 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.

REFERENCES

Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association* **58,** 365–383.

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

Berger, J. O. and 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,** 1197.

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

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

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.

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

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

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

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

Freedman, L. S. and Spiegelhalter, D. J. (1992). Application of bayesian statistics to decision making during a clinical trial. *Statistics in Medicine* **11,** 23–35.

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.

Gsponer, T., Gerber, F., Bornkamp, B., Ohlssen, D., Vandemeulebroecke, M., and Schmidli, H. (2014). A practical guide to bayesian group sequential designs. *Pharmaceutical Statistics* **13,** 71–80.

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.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35,** 549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–199.

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. (1993). Applying bayesian ideas in drug development and clinical trials. *Statistics in Medicine* **12,** 1501–1511.

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].

SUPPORTING INFORMATION

APPENDIX SUPPLEMENTARY MATERIAL

*Generalized Normal Cumulative Distribution Function*

The CDF of a generalized normal random variable $\theta \sim \mathcal{GN}(\mu, \alpha, \beta)$ can be expressed as (Griffin, 2018)

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

*Parameterizing the $\mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ Distribution*

The restrictions on the choice of $(p, \tilde{\mu}, q, \gamma)$ for the $\theta \sim \mathcal{GN}_p(\tilde{\mu}, q, \gamma)$ distribution to be well-defined:

*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 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 7 about here.]

*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 $\theta_0$ 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 8 and 9, 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 8 about here.]

[Figure 9 about here.]
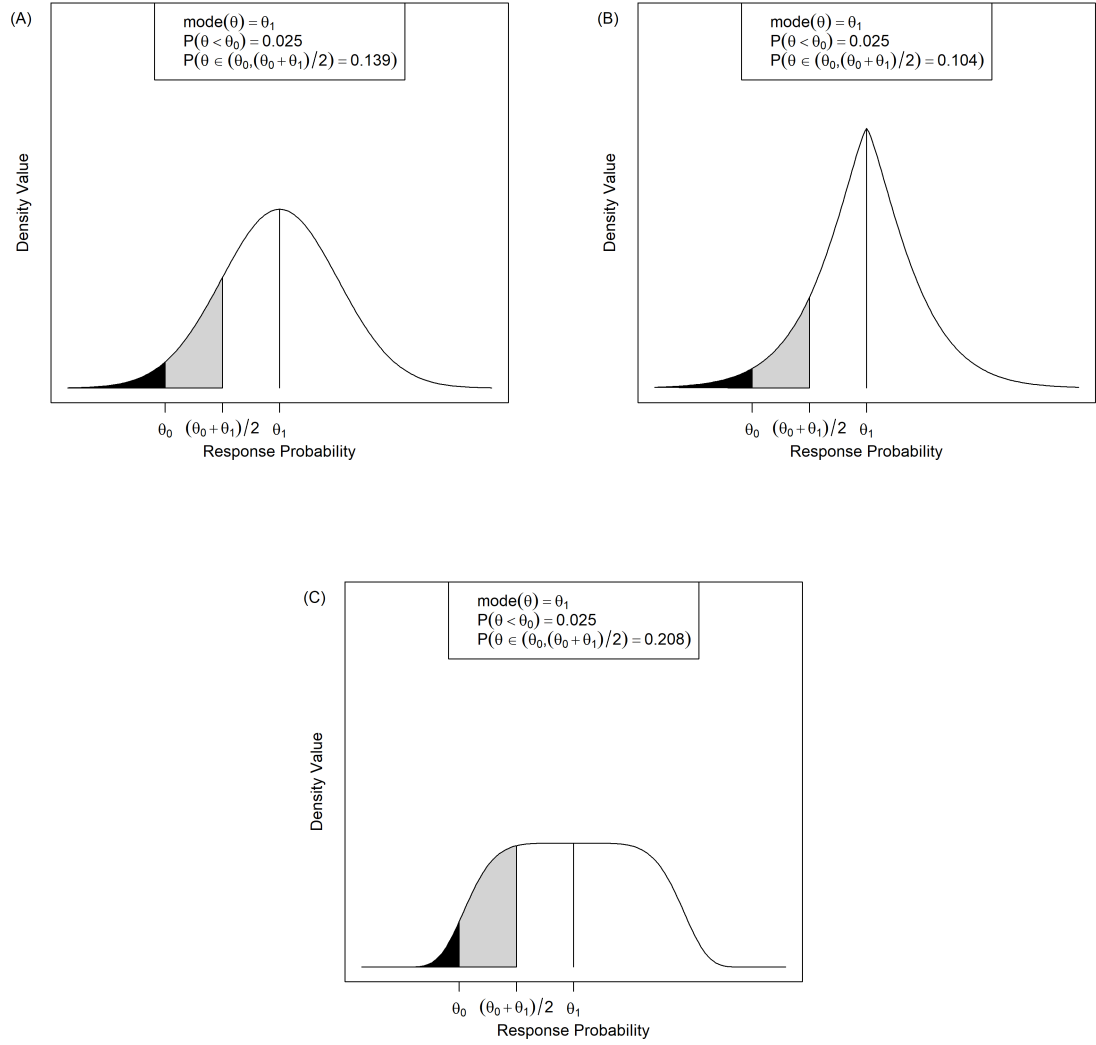
**Figure 1.** A, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1)$. B, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 0.75)$, C, $\mathcal{GN}_{p=0.025}(\tilde{\mu} = \theta_1, q = \theta_0, \gamma = 1.5)$

**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 $\pi(\eta|\theta) \sim \mathcal{GN}_{p=0.975,H=[\max(-\theta,0),\min(1,1+\theta)]}(\tilde{\mu} = \eta_0, q = \eta_1, \gamma = 1)$.

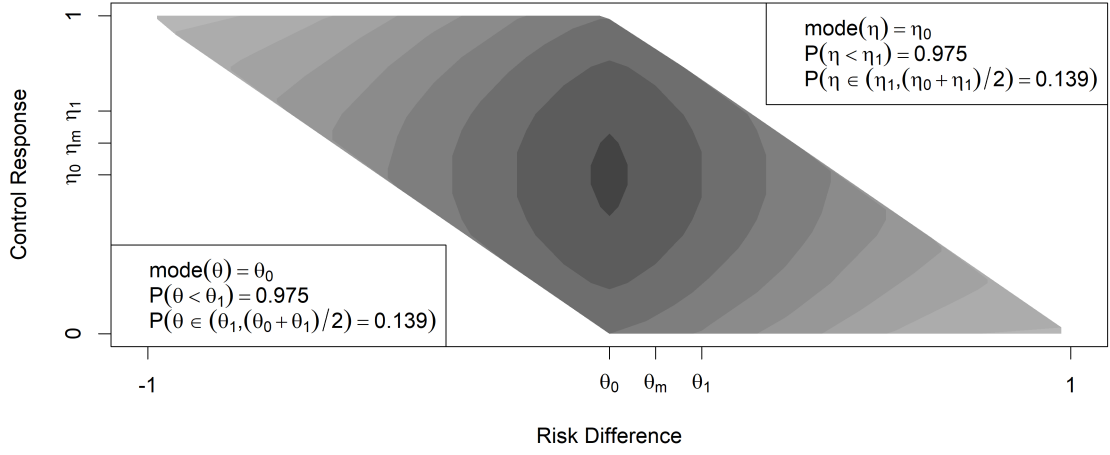| | | | | | |
|---|---|---|---|---|---|
| Enrolled | 0 | 12 | 23 | 34 | 34 |
| Completed | 0 | 10 | 20 | 30 | 34 |
| Response Rate | n/a | 70.0% | 60.0% | 70.0% | 61.8% |
| Efficacy Criteria | 0.504 | 0.852 | 0.876 | 0.994 | 0.970 |
| Futility Criteria | 0.165 | 0.079 | 0.148 | 0.020 | 0.094 |

| | | | | | |
|---|---|---|---|---|---|
| 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 |

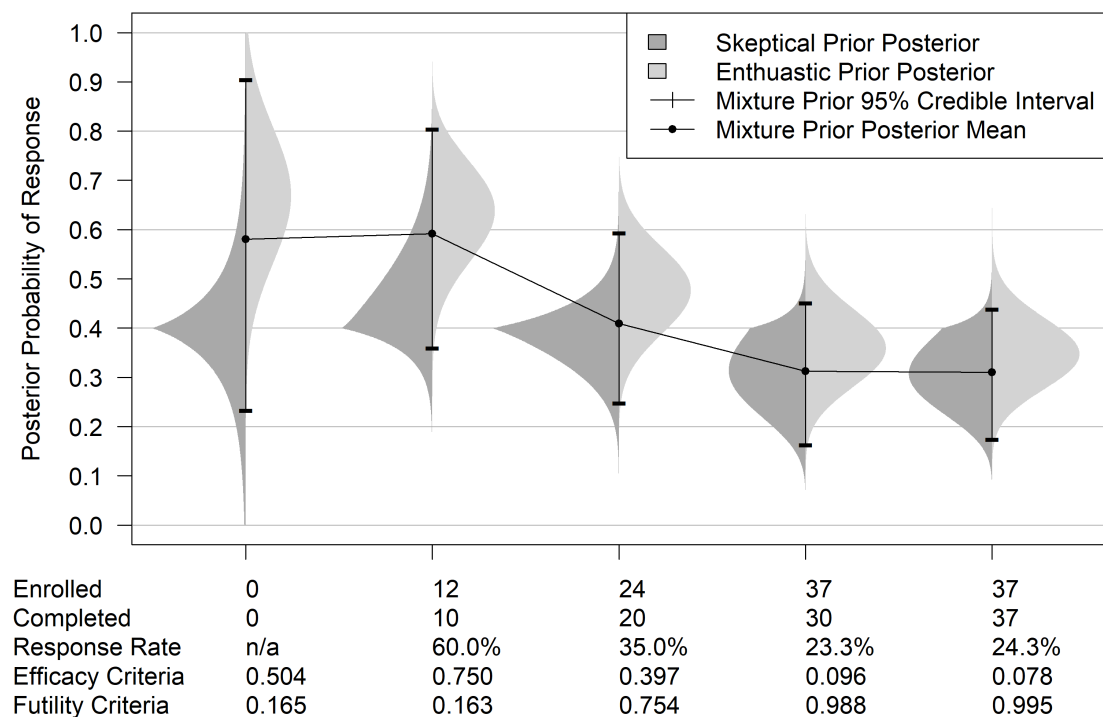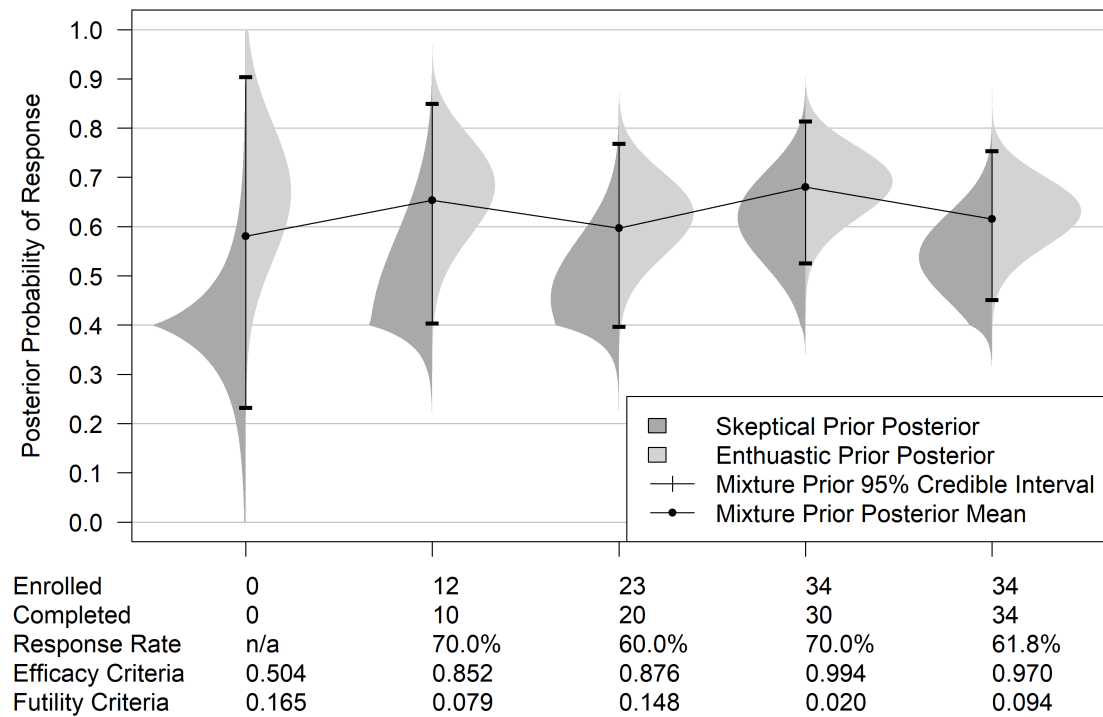**Figure 3.** Example paths for the trial described in Section 3.1.2. A, Early stoppage for efficacy. B, early stoppage for futility.

**Plot A**

Probability (y-axis): 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0

Data labels: 0.039, 0.315, 0.808, 0.988, 1.000

Legend:
— Stop Early for Efficacy
--- Stop Early for Futility
······ Inconclusive Findings

| $\theta$ | 0.4000 | 0.4675 | 0.5350 | 0.6025 | 0.6700 |
|---|---|---|---|---|---|
| SS | 73.0 + 6.2 = 79.2 | 87.2 + 4.2 = 91.4 | 66.8 + 6.4 = 73.2 | 41.8 + 7.9 = 49.6 | 27.8 + 8.0 = 35.8 |
| PM | (I) 0.400 (F) 0.402 | (I) 0.480 (F) 0.477 | (I) 0.567 (F) 0.558 | (I) 0.629 (F) 0.621 | (I) 0.678 (F) 0.672 |
| CP | (I) 0.957 (F) 0.958 | (I) 0.914 (F) 0.929 | (I) 0.920 (F) 0.931 | (I) 0.988 (F) 0.968 | (I) 0.987 (F) 0.978 |

**Plot B**

Final Efficacy Criteria (y-axis): 0.97, 0.96, 0.95, 0.94, 0.93, 0.92, 0.91, 0.90

Data labels: 0.959, 0.964, 0.965, 0.965, 0.967

Legend:
—•— Median
▫ P10-P90
▪ P25-75
▨ P90-P100

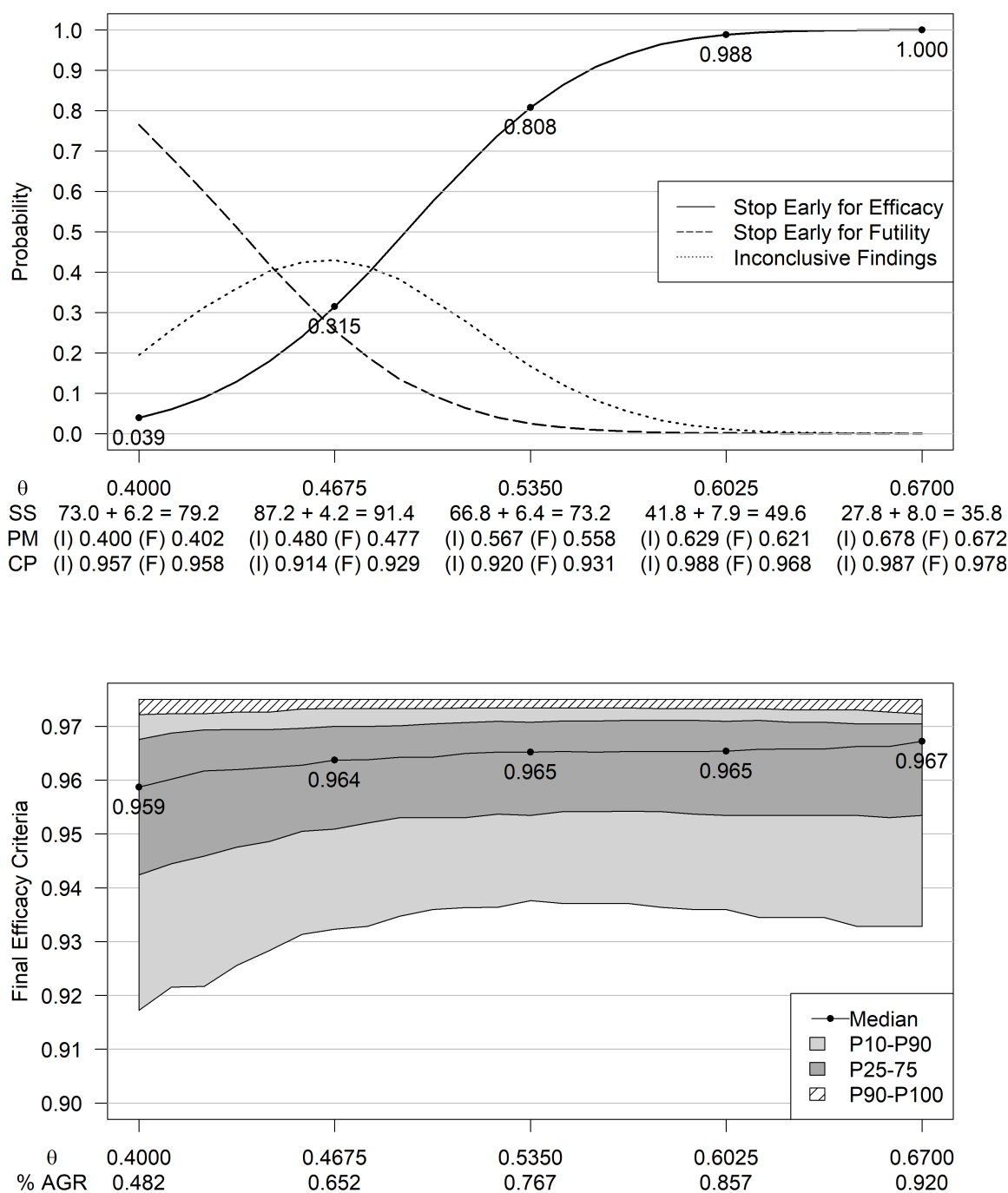| $\theta$ | 0.4000 | 0.4675 | 0.5350 | 0.6025 | 0.6700 |
|---|---|---|---|---|---|
| % AGR | 0.482 | 0.652 | 0.767 | 0.857 | 0.920 |

**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)
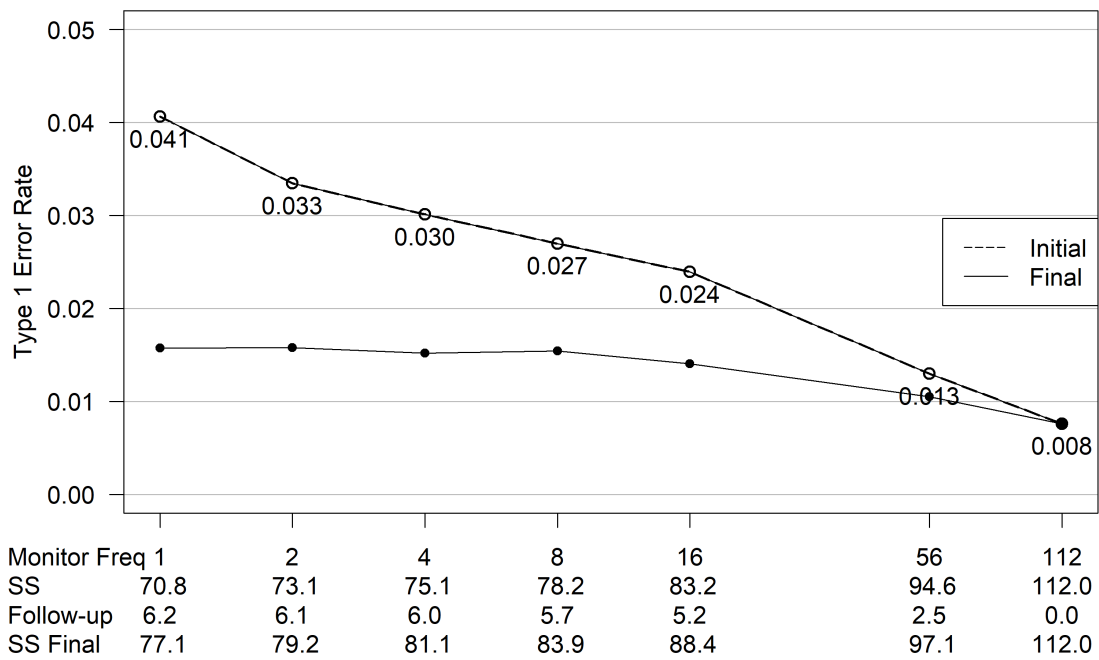
**Figure 5.** Probability of efficacy criteria being satisfied when $\theta = \theta_0$. SS; sample size. Monitor Freq; monitoring frequency.
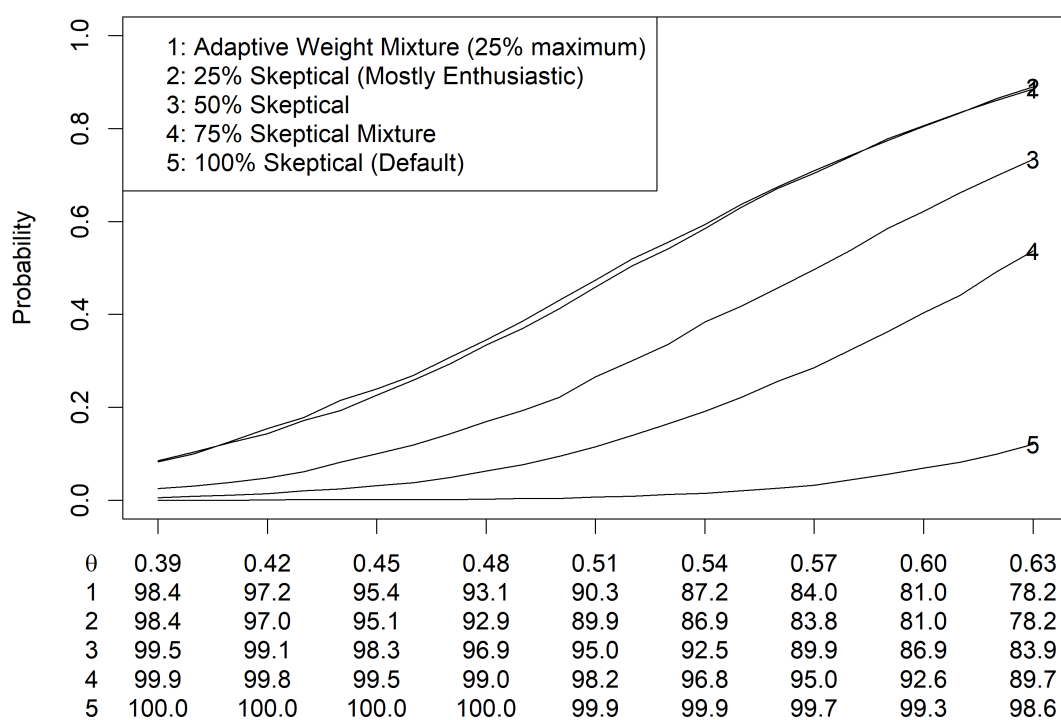
| θ | 0.39 | 0.42 | 0.45 | 0.48 | 0.51 | 0.54 | 0.57 | 0.60 | 0.63 |
|---|------|------|------|------|------|------|------|------|------|
| 1 | 98.4 | 97.2 | 95.4 | 93.1 | 90.3 | 87.2 | 84.0 | 81.0 | 78.2 |
| 2 | 98.4 | 97.0 | 95.1 | 92.9 | 89.9 | 86.9 | 83.8 | 81.0 | 78.2 |
| 3 | 99.5 | 99.1 | 98.3 | 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 $\theta$ with different choices of efficacy monitoring prior.

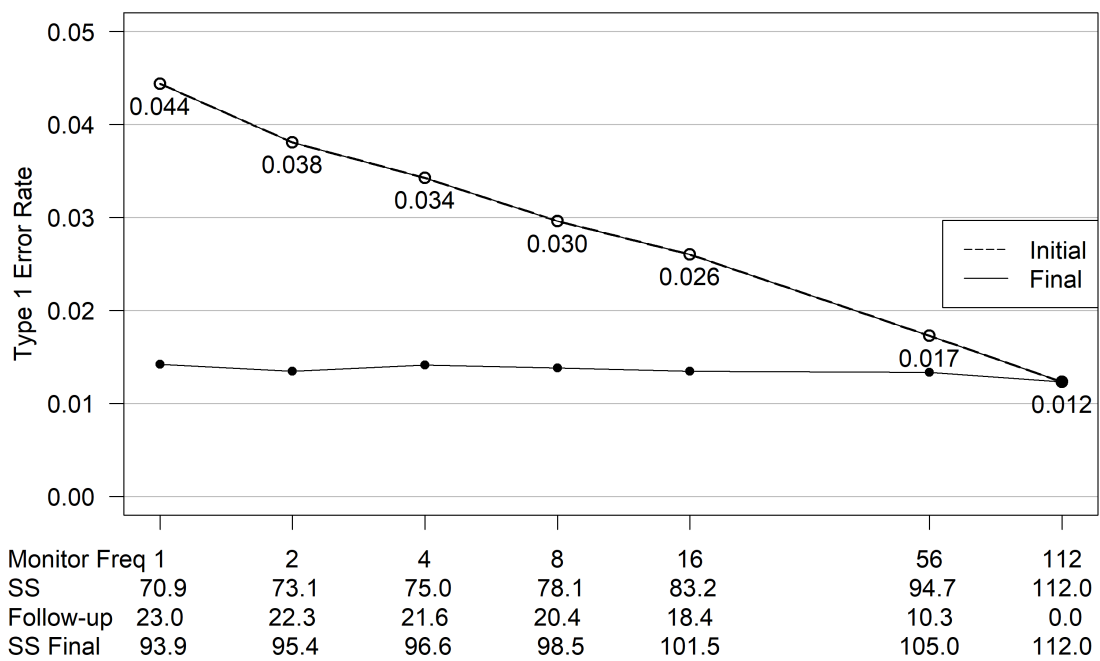| Monitor Freq | 1 | 2 | 4 | 8 | 16 | 56 | 112 |
|---|---|---|---|---|---|---|---|
| SS | 70.9 | 73.1 | 75.0 | 78.1 | 83.2 | 94.7 | 112.0 |
| Follow-up | 23.0 | 22.3 | 21.6 | 20.4 | 18.4 | 10.3 | 0.0 |
| SS Final | 93.9 | 95.4 | 96.6 | 98.5 | 101.5 | 105.0 | 112.0 |

**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.
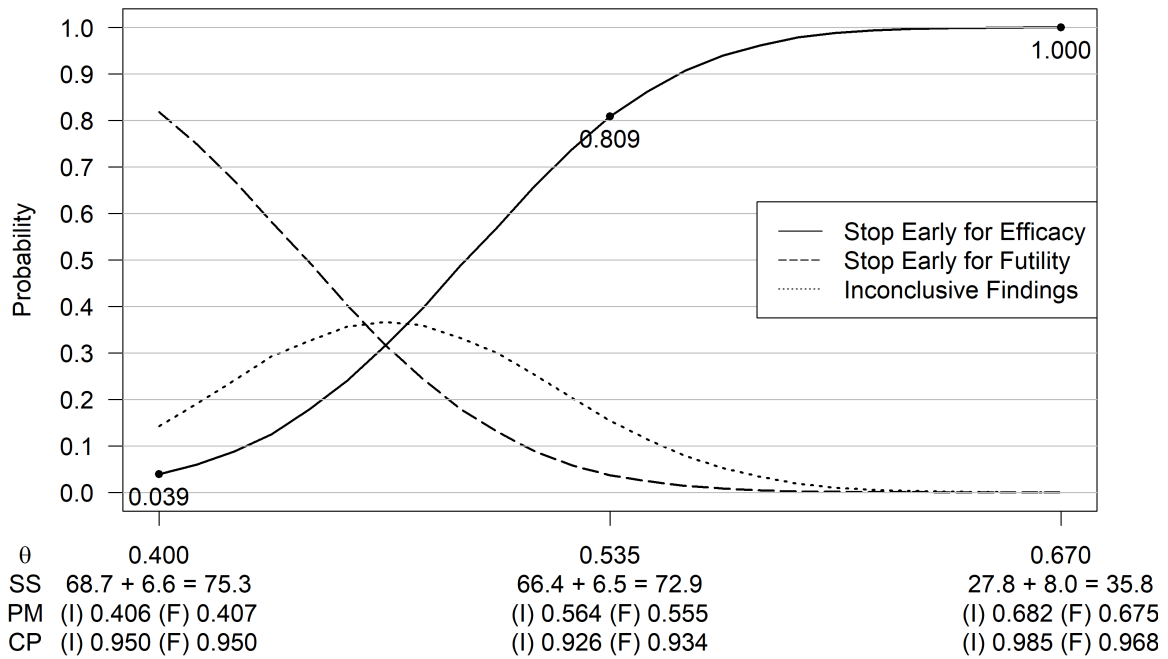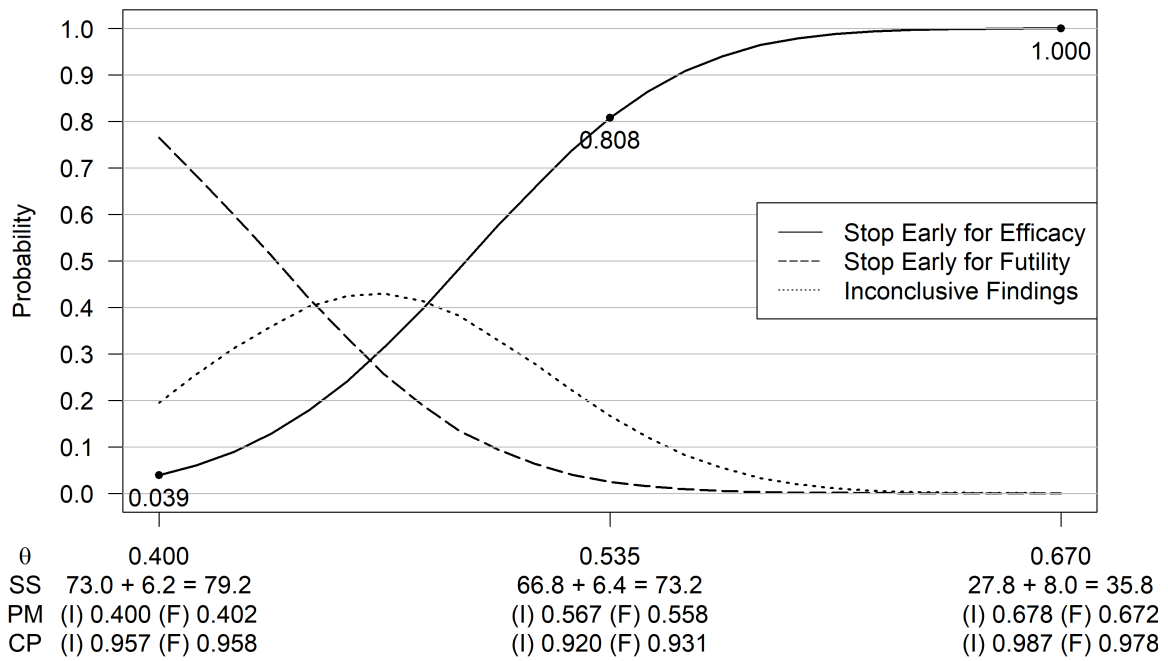
**Figure 8.** Modification of enthusiastic prior parameterization in Example 3.1. A, default enthusiastic prior (Figure 1(c)). B, flattened enthusiastic prior (Figure 1(d)). Both designs use concentrated skeptical prior (Figure 1(b)).
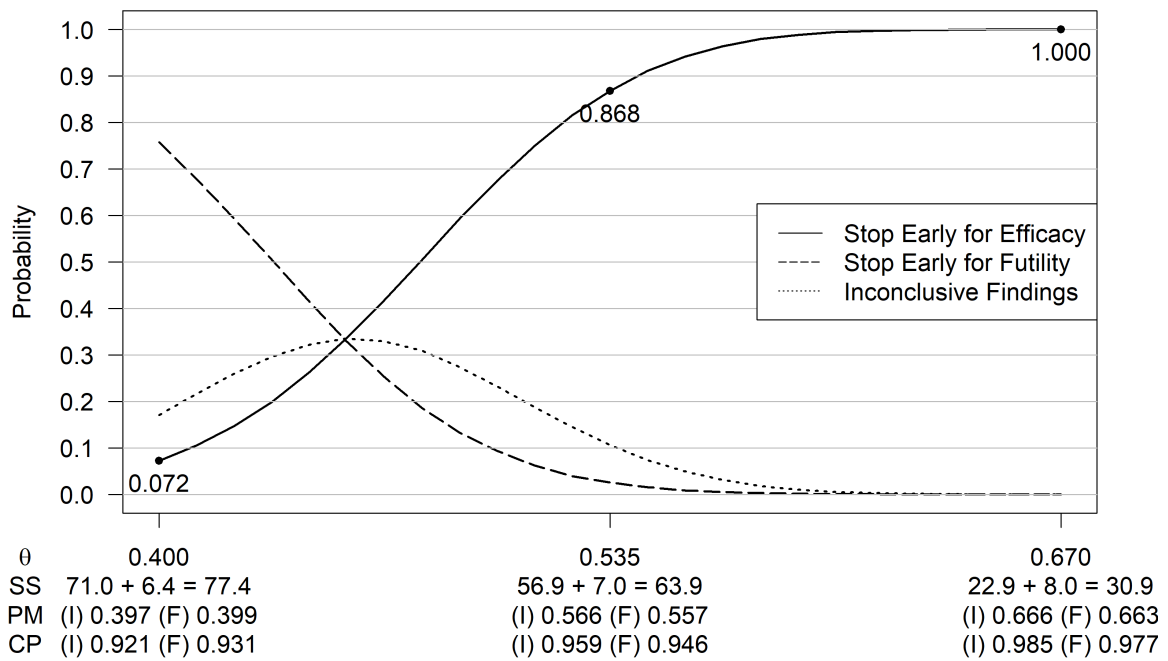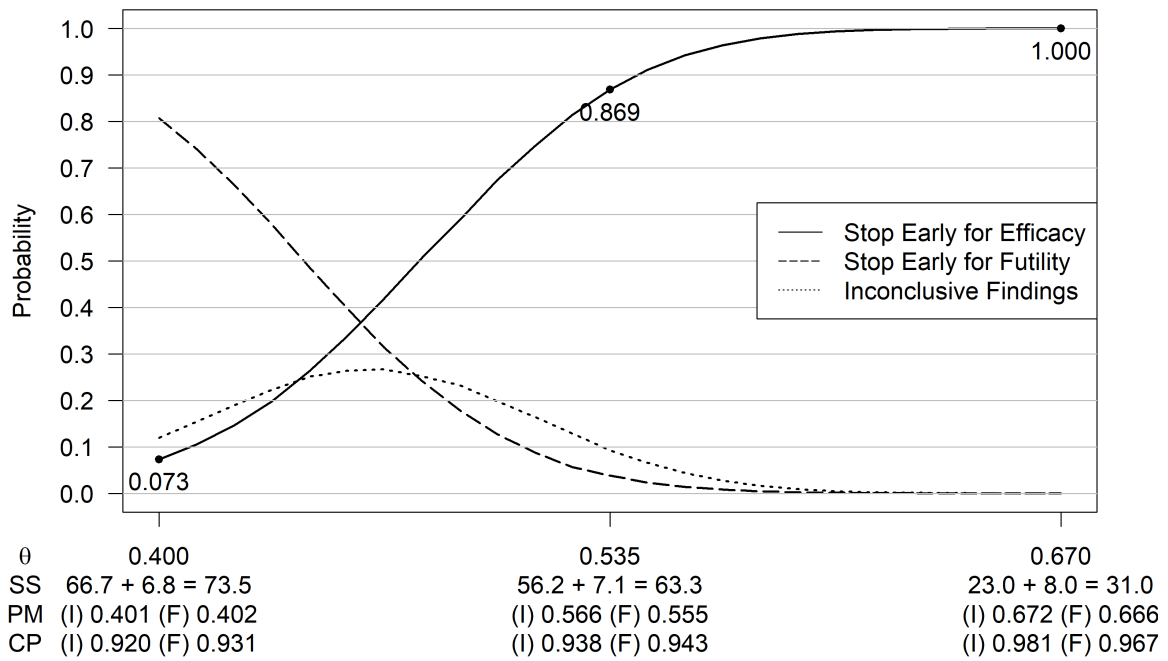
| $\theta$ | 0.400 | 0.535 | 0.670 |
|---|---|---|---|
| SS | 66.7 + 6.8 = 73.5 | 56.2 + 7.1 = 63.3 | 23.0 + 8.0 = 31.0 |
| PM | (I) 0.401 (F) 0.402 | (I) 0.566 (F) 0.555 | (I) 0.672 (F) 0.666 |
| CP | (I) 0.920 (F) 0.931 | (I) 0.938 (F) 0.943 | (I) 0.981 (F) 0.967 |



| $\theta$ | 0.400 | 0.535 | 0.670 |
|---|---|---|---|
| SS | 71.0 + 6.4 = 77.4 | 56.9 + 7.0 = 63.9 | 22.9 + 8.0 = 30.9 |
| PM | (I) 0.397 (F) 0.399 | (I) 0.566 (F) 0.557 | (I) 0.666 (F) 0.663 |
| CP | (I) 0.921 (F) 0.931 | (I) 0.959 (F) 0.946 | (I) 0.985 (F) 0.977 |

**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)).