A Bayesian Adaptive Design for Clinical Trials with Composite Outcomes and Low Event Rates

immediate

Abstract

We propose a Bayesian adaptive design for clinical trials where the outcome of interest is the risk of a relatively rare disease or condition and the incidence rate varies according to different operational definition of events. The proposed design incorporates different case definitions of the outcome of interest that vary in stringency. Composite stopping rules are proposed according to Bayesian superiority and futility results. A variety of stopping rules, and design configurations are compared through an extensive simulation study.

Keywords: composite outcomes, stopping rules, futility, superiority, simulations

1 Introduction

Bayesian adaptive designs for clinical trials (**??**) have become popular over the recent years due to the flexibility and efficiency that they offer over conventional fixed size randomized clinical trials (**???**). These designs can be considered a sequential decision process where adjustments to the trial design components may be made according to the accumulating evidence at preplanned interim analyses. Use of stopping decision rules is facilitated through sequential posterior updates within the Bayesian framework. Stopping decision criteria are defined based on Bayesian probability statements whose validity is not affected by small sample sizes and repeated testing.

Adaptive designs are especially appealing in situations where a considerable amount of uncertainty is associated with the underlying assumptions, namely the effect size and baseline distribution of outcomes. We specifically consider the case where the outcome of interest is an event with a relatively low risk of occurrence at baseline, and where there exist different case definitions of the event, which result in varying baseline risk assumptions. For example, consider neonatal sepsis as the primary outcome of interest. Neonatal sepsis is a clinical syndrome characterized by systemic inflammation, hemodynamic instability and multi-organ dysfunction that is presumed or confirmed to be caused by a systemic invasive bacterial, viral, or fungal infection, and which carries a high risk of morbidity and mortality (**??**). Diagnosis of neonatal sepsis is challenging because of the subtle and protean manifestations of serious illness in young infants. The variability in diagnostic terms and lack of gold-standard case definitions for neonatal sepsis make it difficult to compare incidence, severity, etiology or outcomes of disease across studies and settings (**?**). Sepsis, culture-confirmed sepsis and culture-negative sepsis are widely used in clinical practice and often reported in the literature, but definitions for these syndromes have varied widely (**????**).

We consider a clinical trial with the primary goal of preventing neonatal sepsis. Table 1 contains three case definitions for the event of interest that decrease in stringency from top to bottom. More permissive case definitions correspond to higher base risks of the event. While use of a more stringent definition for the primary outcome may be of interest for clinical precision, the corresponding low event rates can result in low statistical power in detecting the effect of an intervention in a clinical trial setting. In addition to the variability in case definitions, reliable estimates of the baseline risk of sepsis according to each of these definitions are difficult to obtain. This uncertainty makes selecting the primary event definition challenging.

JWCOMMENT: Above the words outcome and event are interchanged (primary outcome, primary event, event of interest, outcome of interest). Is there any benefit to keeping these consistent? There are three case definitions for the primary outcome/event of interest.ENDCOMMENT

In this paper, we propose a flexible Bayesian adaptive design where the primary outcome

Outcome	Definition
Stringent (s)	Non-injury related death Blood culture confirmed sepsis Possibly urine culture confirmed sepsis
Moderately permissive (p_1)	Middle ground between s and p_2 definition of s + Physician confirmed Removal of vague symptoms allowed for p_2/pSBI (e.g. remove fast breathing)
Highly permissive (p_2)	Panigrahi outcome or similar/WHO definition definition of p_1 + clinical sepsis, pSBI Possibly diagnosed by community health worker at onset rather than physician upon hospital visit

Table 1: Description of composite outcomes

is monitored with respect to any number of available (or of interest) case definitions. Early stopping decisions can be defined based on a combination of case definitions. For example, the decision of stopping early for efficacy is made according to the most stringent case definition, while the decision of stopping early for futility may be made with respect to a more permissive outcome definition. The operating characteristics of the proposed design for variations of the stopping rules, as well as for a variety of other design parameters and assumptions, are explored through an extensive simulation study which will be used as a guide to inform the final clinical trial design.

The remainder of the paper is organized as follows: we describe the Bayesian adaptive trial design in Section 2; Section 3 includes the simulation study which explores a variety of design operating characteristics; a discussion of potential deviations from the proposed approach is provided in Section 4.

2 Design

Consider a two-arm clinical trial to assess the effect of intervention A on decreasing the risk of a relatively rare syndrome such as sepsis, which has the three case definitions described in Table 1. Equally distanced interim analyses are planned according to the accumulating number of enrollees for whom the primary outcome is observed. For example, the first

interim analysis is performed when the primary outcome is observed for 1000 participants, the second interim analysis is performed when the outcome is observed for 2000 participants and so on. The frequency of interim looks at the data is determined according to the design operating characteristics explored under various simulation scenarios.

JWCOMMENT: From above, the batch size is controlled based on the number of "outcomes" ("for whom the primary outcome is observed")? This sounds to me as if each interim analysis occurs only after 1000 (2000, 3000, etc.) sepsis events have occured... We may want to clarify these interim analyses aren't based on the number of "events" of sepsis, but on us being able to observe the "outcome" (sepsis: yes or no") for 1000, 2000, 3000 patients. The difference is subtle, but I think for someone who has no insider knowledge of this trial, this could be confusing. This is re-iterated in Section 3 (not basing interim looks on the number of events but rather the number of participants. I have a comment down there as well). You write elsewhere (Section 2) looks can be planned when "responses are available for a certain number of participants". This might be nice language to use here. The equations below clarify this dichotomous outcome, but could be motivated above.ENDCOMMENT

The interim analyses are performed within the Bayesian framework. Given that the primary outcome is dichotomous, the likelihood can be written as,

$$l(\mathbf{y} \mid \pi_1, \pi_2) = \prod \left(\pi_1^{y_n}(1-\pi_1)^{1-y_n}\right)^{1-a_n} \left(\pi_2^{y_n}(1-\pi_2)^{1-y_n}\right)^{a_n} \tag{1}$$

where π_1 is the probability of the event under the control arm, π_2 (where $\pi_2 < \pi_1$ under the alternative hypothesis) is the probability of the event under the intervention arm, and y_n and a_n denote the outcome and arm assignment for participant n. If an event is observed for participant n, $y_n = 1$, otherwise, $y_n = 0$. If participant n is assigned to the intervention arm $a_n = 1$, otherwise $a_n = 0$.

JWCOMMENT: For the likelihood equation above, I don't know the convention in papers, but I've seen lowercase "l" used for the log-likelihood and uppercase "L" used for likelihood. Not sure if this matters here since it's clear we are discussing the likelihood.ENDCOMMENT

The quantity of interest is the relative risk (RR) of the event among those who receive the intervention to those in the control group,

$$RR = \frac{\pi_2}{\pi_1}.$$

The null and alternative hypotheses are formulated, respectively, as,

$$H_0 : RR \geq 1, \qquad H_A : RR < 1.$$

Therefore, the stopping decision criteria are defined based on the posterior probability of the alternative hypothesis, which is derived from the posterior distribution of the RR, which is obtained from the posterior distributions of π_1 and π_2 in a Beta-binomial model. At any interim analysis, efficacy is concluded if the posterior probability that the RR is below 1 is higher than a specified threshold,

$$P\left(RR < 1 \mid \mathbf{y}\right) > t_s. \tag{2}$$

Similarly, futility can be defined based on the posterior distribution of RR. For example, one may consider $RR > 0.9$ $(= RR_f)$ an unimportant effect and therefore define futility with respect to the posterior probability of such a result,

$$P\left(RR > RR_f \mid \mathbf{y}\right) > t_f \tag{3}$$

where $RR_f < 1$ is referred to as the futility boundary and t_f as the futility probability threshold. While unlikely to be observed in practice, superiority and futility as defined above are theoretically not mutually exclusive. Therefore, to be rigorous, we define futility as

$$P\left(RR > RR_f \mid \mathbf{y}\right) > t_f \quad \text{and} \quad P(RR < 1 \mid \mathbf{y}) \leq t_s. \tag{4}$$

2.1 Composite stopping rules

The posterior probability of the RR may be obtained according to any of the case definitions of events. Denoting the vector of outcomes with respect to the case definitions presented in Table 1 by $\mathbf{y}_s$, $\mathbf{y}_{p_1}$ and $\mathbf{y}_{p_2}$, we may define one possible set of stopping criteria at the interim analyses as follows:

- Stop for superiority with respect to $\mathbf{y}_s$, i.e. if $P(RR < 1 \mid \mathbf{y}_s) > t_s$;
- Stop for futility with respect to $\mathbf{y}_{p_1}$, i.e. if $P(RR > RR_f \mid \mathbf{y}_{p_2}) > t_f$.

The advantage of a composite decision rule is being able to prioritize certain decisions with respect to their importance. For example, the above stopping rule requires a significant amount of evidence with respect to the most stringent outcome to conclude efficacy. On the other hand, futility may be assessed with respect to a more permissive outcome since, if the trial is futile with respect to a more probable event, it is likely futile according to a more stringent outcome. However, meeting the decision criteria with respect to the stringent outcome requires a much larger sample size.

A variety of stopping rules are explored via simulations. In Section 3, we present results for stopping criteria which are defined according to the moderately permissive case definition p_1 only.

2.2 Sample size and frequency of interim analyses

In an adaptive design that allows for early stopping, the final trial size is a random variable. Sample size considerations are therefore different from those in fixed size trials. Factors that affect the final trial size include the number and spacing of interim analyses. Frequent interim analyses result in more flexibility, since this allows for more opportunities to make stopping decisions. However, a greater chance of stopping corresponds to a higher probability of obtaining a false positive result.

Another challenge in planning interim analyses is the follow-up time, the time it takes to obtain the primary outcome. It is common to plan interim analyses when responses are

available for a certain number of participants. This approach is preferred from a statistical power analysis perspective since it keeps the interim sample sizes fixed. However, it results in uncertainty in the actual timing of interim analyses since it depends on enrolment progress. Planning interim analyses according to calendar time may be appealing and appear more straightforward from a planning perspective, but it can result in unequal spacing of interim analyses and variable interim sample sizes that have not been accounted for in a power analysis. We take the former approach and keep the interim sample sizes fixed.

Although the final trial size is a random quantity in an adaptive design, specifying a maximum allowable sample size prevents prolongation of the trial beyond budget constraints. If the stopping criteria are not met at any of the planned interim analyses, the trial is stopped when the specified maximum allowable sample size is reached.

Therefore, the frequency of interim analyses is specified by the maximum allowable sample size and the interim sample size. For example, if a total of 12,000 participants is the largest affordable trial size, and interim analyses are to be performed when the outcome is available in batches of size 3,000, the design will allow up to three interim analyses in addition to the final analysis. The optimal frequency of interim looks is explored in the simulation study.

2.3 Design operating characteristics

Similar to fixed trial designs, Bayesian adaptive designs are assessed with respect to their operating characteristics including, but not limited to, power and false positive rate. The definition of power and false positive rate for Bayesian adaptive trials is, in principle, unchanged. However, statistical significance is defined with respect to posterior probabilities whose sampling distributions are generally not known, resulting in a power function that is not analytically tractable.

Power is defined as the probability of concluding efficacy, at any of the preplanned interim looks or at the final analysis, for a given value of RR under the alternative hypothesis, e.g. $RR = 0.8$. Analogously, the false positive rate is the probability of concluding efficacy, at any of the preplanned interim looks or at the final analyses, for the value of RR under the

null hypothesis, i.e. $RR = 1$.

In addition to power and false positive rate, other operating characteristics relevant in a Bayesian adaptive trial might include: the probability of concluding futility under the null and alternative hypotheses, the probability of stopping early or stopping at a specific interim look, and probability of reaching the maximum affordable sample size. We emphasize that these probability statements should not be confused with the posterior probability statements that are the basis of decision making. In other words, power analysis for Bayesian adaptive trials involves evaluating functions that are defined as two layer probability statements; the inner layer is driven from the posterior distribution of the model parameters, while the outer layer is obtained from the sampling distribution of the test statistics, which are the posterior probabilities of superiority or futility.

Therefore, to evaluate operating characteristics for a variety of design parameters and assumptions, extensive simulation studies are required. In the following section, we describe the underlying assumptions, design configurations and data generating procedures for a typical simulation study required for the design of Bayesian adaptive trials.

3 Simulation study

3.1 Inputs

As explained above, the operating characteristics of the proposed design need to be investigated through simulations. An effective simulation study is one that explores a wide range of assumptions and design parameters, especially those that parameterize the decision rules. These can be considered as the simulation inputs. The set of assumptions and parameters that constitute our simulation scenarios are described below.

Considering that the outcome of interest has a relatively low rate of occurrence, especially with respect to some case definitions, the assumed control event rate (CER) plays an important role in power and other important operating characteristics. We consider three sets of assumptions for each case definition: a worst case scenario corresponding to very low

scenario	π_s	π_{p_1}	π_{p_2}
worst	0.002	0.02	0.09
base	0.01	0.05	0.12
best	0.015	0.08	0.15

Table 2: Control event risk (CER) assumptions.

event probabilities, a base case scenario representing relatively moderate event probabilities, and a best case scenario that represents the largest realistic event probabilities that can be assumed. Table 2 shows these three sets of values for CER.

JWCOMMENT: For CER, would you like "control event rate" or "control event risk"? I've changed the "risks" to "rates" to match the incidence rate mentioned in the abstract, though now I'm thinking that may have been premature of me (these are defined over a set period of time for neonatal sepsis, 90 days, right?)

In addition, a range of RR values, including the "no effect" assumption, are explored. Specifically, $RR \in (1, 0.9, 0.8, 0.6, 0.4)$.

The superiority and futility probability thresholds and futility bounds are important design parameters whose optimal values should be specified to achieve desired operating characteristics. The simulation design should therefore explore a range of values for these parameters. Our simulation study includes three values for the superiority probability threshold, $t_s = 0.95, 0.975, 0.99$ and two values for the futility bound $RR_f = 0.1, 0.2$. The futility probability threshold is held fixed at $t_f = 0.99$.

In addition, the frequency of interim looks is varied by performing the analyses at batches of 1,000, 2,000 and 3,000 observed outcomes. The maximum allowable sample size is held fixed at 12,000.

JWCOMMENT: Observed "outcomes", might clarify here with "observed responses of 1,000 participants" or something that matches the language above and below in Section 3.

As for the stopping criteria, the following set of rules are considered:

1. Stop for superiority with respect to s and for futility with respect to p_2

2. Stop for superiority and futility with respect to p_2

3. Stop for superiority and futility with respect to p_1

The combination of the above configurations results in 810 simulation scenarios. The trial is simulated for 500 iterations for each of these scenarios to estimate the design operating characteristics. We present results only for a subset of these simulation scenarios in the manuscript and provide the complete results in Appendix A.

3.2 Correlation structure and data generating process

Data for the simulation study are generated according to the underlying assumptions regarding the three composite outcomes described in Table 1. Based on the case definitions, the highest event rates are expected to be observed under the highly permissive outcome Y_{p_2}, while the strict definition of the stringent outcome Y_s results in low event rates. In addition, every case that is considered an event under the definition of the stringent outcome is also an event under the more permissive outcomes, and likewise, any event under Y_{p_1} is an event under Y_{p_2}. Suppose that the event rates for the three outcomes, Y_s, Y_{p_1} and Y_{p_2}, are denoted by π_s, $\pi_{p_1}$and π_{p_2}, respectively. The outcomes with respect to the three case definitions are generated as follows,

$$
\begin{aligned}
&Y_s \sim \text{Bernoulli}(\pi_s), \\
&Z_{p_1} \sim \text{Bernoulli}\left(\frac{\pi_{p_1} - \pi_s}{1 - \pi_s}\right), \qquad Y_{p_1} = Y_s + (1 - Y_s)Z_{p_1}, \\
&Z_{p_2} \sim \text{Bernoulli}\left(\frac{\pi_{p_2} - \pi_{p_1}}{1 - \pi_{p_1}}\right), \qquad Y_{p_2} = Y_{p_1} + (1 - Y_{p_1})Z_{p_2}.
\end{aligned} \tag{5}
$$

3.3 Outputs

The simulation output includes a variety of measures that guide the selection of design parameters. A number of these measures were discussed in Section 2.3 as design operating characteristics. In the following, we provide a complete list of measures generated by the simulation study.

Power is estimated as the proportion of times that a statistically significant result was obtained under the scenarios where $RR < 1$, regardless of the trial being terminated early or reaching the maximum sample size. **False positive rate** is estimated in the same fashion, but for the case that $RR = 1$. While power and false positive rate refer to the probability of significance with respect to the primary outcome, i.e., the case definition with respect to which the corresponding stopping criteria are defined, these probabilities are estimated for all three case definitions for each simulation scenario.

Probability of futility is estimated as the proportion of times that futility is concluded according to the futility criterion defined in (4), either at an interim look or at the final analysis. Probability of concluding futility becomes negligible for smaller values of RR and therefore cannot be estimated with precision using simulations.

Probability of stopping early is estimated as the proportion of times that the trial is terminated before reaching the maximum sample size, due to either a superiority or futility result. This probability is further broken down into the **probability of stopping early for superiority** and **probability of stopping early for futility**, the latter of which suffers from lack of precision in some cases. Similarly, **probability of reaching maximum sample size** can be obtained as the proportion of times that the trial could not be stopped at any interim look, i.e.

$$1 - P(\text{stopping early})$$

Expected sample size at trial termination is obtained as the average of trial sizes for each simulation scenario. Note that the trial size at the time of termination is a multiple of interim sample sizes. Therefore, the expected sample size represents the trial size that is expected to arise on average and not in a single trial.

Sampling distribution of relative risk estimates is obtained as the distribution of posterior means of relative risks. The posterior mean of RR is estimated as the average of a sample of RR values drawn by sampling the risks under each arm from the respective posterior distributions. The estimates are obtained for the three case definitions for each simulation scenario.

P-values for a Fisher's exact test are obtained for every iteration. While the analysis and decision making primarily relies on posterior distributions and probabilities, the result of a frequentist test can provide insight regarding the performance of the Bayesian adaptive design.

3.4 Results

The results of the simulation study, including all simulation scenarios and outputs discussed above, are provided in Appendix A. In this section, we provide select graphs for a design that employs the third set of stopping rules – stopping for superiority and futility with respect to p_1 – and allows for interim analyses to occur after the collection of responses from every 3,000 participants. Key conclusions that can guide specification of design components are discussed.

JWCOMMENT: Is the 3,000 right? I only ran/uploaded the batch sizes of 1,000 and 2,000 to Github.ENDCOMMENT

Figure 1 shows the estimated power for the three outcomes across different simulation scenarios arising from the combination of values for RR, CER and superiority thresholds. The futility threshold is held fixed at 0.1, as it appeared to not make a difference in the estimated power. As expected, power decreases as the effect diminishes, i.e., RR gets closer to 1. The rate of decline is higher under lower CER, with practically no power for the stringent outcome s. A larger superiority probability threshold, in general, results in lower power. However, the difference is visible only under weaker effects ($RR = 0.8, 0.9$).

Based on these observations, it would be unrealistic to power the trial with respect to the stringent case definition. For an RR of 60% or smaller, the design has sufficient power with respect to either of the more permissive outcome definitions. For smaller effect sizes, the choice between p_1 and p_2 can make a considerable difference, especially if a more stringent superiority stopping rule is employed.

The estimates of false positive rates are presented in Figure 2. An immediate observation is that the probability of a false positive result is higher for p_1, i.e., the outcome with respect

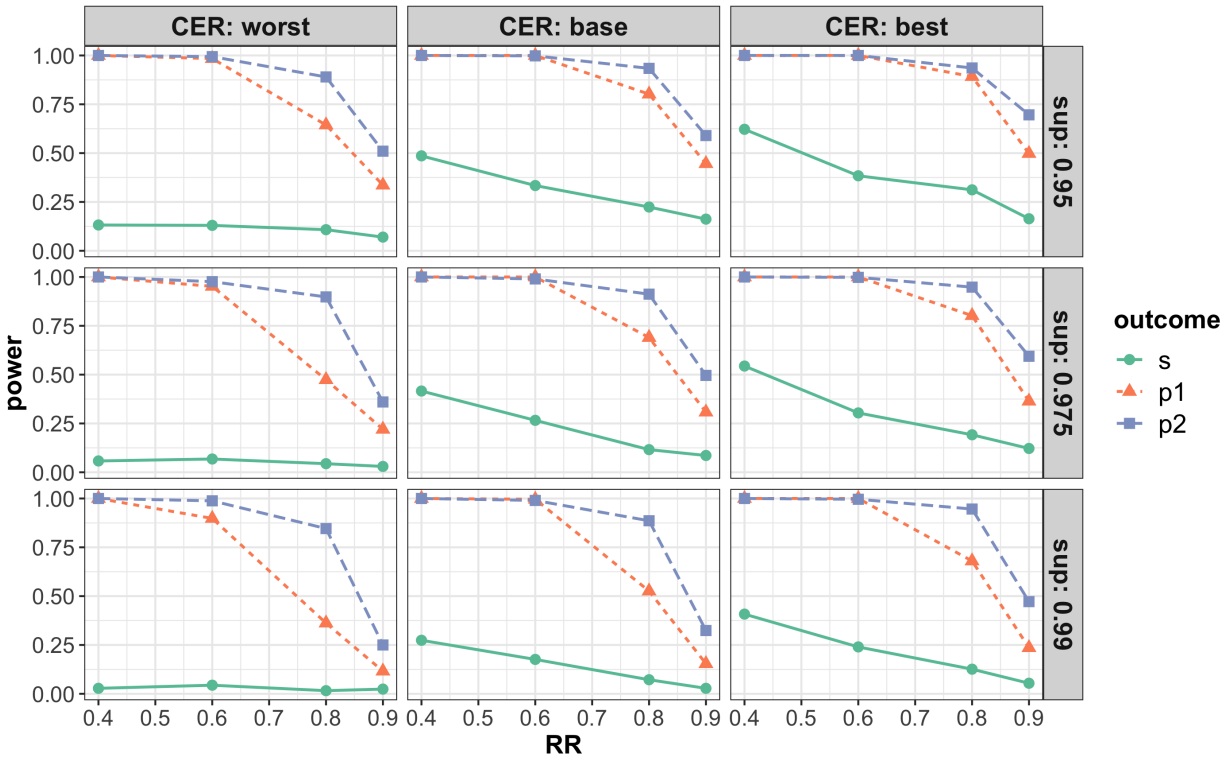

Figure 1: Estimated power across simulation scenarios. On the X axis are the increasing values of RR (decreasing effect size) and on the Y axis is the estimate of power. The column panels represent the three assumed sets of values for CER presented in Table 2 and the row panels represent three values for the superiority probability threshold. The colors/line types/dot shapes refer to the three outcome definitions.

to which a superiority stopping decision is made. The inflation of false positives for p_1 is the result of interim stopping for superiority at "random highs" of probability of superiority for this outcome. Note that interim false positives are as likely for the other two outcomes but the trial is not stopped according to these outcomes. Only false positives at the final analysis are counted for s and p_2. Clearly, increasing the superiority probability threshold helps control the type I error rate at a lower level. A superiority threshold of 0.99% guarantees a false positive rate below 5% across simulation scenarios. A higher futility bound is expected to decrease the chance of a false positive result, as it increases the chance of stopping the trial early for futility. This appears to be the case for p_1. However, since superiority is prioritized to futility, the change in futility criterion does not affect the false positive rate in the other two outcomes. We emphasize that precisely estimating false positive rates through

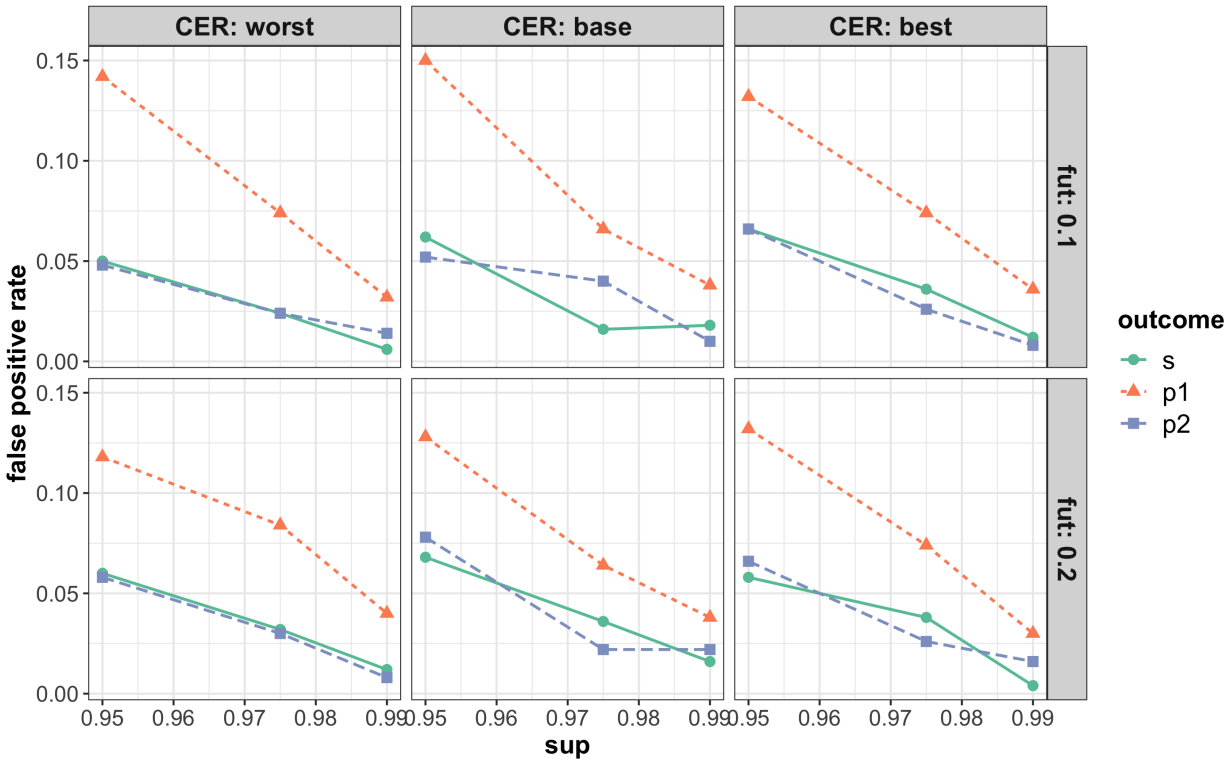

Figure 2: Estimated false positive rate across simulation scenarios. On the X axis are the increasing values of superiority probability threshold and on the Y axis is the estimate of type I error rate. The column panels represent the three assumed sets of values for CER presented in Table 2, the row panels represent two values for the futility bounds. The colors/line types/dot shapes refer to the three outcome definitions.

simulations is challenging and the presented graph should not be overinterpreted.

Figure 3 shows the expected sample size at trial termination over different simulation scenarios. The trial size increases for smaller assumed effect sizes. In the absence of a true effect, the trial is either stopped early due to a false positive result or for enough evidence for futility. Therefore, in this case, the futility bound plays a role in trial size; a higher futility bound increases the chance of stopping early for futility and thus decreases the expected trial size. A more stringent superiority criterion (higher superiority probability threshold) results in larger trials on average, as the chance of stopping early is lowered.

Finally, Figure 4 presents the estimated probability of concluding futility for relevant simulation scenarios. The simulation results are filtered to include only small or zero effect assumptions, i.e., $RR = 0.8, 0.9, 1$, since the probability of concluding futility is negligible

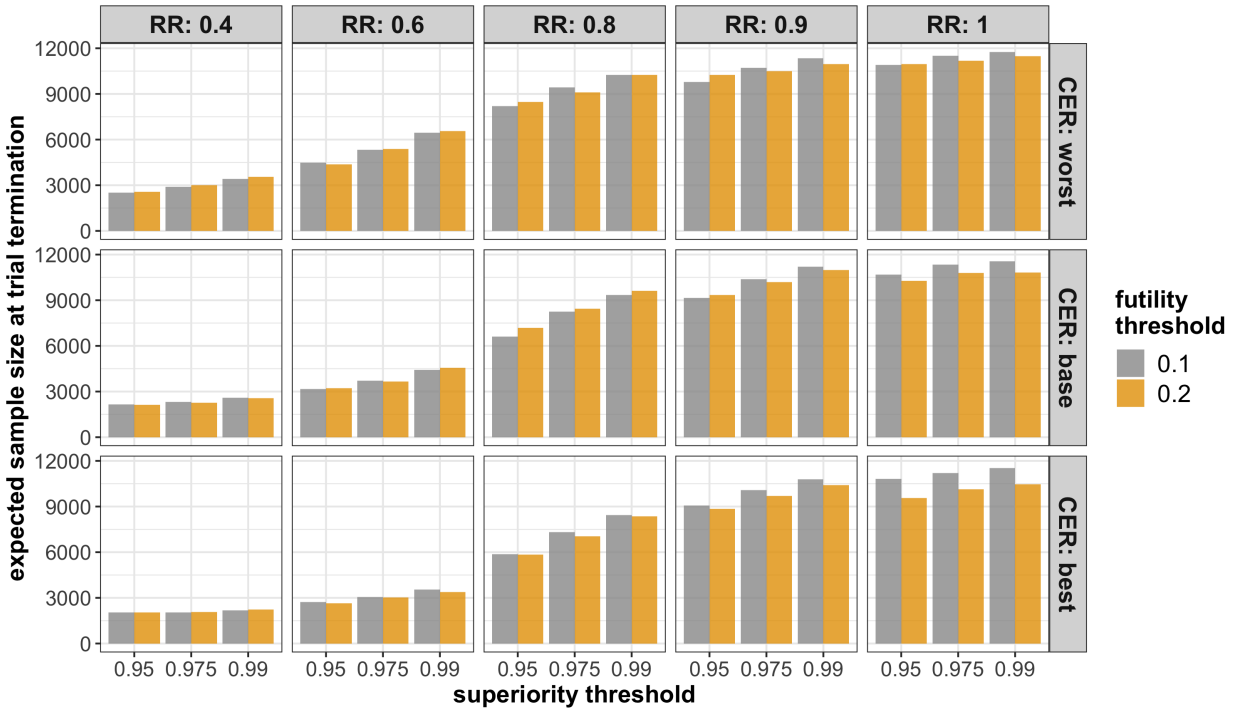

Figure 3: Expected sample size at trial termination across simulation scenarios. On the X axis are the increasing values of the superiority probability threshold and on the Y axis is the average trial size. The column panels represent the assumed RR values with RR: 1 representing no effect. The row panels represent the three assumed sets of values for CER presented in Table 2 and the color of the bars refers to the two values for the futility bounds.

in other cases. Clearly, using a higher futility bound results in higher chance of concluding futility. The futility rule is sufficiently stringent such that the chance of concluding futility is only non-negligible when the trial is in fact futile, i.e., $RR \leq RR_f$. The probability of stopping for futility is highest under the largest control event rates and is at best 30%.

4 Discussion

We have proposed a Bayesian adaptive design with both futility and superiority stopping rules for the scenario where multiple case definitions are available for the event of interest, and where the event rates vary over these definitions. The challenge is to choose various components of the design and decision criteria according to a variety of design operating characteristics, each of which depends on at least a subset of design and decision parameters.

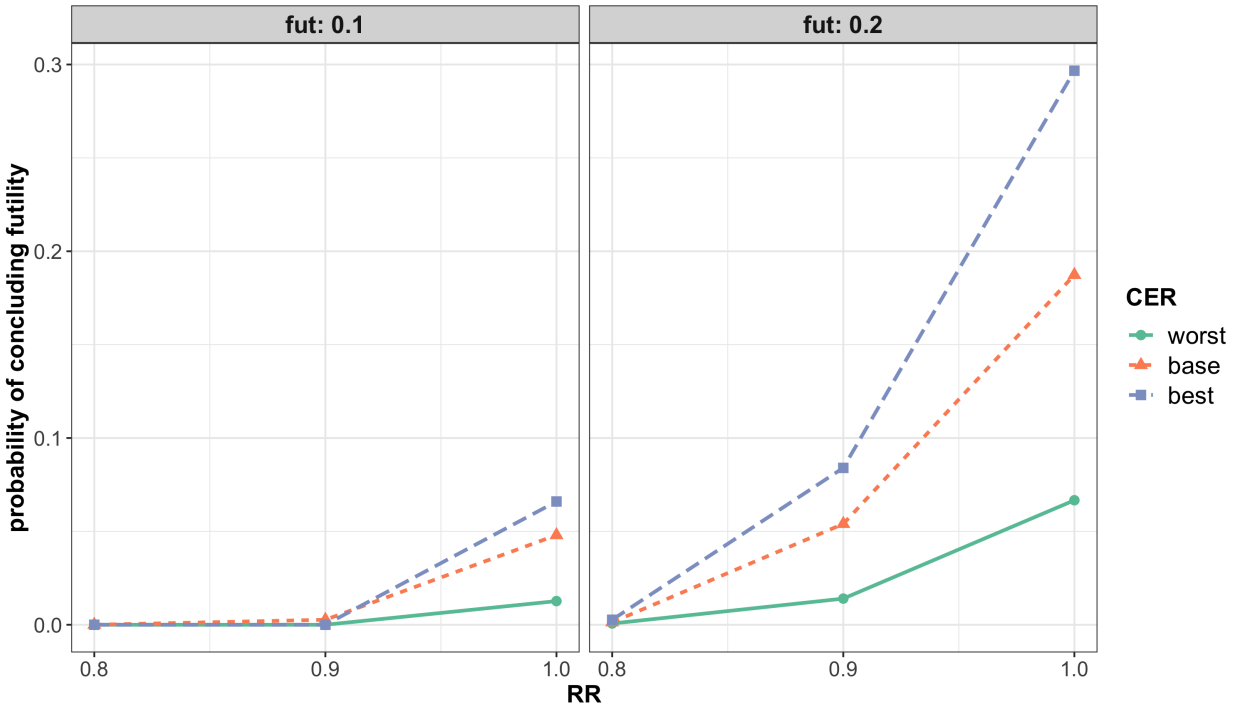

Figure 4: Estimated probability of concluding futility across simulation scenarios. On the X axis are the RR values filtered to small or no effect, and on the Y axis is the estimated probability of concluding futility. The panels represent the values of the futility bound. The colors/line types/dot shapes refer to the three assumed sets of values for CER presented in Table 2.

We showcase a simulation study that explores a variety of plausible options and the behaviour of design operating characteristics across a large set of scenarios. The goal is to provide the team of investigators with insight regarding the combination of assumptions, decision criteria and sample sizes that result in acceptable operating characteristics, and additionally, to choose a trial design which optimizes the balance between acceptable operating characteristics, feasibility and resources (time and cost).

The simulation study presented in this paper assumes the risk of each event, defined by their specific case definitions, are reduced equally by the treatment. However, this may not be a realistic assumption. For example, the treatment may reduce the risk of non-injury death but not the risk of sepsis diagnosed by any means. Under this hypothetical scenario, using p_1 or p_2 instead of s, results in a set of events whose risk of occurrence is unchanged by the intervention. Such an approach yields no additional power, but can result in the loss of

power by introducing noise on top of the event of interest. Therefore, case definitions with higher event rates should only be used where there is scientific justification for the relevance of the additional events.

Another assumption underlying the presented simulations is that the control event rates remain constant over time. In reality this assumption may not hold. Incorporating a time-varying control rate into the simulation study is straightforward. The challenge is to come up with reasonable assumptions for the functional path along which the rate varies. A possible alteration to the design that removes the requirement of making assumptions regarding the CER, is to schedule the interim analyses with respect to the number of events rather than number of participants. This approach, however, adds a significant level of uncertainty to the timing of interim analyses and the overall length of the trial. With low event rates, specifically, such an approach can result in an infeasible and seemingly endless trial.

JWCOMMENT: I've changed the "control event risks" to "rates" to stay consistent with the beginning of the paper. Is this ok? This number of events vs number of participants is what one of my first comments was about, when discussing the batch sizes of interim analysis (I think the word "outcome" was used up there, after 1000 or 2000 outcomes were observed). Perhaps we can use this wording at top "number of participants"? Or number of participants/responses assessed for outcome at 90 days post birth, or whatever the definition we are using for the trial.ENDCOMMENT

References