Reviewers' Comments to Authors

Reviewer 1

Comments to the Authors:

This manuscript proposes an adaptive strategy for designing sequentially monitored clinical trials based on a Bayesian hypothesis testing framework. A mixture of two priors under null/skeptical and alternative/enthusiastic values for the parameter of interest are the main assumptions be incorporated in the monitoring plans. The manuscript proposed the generalized normal distributions for the priors. The mixing proportion w is set by Box's p-value and capped by a pre-specified parameter δ.

Overall concerns regarding the proposed approaches need further clarifications that may improve this manuscript. Comments are provided with notations "P#,L#" that point the page and line numbers, or so.

[Comment 1]

- Title, Abstract and Introduction:
  The external evidence was mentioned in the title, abstract, and introduction. However, it is not clear what external evidence is and how it is adaptively incorporated through the proposed strategy. Do you mean that the θ_0 and θ_1 are set according to the external evidence? Or, do the skeptical and enthusiastic priors somehow represent the external evidence? Or, does the mixing proportion w of the two priors indicate the proportion of the external evidence?

[Response by Authors] The authors added the paragraph below to the introduction to describe how external evidence is incorporated through the prior distributions used for analysis. Based on our framework, θ_1 is set according to external evidence, the enthusiastic monitoring prior represents external evidence about the treatment effect, and the mixing proportion ω represents the applicability of external evidence in the analysis. Typically, θ_0 represents a null treatment effect that is not informed by external evidence.

For traditional frequentist approaches, prior information (e.g., historical data) is often used to provide external evidence about the magnitude of a plausible, clinically meaningful value for the treatment effect to use for power calculations, but this external evidence generally is not used in analysis once data are actually available. In contrast, the Bayesian paradigm provides a natural framework for incorporating information into both the design and analysis of a future trial. For analysis purposes, external evidence is typically translated into a prior distribution that characterizes what is currently believed about the treatment effect. See, for example, Psioda & Ibrahim (2018) and the references therein for recent work on general Bayesian methods for trial design and analysis using historical data.

[Comment 2]

- P14, Equation (3) and P15, Equation (6):
  It is not clear how this adaptive monitoring prior $\pi_{AE}(\theta)$ can be used in the proposed approach. For example, which formulas/equations/priors in the paper can be replaced by this adaptive monitoring prior? The rest of the method section only focuses on how this prior is formed, but does not mention where it can be applied to the proposed strategy. Does $\pi_{AE}(\theta)$ involve in any calculation for efficacy or futility criteria and how? For example, is a different probability needed for decision making, such as $P_{AE}(\cdots)$ (similar with the $P_E(\cdots)$ and $P_S(\cdots)$)?

[Response by Authors]

To show how the monitoring prior, $\pi_{AE}(\theta)$ is used in the proposed approach, the stopping criteria was given its own line in Equation (3) of Section 2.1.4 "Maximum Sample Size and Formal Stopping Criteria," and was generally defined with respect to any prior distribution:

Stopping criteria are based on whether the posterior probability that the treatment effect is in a particular region is sufficiently large. For region Θ_i, this is formalized as

$$P_\pi(\theta \in \Theta_i|\mathbf{D}) > 1 - \epsilon, \tag{3}$$

The following sentence was added to Section 2.2.3 "Incorporating Prior Information in the Monitoring Priors" to show what calculation uses the adaptive monitoring prior:

The adaptive monitoring prior $\pi_{AE}(\theta)$ is used to make determinations regarding treatment efficacy for monitoring purposes, and is a replacement for the traditional skeptical prior $\pi_S(\theta)$ in (3).

Equivalently stated, the probability $P_S(\cdots)$ for decisions about efficacy is replaced with $P_{AE}(\cdots)$.

[Comment 3]

- P15, Equations (4) and (5):
  Please give examples in appendix or supplementary on how the Equations (4) and (5) can be utilized for a randomized controlled trial. In addition, please clarify if the D_{obs} is used twice in each time point for the ongoing trial. The D_{obs} seems to be involved in w (Equation (6)) for the mixing proportion and in $P(\theta \in \Theta_i|D)$ (Equation (1)). Would the D_{obs} be a part of or equivalent to D?

[Response by Authors]

Yes, as originally formulated, the adaptive monitoring prior uses D_{obs} in computing the mixing weight ω given to the enthusiastic prior component Equation (4), and in the assessment of the stopping criteria Equation (5). Note that the new Equation (3) introduced in response to Reviewer #1 [Comment 2] has shifted the original equation numbers one number higher.

The authors appreciated the chance to clarify the use of D_{obs} and D, and it was correctly pointed out that D_{obs} would be equivalent to D in that instance. This presented an opportunity to rectify the ambiguous use of notation throughout the entire manuscript. The term D_{obs} was dropped completely; instead, D always refers to observed data and the term D_{rep} refers to hypothetical data used in the expression for the predictive distribution and Box's p-value. The revised Equations (5) and (6) are below:

which is defined (in our case) using the enthusiastic prior. The prior-predictive distribution for replicated data denoted by $\mathbf{D}_{\text{rep}}$ reflects the probability of observing this hypothetical data given the assumed data generating mechanism and prior for θ, and is defined formally as

$$p(\mathbf{D}_{\text{rep}}) = \int p(\mathbf{D}_{\text{rep}}|\theta)\pi(\theta)d\theta. \tag{5}$$

Box's p-value is defined as the following:

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

This comment, in addition to Reviewer #2 [Comments 3] led to the creation of Appendix C "Step-by-Step Implementation Guide" to address how exactly the formulas used can be applied in a trial. This includes references to how these equations are utilized in a trial:

3. Sequentially monitor the clinical trial:

   (a) Iteratively conduct monitoring according to (3). If the adaptive monitoring prior is used, then compute the mixture weight ω using (7) at each iteration, which is constructed using the computation of Box's p-value (6) which in turn uses the predictive distribution of the data (5).

[Comment 4]

- P16,L13,Section2.2.4:
  Do you mean "mode $\pi(\eta|\theta = \theta_0) = \eta_0$"?

[Response by Authors]

The authors agree that this is a vague expression. To make this statement consistent with the existing notation for the mode of a probability distribution first used in Section 2.1.3, the following change was made:

$\text{mode}(\eta|\theta = \theta_0) = \theta_0$ was changed to $argmax_\eta \pi(\eta|\theta = \theta_0) = \eta_0$.

Reviewer 2

Comments to the Authors:

This article provides a Bayesian framework for sequential monitoring of clinical trials using external data. It defines skeptical and enthusiastic priors in the context of Bayesian hypothesis testing and proposes a two-component mixture prior to combine these two types of priors. Weight of each component is determined based on an assessment of prior data conflict, for example, a higher weightage will be given the skeptical component if the observed data are not compatible with the enthusiastic prior. Section 2 of the article presents a thorough discussion on the elicitation of prior and a computation strategy to determine the weights.

Overall, the article is interesting but there are scopes of further improvement. My comments are provided below:

[Comment 1]
Authors provide a very limited discussion on the existing Bayesian approaches for sequential monitoring of clinical trials using noninformative prior and other approaches of elicitation of informative prior (such as power prior). It will benefit the readers if the authors discuss how their proposed method differs from other existing approaches through examples/simulation study.

[Response by Authors]

A comparison of our method with a Bayesian approach using a noninformative prior is now given in the new Section 3.3 “Comparison to Single Analysis with Non-Informative Prior.” (See Reviewer #2 [Comment 2]).

A reference comparing a feature of our method to other approaches of eliciting an informative prior is given in the discussion:

the case in pediatric settings). Other techniques for creating a dynamic borrowing prior for monitoring purposes would require manual verification that the information contributed by the prior is consistent with the residual uncertainty that must be present to justify the trial to begin with. For example, using a power prior (Ibrahim & Chen 2000) or a normalized power prior (Duan et al. 2006) would require that the value of the borrowing parameter (or its distribution) be chosen such that analysis of new data does not result in a compelling demonstration of efficacy in cases where the new data suggest otherwise. Thus, the proposed

Existing Bayesian approaches to sequential monitoring are discussed in the introduction with cited works including Spiegelhalter et al. 1993 and Jennison & Turnbull 2000. Comparison

to existing Bayesian approaches to sequential monitoring is provided by using the default skeptical prior for determination of treatment efficacy in Section 3.1 “Single-Arm Trial with Binary Endpoint.” Articles cited that discuss operating characteristics of existing Bayesian approaches include Psioda & Ibrahim 2018, Kopp-Schneider et al. 2020, etc.

[Comment 2]
This paper notes that strict type 1 error control may not be achievable when prior information is incorporated into the analysis. I think it will be useful for the readers if the authors provide a more detailed discussion on this topic. I recommend the authors to consider performing a simulation study, comparing the operating characteristics based on a standard design with noninformative prior and the proposed design using the mixture prior.

[Response by Authors] The authors provide a detailed discussion of this topic along with results from the suggested simulation study in the new Section 3.3 “Comparison to Single Analysis with Non-Informative Prior.”

3.3 Comparison to Single Analysis with Non-Informative Prior

We consider a simplified design without interim analyses that compares the use of the adaptive monitoring prior to a non-informative prior for the determination of treatment efficacy. The sample size of 100 patients and randomization scheme are the same as described in Section 3.2. The non-informative prior, denoted by $\pi_{NI}(\theta)$, is a uniform prior over the joint parameter space of (θ, η) so that the distribution of the risk difference θ and placebo response rate η are both marginally uniform. As was done in Section 3.2, the placebo response rate η is fixed at 0.39, and different possible values of the risk difference θ are considered.

For comparison purposes, we modify the skeptical and enthusiastic components which contribute to the adaptive monitoring prior to each having a marginally uniform distribution on η. We note that it is possible to achieve a simultaneous increase in power and preservation of nominal type I error rate as compared with a non-informative prior on the treatment effect if there is still an informative prior providing accurate information on nuisance parameters (e.g., see Psioda et al. (2018)), but that is not the focus in this example. In the following analyses, we also consider comparisons with a non-informative prior with a modified threshold for compelling evidence (referred to as modified critical value method in Psioda & Ibrahim (2018)) to identify any differences in null hypothesis rejection rates beyond a simple rescaling of the rejection rate at $\theta = \theta_0$.

Figure 6(A) demonstrates that the null hypothesis rejection rate using a fixed weight mixture prior with $\omega = 0.8$ in (4) is uniformly greater than the non-informative prior, including at $\theta = 0$. However, if we modify the threshold for compelling evidence in (2) so that the rejection rates at $\theta = 0$ are equal, then the rejection rates for all values of θ become identical, consistent with findings from Psioda & Ibrahim (2018). Although possible to simply modify the critical value used for hypothesis testing within a traditional design framework, we do not advocate for this approach. By doing so, one is implicitly deciding what level of evidence the pediatric data provide on their own, whereas a compelling demonstration of treatment effectiveness after synthesis of all pertinent information still requires additional post hoc evaluation. Instead, the proposed approach focuses on the end result, whether or not there is a compelling demonstration of treatment efficacy once all evidence has been synthesized, and provides a clear and rigorous framework for synthesis that balances information borrowing with the need to act sensibly in the presence of prior-data conflict.

Figure 6(B) demonstrates that the adaptive monitoring prior has a higher null hypothesis rejection rate at $\theta = 0$, and a higher rejection rate for values of θ less than 0.3. However, if we modify the threshold for compelling evidence in (2) so that the rejection rates at $\theta = 0$ are identical, then the rejection rates for the adaptive monitoring prior are higher for the adaptive monitoring prior for fewer values of theta, in particular for values of θ greater than 0 and less than 0.19, an interval containing effects much larger than those best supported by the adult data. This showcases desirable operating characteristics of the adaptive monitoring prior as compared to a non-informative prior, even when the required evidence threshold for analysis with the non-informative prior is modified (i.e., lowered).

[Comment 3]
I recommend the authors to provide a step-by-step algorithm for implementing their approach. This will be useful for the practitioners.

[Response by Authors]

The authors added Appendix C "Step-by-Step Implementation Guide" to address this comment.

C: Step-by-Step Implementation Guide

Below are step-by-step instructions for implementing this method in the context of a superiority trial.

1. Identify parameters for the trial design:

   (a) Specify null treatment effect θ_0 which is used to define the null and alternative hypotheses H_0 and H_1.

   (b) Specify threshold for a compelling demonstration $1 - \epsilon$.

   (c) Specify the plausible, clinically meaningful value for the treatment effect θ_1.

2. Create monitoring priors:

   (a) Choose the prior shape for the skeptical and enthusiastic monitoring priors $\pi_S(\theta)$ and $\pi_E(\theta)$ (e.g. choose k as described in Section 2.2.1). Our recommendation is to use a concentrated specification (i.e. $k = 1.5$) for $\pi_S(\theta)$ and a default specification (i.e. a normal distribution which has asymptotic justification as belief arriving from a hypothetical dataset) for $\pi_E(\theta)$.

   (b) Solve for the parameters in the generalized normal distributions $\pi_S(\theta)$ and $\pi_E(\theta)$ as described in Section 2.2.1 and Appendix B. These parameters are determined by the quantities provided in Step 1 and Step 2(a). The code for this paper shows how these parameters were computed.

   (c) Repeat Steps 2(a,b) for nuisance parameters as described in Section 2.2.4.

   (d) If the adaptive monitoring prior $\pi_{AE}(\theta)$ is to be used, specify the minimum possible mixing weight δ assigned to the skeptical prior. Our recommendation is $\delta = 0.1$ so that a modest weight of at least 0.1 is always given to the skeptical prior.

3. Sequentially monitor the clinical trial:

   (a) Iteratively conduct monitoring according to (3). If the adaptive monitoring prior is used, then compute the mixture weight ω using (7) at each iteration, which is constructed using the computation of Box's p-value (6) which in turn uses the predictive distribution of the data (5).

[Comment 4]
In my opinion, Section 5 of the paper can be improved by the authors taking a more critical view of their writing, sharpening the arguments, and including recommendations for the practitioners.

[Response by Authors] The authors removed text from the discussion that was redundant with previous exposition, and focused the discussion on two main conclusions: the requirement that observed data must demonstrate some degree of efficacy on their own to justify stopping

enrollment early, and a comparison with a published post-hoc Bayesian analysis using a hierarchical model. A reference to Appendix C which includes an implementation guide for practitioners was included.

In taking a more critical view of our writing, we removed statements that were vague or not fully supported by the results of the paper, such as the following sentence referring to concepts such as “overwhelming treatment benefit,” “in the more likely scenario,” “some evidence of benefit,” and “reasonable compatibility” was removed since those concepts were not defined rigorously:

> 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.
