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

Abstract: We present a Bayesian framework for sequential monitoring that allows for use of external data, and that can be applied in a wide range of clinical trial applications. The basis for this framework is the idea that, in many cases, specification of priors used for sequential monitoring and the stopping criteria can be semi-algorithmic byproducts of the trial hypotheses and relevant external data, simplifying the process of prior elicitation. Monitoring priors are defined using the family of generalized normal distributions which comprise a flexible class of priors, naturally allowing one to construct a prior that is peaked or flat about the parameter values thought to be most likely. External data are incorporated into the monitoring process though mixing an a priori skeptical prior with an enthusiastic prior using a weight that can be fixed or adaptively estimated. In particular, we introduce an adaptive monitoring prior for efficacy evaluation which dynamically weighs skeptical and enthusiastic prior components based on the degree to which observed data are consistent with an enthusiastic perspective. The proposed approach allows for prospective and pre-specified use of external data in the monitoring procedure. We illustrate the method for both single-arm and two-arm randomized controlled trials. For the latter case, we also include a retrospective analysis of actual trial data using the proposed adaptive sequential monitoring procedure. Both examples are motivated by completed pediatric trials, and the designs incorporate information from adult trials to varying degrees. Preposterior analysis and frequentist operating characteristics of each trial design are discussed.

Keywords: Adaptive Trial Design, Bayesian Sequential Monitoring, Information Borrowing, Pediatric Trials, Skeptical Prior.

1 Introduction

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

A key contribution of this work is the provision of structured definitions of skeptical and enthusiastic perspectives that can be used to inform early stopping decisions in favor of efficacy and futility, respectively. Skeptical and enthusiastic priors are developed using the generalized normal family of distributions. This flexible family includes the normal distribution as a special case, and provides the capacity to construct monitoring priors that reflect nuanced prior opinion about the treatment effect. The structured definitions of skeptical and enthusiastic perspectives form the basis for an adaptive monitoring prior to be used for efficacy evaluations monitoring when there is a desire to incorporate prior information into the monitoring process. The prospective use of external data in a pre-specified design provides novelty beyond conducting sensitivity analyses with different priors, and provides a pathway for innovative designs.

This paper is organized as follows: Section 2.1 reviews Bayesian hypothesis testing using posterior probabilities and the use of skeptical and enthusiastic priors for efficacy and futility monitoring. Section 2.2 presents a method for parameterizing monitoring priors the generalized normal distribution and for incorporating prior information into the monitoring priors, and a method to specify priors for nuisance parameters. Section 3.1 presents an example based on a two-arm randomized, controlled trial.

2 Methods

2.1 Preliminaries

2.1.1 Formalizing the Statistical Concept of a Compelling Demonstration

Consider the one-sided hypotheses $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ for fixed θ_0, which we refer to as the boundary null value. Often in Bayesian hypothesis testing, one rejects the null hypothesis when $P(\theta > \theta_0|\mathbf{D})$ exceeds a prespecified threshold. Let ϵ represent *insignificant residual probabilistic uncertainty* regarding a claim. Define $1 - \epsilon$ to be the threshold for posterior probabilities in favor of the claim (e.g., that $\theta > \theta_0$), such that posterior probabilities above $1 - \epsilon$ are considered as providing *a compelling demonstration* that the claim is true. Leveraging common practice, we will use $\epsilon = 0.025$ for the examples presented herein so that $1 - \epsilon = 0.975$ is the threshold that determines when evidence of a claim is compelling.

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

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

where the subscript π in (1) is simply to indicate that the probability is calculated based on $\pi(\theta)$ which could be either a prior or posterior distribution.

2.1.2 Skeptical and Enthusiastic Monitoring Priors

Having formalized concepts for *a compelling demonstration* and being *all but convinced* of a claim, we now can develop a structured framework for constructing skeptical and enthusiastic monitoring priors which will be used to determine early stopping rules for efficacy and futility, respectively. Consider again the hypotheses $H_0 : \theta \leq \theta_0$ versus $H_0 : \theta > \theta_0$ where θ_0 represents a treatment effect of interest and let $\theta_1 > \theta_0$ represent a plausible, clinically

meaningful effect. Define an enthusiastic prior, denoted as $\pi_E(\theta)$, as a prior consistent with θ_1 being the most likely value of θ (i.e., the prior mode) and that reflects the belief of an observer who is *all but convinced* that H_1 is true a priori. Formally, this is defined as satisfying (i) $\text{argmax}_\theta\ \pi_E(\theta) = \theta_1$ and (ii) $P_E(\theta > \theta_0) = 1-\epsilon$, where the subscript E indicates that the probability is based on $\pi_E(\theta)$. Similarly, define a skeptical prior, denoted as $\pi_S(\theta)$, as a prior consistent with θ_0 being the most likely value of θ and that reflects the belief of an observer who is *all but convinced* that $\theta < \theta_1$ is true a priori. Formally, this is defined as the prior $\pi_S(\theta)$ satisfying (iii) $\text{argmax}_\theta\ \pi_S(\theta) = \theta_0$ 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.

2.1.3 Maximum Sample Size and Formal Stoppage Criteria

Stopping criteria for efficacy are defined from the perspective of a skeptical observer. The skeptic becomes convinced that a treatment is effective if at some point the observed data suggest there is a compelling demonstration that the alternative hypothesis is true. Formally, the early stopping criteria are met based on data $\mathbf{D}$ when $P_S(\theta > \theta_0|\mathbf{D}) > 1 - \epsilon$. Note that the evidence must *exceed* the threshold for what defines it as being compelling. When the evidence in favor of the alternative surpasses this threshold, it may no longer be necessary to enroll patients for the purpose of proving treatment efficacy.

Stopping criteria for futility monitoring are defined from the perspective of the enthusiastic observer. At first thought it may seem appealing to stop the trial when the enthusiast becomes convinced that the null hypothesis is true, that is, that $P_E(\theta \leq \theta_0|\mathbf{D}) > 1 - \epsilon$. However, when $\theta = \theta_0$, $P_E(\theta \leq \theta_0|\mathbf{D})$ approaches 0.5 for large sample sizes. Therefore this potential futility criteria would not be satisfiable unless the observed data were consistent with values of θ much less than θ_0. For this reason, we consider a different approach. Recalling that θ_1 represents a plausible, clinically meaningful treatment effect, the early stopping criteria are met based on data $\mathbf{D}$ when $P_E\left(\theta < \theta_1|\mathbf{D}\right) > 1 - \epsilon$. In this case the trial may be

stopped due to there being a compelling demonstration that the treatment effect is much less than hypothesized (i.e., θ_1).

2.2 Specifying Monitoring Priors

2.2.1 Default Monitoring Priors

The skeptical and enthusiastic monitoring priors defined in Section 2.1.2 have mode value and tail-probability constraints. However, these constraints alone do not uniquely determine the priors. There are infinitely many distributions which satisfy these conditions. However, the mode and tail constraints do uniquely determine a pair of normal distributions which might serve as a default set of monitoring priors. A default enthusiastic monitoring prior satisfying (i) $\text{argmax}_\theta\ \pi_E(\theta) = \theta_1$ and (ii) $P_E(\theta > \theta_0) = 1 - \epsilon$ is the normal distribution with location θ_1 and standard deviation $\sigma = \frac{\theta_1 - \theta_0}{\Phi^{-1}(1-\epsilon)}$, where Φ^{-1} denotes the quantile function of a standard normal. The specification of μ and σ completely determine the density at all points, including the value of the density at the mode which is $f(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma}$. The skeptical monitoring prior is similarly defined, satisfying (i) $\text{argmax}_\theta\ \pi_S(\theta) = \theta_0$ and (ii) $P_S(\theta < \theta_1) = 1 - \epsilon$.

Use of normal distributions for the monitoring priors can be motivated by the Bayesian Central Limit Theorem (CLT) (Le Cam & Yang 2000) which states that, under general conditions, the posterior distribution for θ approaches normality as the sample size increases, regardless of the initial choice of prior. Therefore, a normally distributed monitoring prior is consistent with belief derived from a sufficiently large dataset with maximum likelihood estimate equal to the mode value required by the prior.

2.2.2 Generalized Normal Distribution

Despite the aforementioned justification of normally distributed priors, it may be desirable to construct a monitoring prior with different behavior about the mode than what is possible when using the normal distribution. Choosing a flattened distribution is appropriate when

one wishes to reflect more uncertainty regarding the likelihood that θ is near θ_1 (relative to what is permitted by the normal distribution), while maintaining the same residual uncertainty that $\theta < \theta_0$. Similarly, choosing a concentrated distribution is appropriate when one wishes to reflect a higher degree of certainty that θ is near θ_1, while maintaining residual uncertainty that $\theta < \theta_0$.

The family of generalized normal distributions, which contains the normal distribution as a special case, is able to accommodate changes in the density value at the mode while still satisfying the mode value and tail probability constraints. The density for a generalized normal distribution $\mathcal{GN}(\mu, \alpha, \beta)$ is

$$f(\theta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{-\left(\frac{|\theta - \mu|}{\alpha}\right)^{\beta}\right\}$$

where μ is a location parameter, $\alpha > 0$ is a scale parameter, and $\beta > 0$ is a shape 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 density's behavior near the mode. Recall the density at the mode for a default enthusiastic prior is $f(\theta_1) = \frac{1}{\sqrt{2\pi}\sigma}$. An enthusiastic monitoring prior in the generalized normal family of distributions can have density at the mode equal to $k \times \frac{1}{\sqrt{2\pi}\sigma}$, with $k < 1$ indicating a more flattened distribution and $k > 1$ indicating a more peaked distribution at the mode, relative to the default normal distribution.

2.2.3 Incorporating Prior Information in the Monitoring Priors

The monitoring priors are constructed based on the quantities θ_0 and θ_1, as well as the definition of a compelling demonstration. As described previously, prior information may be directly used in the construction of the enthusiastic prior (e.g., choice of θ_1). It also may be desirable to incorporate prior information into the monitoring process when making a determination of when to stop enrollment early for efficacy. To facilitate this, we introduce

a procedure for modifying the monitoring process such that, if the enthusiastic prior is congruent with observed data, the degree of skepticism can be adaptively lessened. We propose incorporating prior information into the monitoring process for efficacy through constructing a mixture prior from the skeptical and enthusiastic priors using a mixing weight that is constructed from a measure of compatibility between the observed data and the enthusiastic prior. We define the *adaptive monitoring prior* for efficacy evaluations as the mixture distribution

$$\pi_{AE}(\theta) = \omega \cdot \pi_E(\theta) + (1 - \omega) \cdot \pi_S(\theta), \tag{2}$$

where $\omega \in (0, 1)$ is an adaptively determined mixing weight. The objective of the proposed approach is to create a mixture prior which favors the enthusiastic prior component in cases where high compatibility is observed between the trial data and the enthusiastic prior, and favors the skeptical prior component if the data observed are incompatible with the enthusiastic prior. This approach is motivated by the rationale that the enthusiastic prior reflects a plausible perspective about the treatment's effect, and one that we assume will typically be informed by data (e.g., from adult trials in the case of a planned pediatric trial).

The adaptive mixing weight ω 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 enthusiastic prior. The prior-predictive distribution for data $\mathbf{D}$ (also called the marginal likelihood) reflects the probability of observing $\mathbf{D}$ given the assumed prior distribution for θ and is defined formally as

$$p(\mathbf{D}) = \int p(\mathbf{D}|\theta)\pi(\theta)d\theta. \tag{3}$$

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

$$\psi(\mathbf{D}_{\text{obs}}) = \int p(\mathbf{D})1[p(\mathbf{D}) \leq p(\mathbf{D}_{\text{obs}})]d(\mathbf{D}) \tag{4}$$

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

Box's p-value can be interpreted as the probability of observing data as or more extreme than $\mathbf{D}_{\text{obs}}$, given the predictive distribution. Small values of $\psi(\mathbf{D}_{\text{obs}})$ indicate a lack of compatibility or congruency between the prior and the data. We propose using the enthusiastic prior $\pi_E(\theta)$ to compute the quantities in (3) and (4) to create a compatibility measurement $\psi^{(E)}(\mathbf{D}_{\text{obs}})$ which is used to determine the mixing weight in (2). Use of Box's p-value as a measure of prior-data conflict has been considered previously (Psioda & Xue 2020), but not in the context of sequential monitoring or using a mixture prior framework as we proposed here.

Define the mixing weight ω given to the *enthusiastic* prior as

$$\omega = (1 - \delta) \cdot \psi^{(E)}(\mathbf{D}_{\text{obs}}) \tag{5}$$

This mixture weight achieves the goal of favoring the enthusiastic component if the trial data are compatible with that prior, and otherwise assigning a higher weight to the skeptical component. The minimum possible mixing weight δ assigned to the *skeptical* prior is achieved when $\psi^{(E)}(\mathbf{D}_{\text{obs}}) = 1$ and is equal to δ. Choices of δ in $\{0, 0.05, 0.10, 0.15, 0.20, 0.25\}$ are explored in Sections 3.1 and 4, and general advice for choosing δ is given in Section 5.

2.2.4 Prior Specification for Nuisance Parameters

If the parameters θ and η are assumed to be independent, then the joint prior can be factored as $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta)$ and the priors $\pi(\theta)$ and $\pi(\eta)$ can be elicited separately. In some cases this is not possible. For example, suppose that θ is the risk difference between response probabilities of a treatment group and the control group, and denote the response probability in the control group by η. In this case θ and η are linked through constrained support (e.g. $0 \leq \theta + \eta \leq 1$). Such a prior specification is demonstrated in Figure 1, and Section 3.1.2 uses this representation of the joint prior. Panel A shows the marginal distribution $\pi(\theta)$, Panel

B shows the conditional distribution $\pi(\eta|\theta = \theta_0)$, and Panel C shows the joint prior $\pi(\theta, \eta)$. In this example, the conditional distribution $\pi(\eta|\theta)$ will look very similar to the marginal distribution of $\pi(\eta)$ except at the boundaries of the parameter space.

3 Example

3.1 Parallel Two-Group Design with Binary Endpoint

3.1.1 Motivating Example

We consider the trial "The Pediatric Lupus Trial of Belimumab Plus Background Standard Therapy (PLUTO)" (NCT01649765) which was conducted between September 2012 and January 2018 (Brunner et al. 2020). The study population was comprised of patients ages 5 though 17 with active systemic lupus erythematosus (SLE), defined as a baseline SELENA SLEDAI score of 6 or above. Patients were randomized to monthly dosing of either belimumab 10mg/kg or placebo, while continuing to receive standard of care therapy regardless of assignment. The primary endpoint was a dichotomous variable reflecting a 4-point or greater reduction in SELENA SLEDAI score from baseline to week 52. The original study design included enrollment of 100 patients, the first 24 patients randomized in a 5:1 allocation ratio (belimumab:placebo) and the remaining 76 patients in a 1:1 ratio, resulting in 58 patients randomized to belimumab and 42 to placebo. The sample size was based on feasibility constraints rather than power considerations. Data from two studies of belimumab in adults having the same disease resulted in a placebo response probability of 0.39, and a 10mg/kg response probability of 0.51. Using these values for the null and hypothesized response probabilities for the treatment group and assuming a response probability of 0.39 for the control group, a frequentist two-sided hypothesis test with confidence level 95% and 80% power would require 266 patients per group. Ultimately, 93 patients were enrolled over approximately 52.5 months (approximately 1 patient enrolled per 17 days). Clinical

response was observed in 28 of 53 (52.8%) of patients randomized to belimumab and in 17 of 40 (43.6%) of patients randomized to placebo.

3.1.2 Model Formulation & Prior Elicitation

We use this trial as a template to demonstrate our framework, in particular the performance of the adaptive monitoring prior defined in Section 2.2.3. The adaptive monitoring prior is necessary since the power analysis in Section 3.1.1 shows the need for many more patients than were available; therefore, a strategy for prospective incorporation of prior information must be implemented for the trial to have a chance of providing a compelling demonstration of efficacy through a pre-specified design. The data $\mathbf{D}$ are assumed to be independent Bernoulli random variables with response probability η_0 for the placebo group and η_1 for the treatment group, with $\theta = \eta_1 - \eta_0$ denoting the difference in response probabilities. This trial has a superiority hypothesis of treatment to control with null difference in response probabilities, denoted by $\theta_0 = 0$. An estimate for the pediatric response probability is denoted by $\eta_0 = 0.39$ (i.e. the sample proportion of responders from the pooled adult studies), and for purposes of monitoring, a plausible, clinically meaningful difference in response probabilities is $\theta_1 = 0.12$ (i.e. based on the pooled adult study's treatment response probability of 0.51).

The skeptical monitoring prior is $\pi_S(\theta, \eta_0) = \pi_S(\theta) \times \pi(\eta_0|\theta)$, where $\pi_S(\theta)$ is a concentrated skeptical prior. The enthusiastic monitoring prior is $\pi_E(\theta, \eta_0) = \pi_E(\theta) \times \pi(\eta_0|\theta)$, where $\pi_E(\theta)$ is a default enthusiastic prior. The conditional prior for the nuisance parameter $\pi(\eta_0|\theta)$ is specified as a flattened prior around the conditional modal value of $\eta_0 = 0.39$. The probability of concluding efficacy at an interim analysis is made using the adpative monitoring prior as described in Section 2.2.3.

A maximum sample size of $n_{\max} = 100$ was chosen based on the original trial protocol. A minimum sample size of $n_{min} = 50$ was chosen to provide an adequate number of placebo controls to be enrolled given the initial 5:1 allocation to the treatment group. An interim

analysis is completed after every two patients have outcomes beginning at n_{min}.

3.1.3 Preposterior Analysis of Operating Characteristics

Figure 2(A) shows the enthusiastic mixture weights ω by choice of δ in (5) for all combinations of response difference between the IP and PC groups, when the PC group is fixed at a 38% response rate (16/42 responses). Observe that the highest mixture weights ω are observed when the response differences are observed to be around 0.12, which was the mode value for the enthusiastic prior, and have maximum values of $1 - \delta$.

The operating characteristics presented in this section are estimated using 2,500 simulated trials per value of θ using the trial design as described in Section 3.1.2. The generating response probability in the placebo group was assumed to be 0.39, and the generating response probability in the treatment group was determined based on risk differences θ in $\{0, 0.03, 0.06, 0.09, 0.12\}$. Figure 2(B) shows the probability of stopping early for efficacy and the associated sample sizes when using the adaptive monitoring prior (2) with different choices of δ in (5). When $\delta = 0$ or $\delta = 0.05$, a conclusion of efficacy is made at an interim analysis 24% and 14% of the time respectively, while this value is 7% or lower when $\delta \geq 0.1$. Reductions in expected sample size are seen with lower choices of δ and higher generated risk differences. When $\delta = 0.1$, a demonstration of efficacy is observed in 53% of simulated trials, with an expected sample size of 90.1. Even though this is a modest reduction from the maximum sample size of 100 for this case, even more favorable reductions are possible when enrollment is comparatively slower and/or when follow-up times are comparatively shorter.

4 Real Data Example

We consider applying the adaptive monitoring prior of (2) to the observed outcomes of the PLUTO trial presented in Section 3.1. Responses were available for 92 patients (one subject in the placebo group had no outcome available due to a protocol violation). Sequential

monitoring after every two completed outcomes was conducted after a minimum sample size of 50 had been reached. Results of this analysis by different choice of δ are shown in Table 2. When $\delta \leq 0.1$, a conclusion of efficacy is made before the maximum sample size of 92. Figure 3(A) shows Box's p-value at the observed data with 90 completed outcomes to be 0.965 which translates directly to the value of ω in the case that $\delta = 0$. Figure 3(B) shows the efficacy posterior probability of 0.979 when $\delta = 0$ so that $\omega = \psi^{(E)}(\mathbf{D}_{\text{obs}})$.

We note that the final sample size is ≥ 90 for all choices of δ. Thus, in this application, due to the 52-week period of follow-up for the primary outcome and despite the slow enrollment, the impact of sequential monitoring would not have been substantial in terms of shortening the overall trial or reducing the number of patients enrolled. However, it would have nonetheless provided a mechanism for prospective incorporation of external evidence in a pre-specified manner for the trial. In situations where the time-to-outcome ascertainment is shorter and/or enrollment is slower relative to the time-to-outcome ascertainment, greater reductions in sample size would be expected.

5 Discussion

Our formulation of the enthusiastic prior enforces that there be residual uncertainty that the null hypothesis is true; it demonstrates strong belief about effectiveness of the treatment yet is still consistent with a degree of equipoise. In the extreme case that interim data are observed to be perfectly consistent with the enthusiastic prior, the residual uncertainty that the null hypothesis is true reflected in the adaptive monitoring prior cannot be less than that reflected in the enthusiastic prior itself. This is a critical feature of the design as it enforces the requirement that observed data must demonstrate some degree of efficacy on their own to justify stopping enrollment early. Without maintaining residual uncertainty as we have done when constructing the enthusiastic prior, it would be possible to conclude benefit in cases where observed data are somewhat consistent with that prior (i.e., $\psi^{(E)}(\mathbf{D}_{\text{obs}}) > 0$) but

also consistent with no benefit (or even harm). This is particularly problematic when the observed data contain little information compared to the source that informs the enthusiastic prior (as is often the case in pediatric settings). Thus, the proposed approach provide a desirable assurance that evidence of efficacy must come, at least in part, from both the prior information and the trial data. A conclusion of treatment efficacy is possible only when there is overwhelming treatment benefit observed in the trial data so as to convince a skeptic on that data's own merit, or, in the more likely scenario, some evidence of benefit from the trial data along with reasonable compatibility with the enthusiastic prior.

Our results in Section 3.1 can be compared to a published post-hoc Bayesian hierarchical analysis (Brunner et al. 2020) which used data from two studies of the use of belimumab in adults. Patients in the pediatric trial had 1.5 times the odds of clinical response with 95% CI (0.6, 3.5), and a meta-analysis of the two adult studies showed an odds ratio of 1.6 with 95% CI (1.3, 2.1). The analysis used a mixture prior which was a weighted sum of a skeptical prior centered at null effect with effective sample size equal to two pediatric patients and an informative prior resulting from the meta-analysis. When the weight of the informative component was 0.55 and above, efficacy was concluded based on a 95% credible interval excluding one. The 0.55 weight of the informative component, interpreted as a 55% weight on the relevance of the adult information to the pediatric population, was determined to be reasonable by the clinical team. Our method contrasts such a post-hoc analysis with the prospective use of a monitoring prior for efficacy which gives weight to the adult data at interim analyses, although both methods show the necessity of information borrowing.

References

Box, G. E. P. (1980), 'Sampling and Bayes' Inference in Scientific Modelling and Robustness', *Journal of the Royal Statistical Society. Series A (General)* **143**(4), 383–430.

Brunner, H. I., Abud-Mendoza, C., Viola, D. O., Calvo Penades, I., Levy, D., Anton, J. & et al. (2020), 'Safety and efficacy of intravenous belimumab in children with systemic lupus erythematosus: results from a randomised, placebo-controlled trial', *Annals of the Rheumatic Diseases* **79**(10), 1340 LP – 1348.

Jennison, C. & Turnbull, B. W. (2000), *Group sequential methods with applications to clinical trials*, Chapman & Hall/CRC, Boca Raton.

Kopp-Schneider, A., Calderazzo, S. & Wiesenfarth, M. (2020), 'Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control', *Biometrical Journal* **62**(2), 361–374.

Le Cam, L. & Yang, G. L. (2000), *Asymptotics in Statistics: Some Basic Concepts*, Springer, New York.

Nadarajah, S. (2005), 'A generalized normal distribution', *Journal of Applied Statistics* **32**(7), 685–694.

Psioda, M. A. & Ibrahim, J. G. (2018), 'Bayesian design of a survival trial with a cured fraction using historical data', *Statistics in Medicine* **37**(26), 3814–3831.

Psioda, M. A. & Xue, X. (2020), 'A Bayesian Adaptive Two-Stage Design for Pediatric Clinical Trials', *Journal of Biopharmaceutical Statistics* .

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

Stallard, N., Todd, S., Ryan, E. G. & Gates, S. (2020), 'Comparison of Bayesian and frequentist group-sequential clinical trial designs', *BMC Medical Research Methodology* **20**(1), 4.

U.S. Congress (2016), '21st Century Cures Act (Pubic Law 114-255, 130 STAT 1033-1344)'.

U.S. Food and Drug Administration (2006), 'Establishment and Operation of Clinical Trial Data Monitoring Committees'.

Ventz, S. & Trippa, L. (2015), 'Bayesian designs and the control of frequentist characteristics: A practical solution', *Biometrics* **71**(1), 218–226.

Zhu, H. & Yu, Q. (2015), 'A Bayesian sequential design using alpha spending function to control type I error', *Statistical Methods in Medical Research* **26**(5), 2184–2196.

Zhu, L., Yu, Q. & Mercante, D. E. (2019), 'A Bayesian Sequential Design for Clinical Trials with Time-to-Event Outcomes', *Statistics in biopharmaceutical research* **11**(4), 387–397.

Tables

Table 1: Summary characteristics of re-analysis of PLUTO trial. I/F = Interim/Final, $\psi^{(E)}(\mathbf{D}_{\text{obs}})$ = Box's p-value using enthusiastic prior, ω = Enthusiastic mixing weight in adaptive monitoring prior, Efficacy Post Prob = Posterior probability of treatment efficacy.

δ	Sample Size (I/F)	$\psi^{(E)}(\mathbf{D}_{\text{obs}})$ (I/F)	ω (I/F)	Efficacy Post Prob (I/F)
0.00	62 / 90	0.914 / 0.965	0.914 / 0.965	0.980 / 0.979
0.05	64 / 92	0.876 / 0.934	0.833 / 0.887	0.976 / 0.962
0.10	76 / 92	0.941 / 0.934	0.847 / 0.841	0.975 / 0.951
0.15	92 / 92	0.934 / 0.934	0.794 / 0.794	0.940 / 0.940
0.20	92 / 92	0.934 / 0.934	0.747 / 0.747	0.928 / 0.928
0.25	92 / 92	0.934 / 0.934	0.701 / 0.701	0.917 / 0.917

Figures

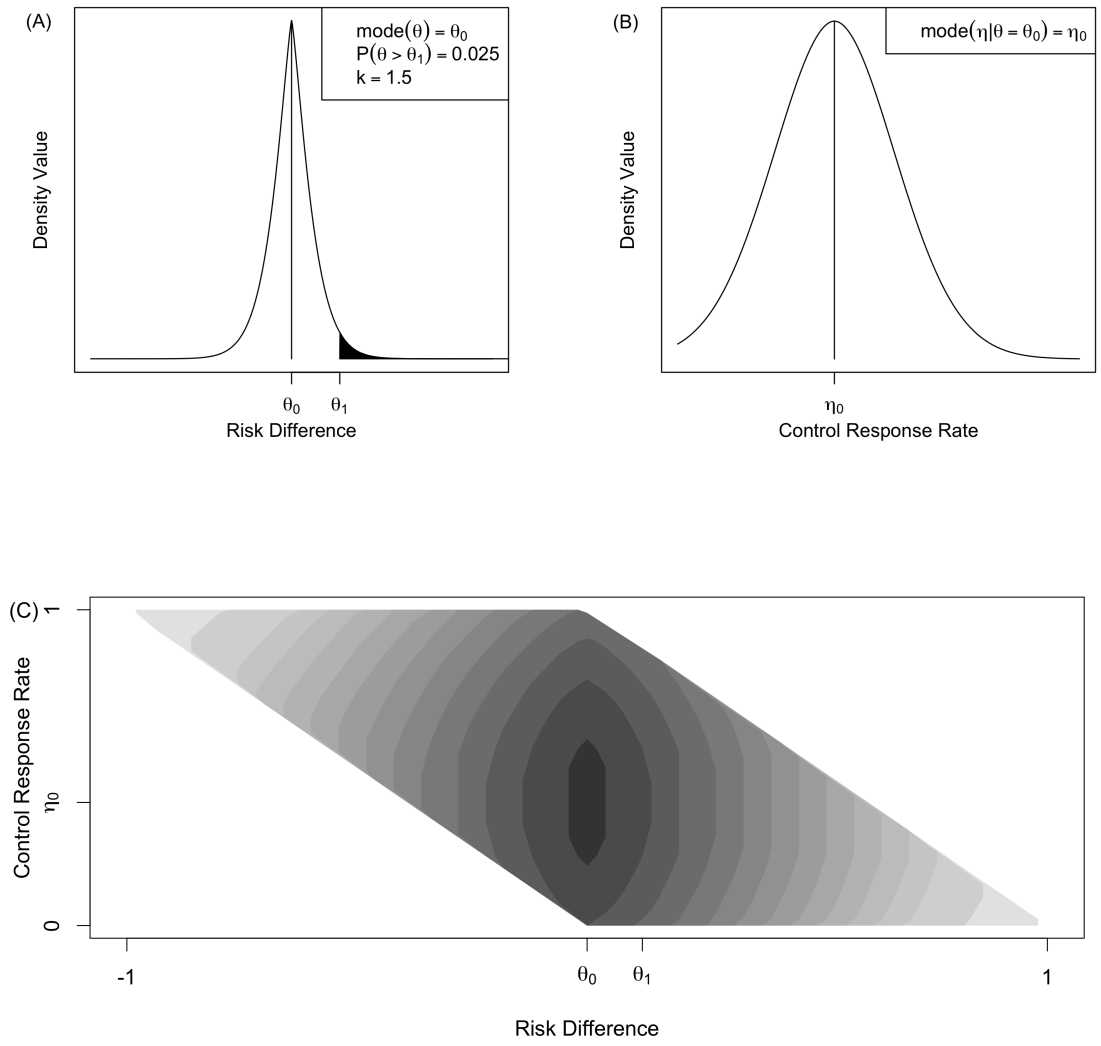
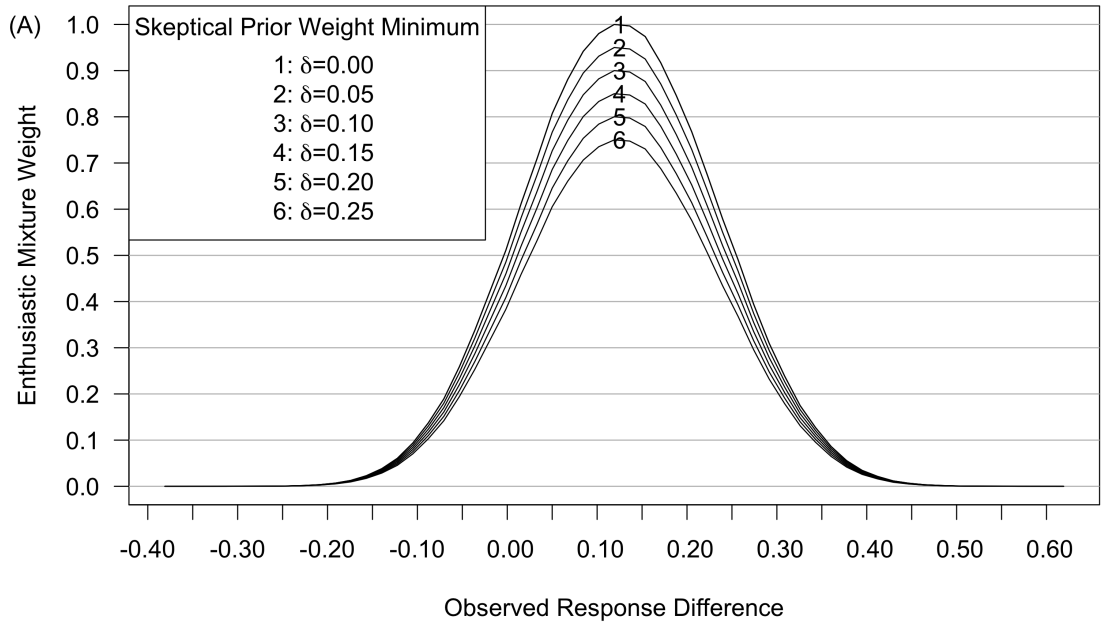
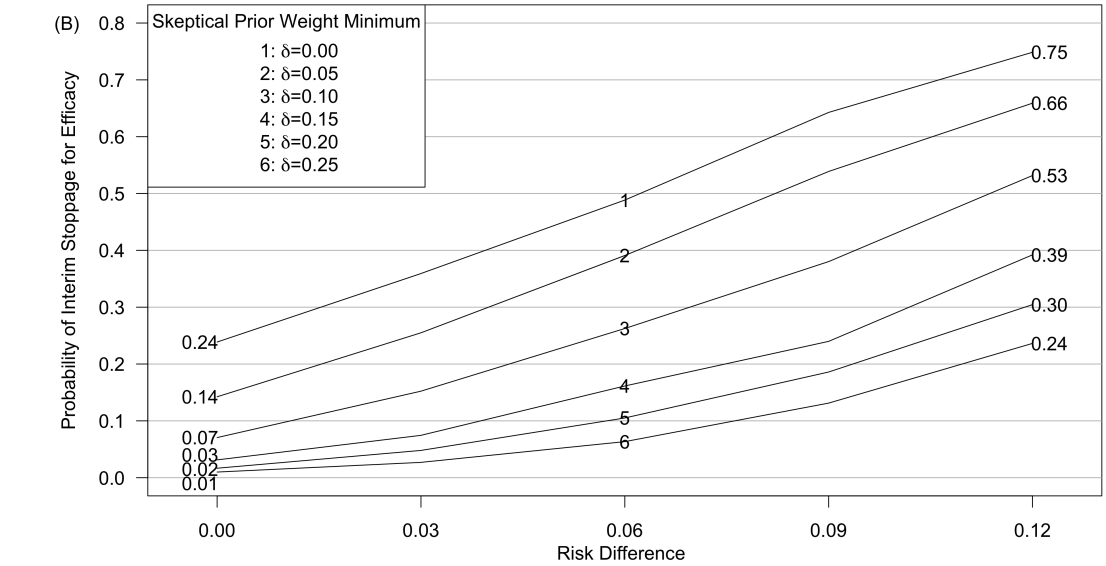


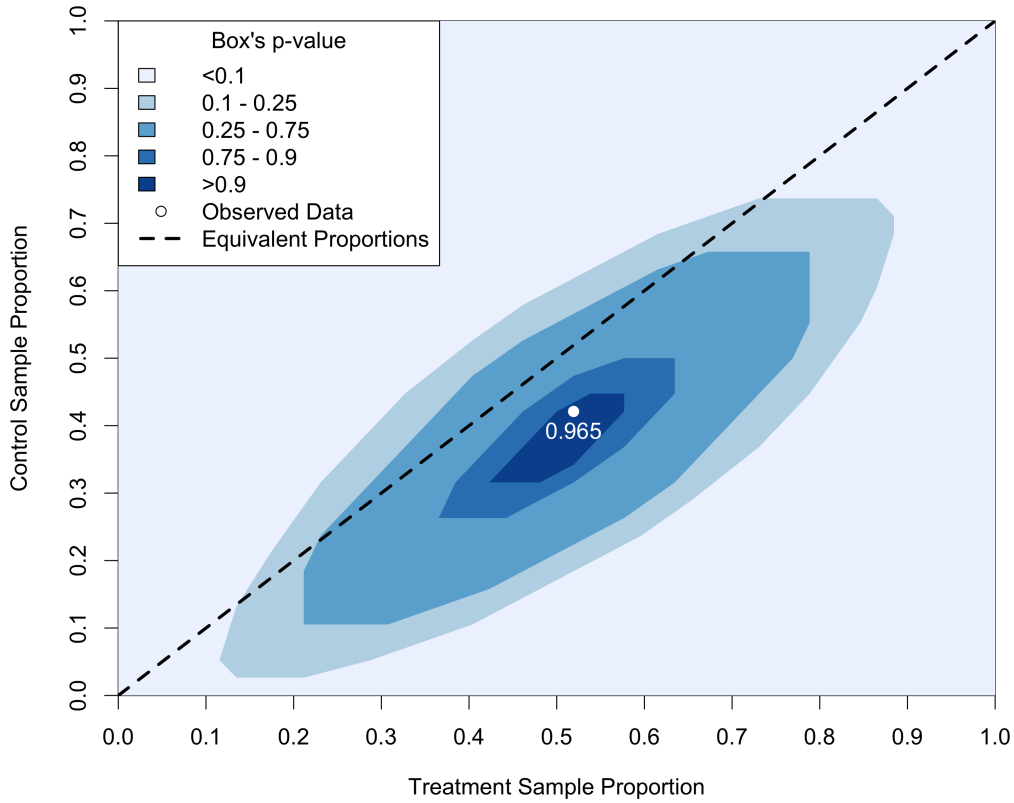
Figure 1: A, Concentrated skeptical prior $\pi_S(\theta)$ truncated to $[-1, 1]$. B, Conditional prior $\pi(\eta|\theta = \theta_0)$. C, Joint prior $\pi(\theta, \eta) = \pi(\theta) \times \pi(\eta|\theta)$ truncated based on the conditions $-1 < \theta < 1$ and $0 < \theta + \eta < 1$.





Expected Sample Size (Interim/Final)					
1: δ=0.00	91.5 / 96.1	87.1 / 94.0	82.6 / 92.0	77.4 / 89.7	73.4 / 88.0
2: δ=0.05	95.5 / 98.3	92.6 / 97.3	89.3 / 96.1	84.5 / 94.3	80.2 / 92.8
3: δ=0.10	98.7 / 99.8	97.6 / 99.8	95.5 / 99.5	93.2 / 99.2	90.1 / 98.7
4: δ=0.15	99.6 / 100.0	99.2 / 100.0	98.2 / 100.0	97.1 / 99.9	94.9 / 99.9
5: δ=0.20	99.7 / 100.0	99.6 / 100.0	99.0 / 100.0	98.4 / 100.0	96.7 / 99.9
6: δ=0.25	99.8 / 100.0	99.6 / 100.0	99.5 / 100.0	99.0 / 100.0	97.9 / 100.0

Figure 2: A, Enthusiastic prior mixing weight ω associated with skeptical prior weight minimum δ in (5) by observed response difference between IP and PC groups, when the PC response rate is fixed at 38% (16/42 responses). B, Operating characteristics for designs having with skeptical prior weight minimum δ in (5) by true risk difference when the PC response rate generated at 39%.



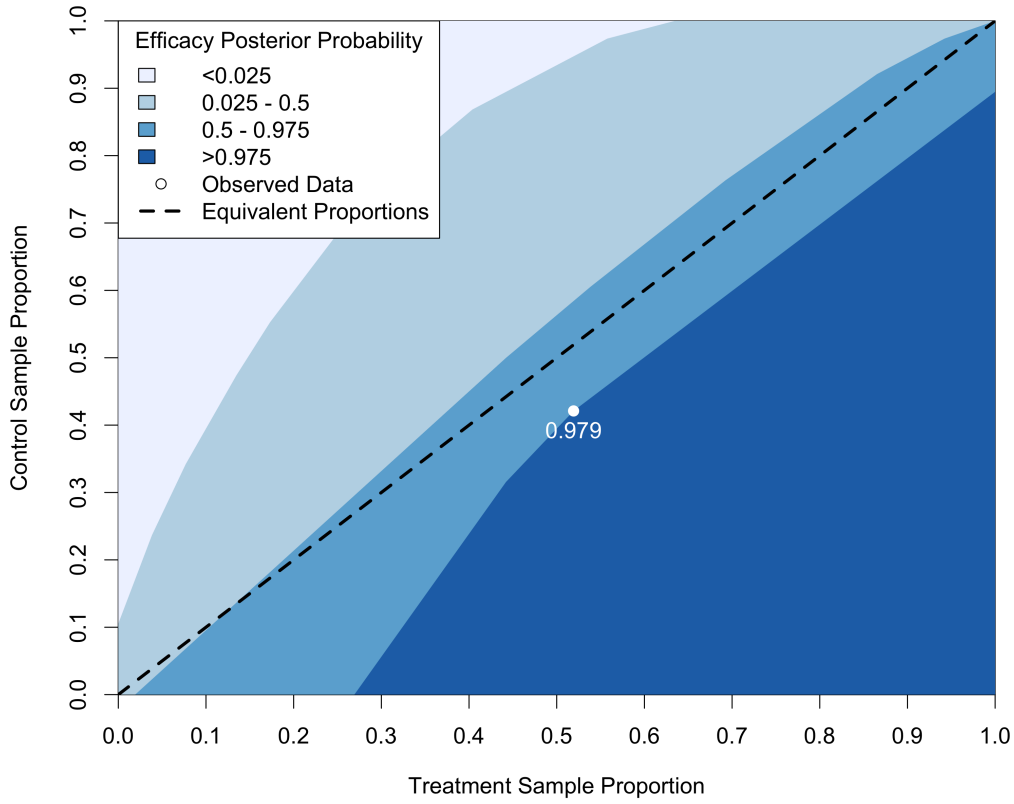


Figure 3: A, Box's p-value by control and treatment sample proportions at the final analysis with 90 subjects when $\delta = 0$ is used (5) for the adaptive monitoring prior. B, Posterior probability of efficacy by control and treatment sample proportions.