A Bayesian adaptive design for clinical trials of rare efficacy outcomes with multiple definitions

Shirin Golchi\textsuperscript{1}, James J Willard\textsuperscript{1}, Eleanor Pullenayegum\textsuperscript{2,3}, Diego G Bassani\textsuperscript{2,3,4,5}, Lisa G Pell\textsuperscript{4}, Kristian Thorlund\textsuperscript{6} and Daniel E Roth\textsuperscript{2,3,4,5}

Abstract

Introduction: Bayesian adaptive designs for clinical trials have gained popularity in the recent years due to the flexibility and efficiency that they offer. We consider the scenario where the outcome of interest comprises events with relatively low risk of occurrence and different case definitions resulting in varying control group risk assumptions. This is a scenario that occurs frequently for infectious diseases in global health research.

Methods: We propose a Bayesian adaptive design that incorporates different case definitions of the outcome of interest that vary in stringency. A set of stopping rules are proposed where superiority and futility may be concluded with respect to different outcome definitions and therefore maintain a realistic probability of stopping in trials with low event rates. Through a simulation study, a variety of stopping rules and design configurations are compared.

Results: The simulation results are provided in an interactive web application that allows the user to explore and compare the design operating characteristics for a variety of assumptions and design parameters with respect to different outcome definitions. The results for select simulation scenarios are provided in the article.

Discussion: Bayesian adaptive designs offer the potential for maximizing the information learned from the data collected through clinical trials. The proposed design enables monitoring and utilizing multiple composite outcomes based on rare events to optimize the trial design operating characteristics.

Keywords

Composite outcomes, stopping rules, futility, superiority, simulations

Introduction

Bayesian adaptive designs for clinical trials\textsuperscript{1,2} have become popular in recent years due to the flexibility and efficiency that they offer over conventional fixed-size randomized clinical trials.\textsuperscript{3–5} These designs can be considered sequential decision processes 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 the 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, and where there exist different case definitions of

\textsuperscript{1}Department of Epidemiology and Biostatistics, McGill University, Montreal, QC, Canada
\textsuperscript{2}Child Health Evaluative Sciences, Hospital for Sick Children, Toronto, ON, Canada
\textsuperscript{3}Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
\textsuperscript{4}Centre for Global Child Health, Hospital for Sick Children, Toronto, ON, Canada
\textsuperscript{5}Department of Paediatrics, University of Toronto, Toronto, ON, Canada
\textsuperscript{6}Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada

Corresponding author:
Shirin Golchi, Department of Epidemiology and Biostatistics, McGill University, 2001 McGill College, Suite 1200, Montreal, QC H3A 1G1, Canada.
Email: shirin.golchi@mcgill.ca
the event, which result in varying control group risk assumptions. This is a scenario that occurs frequently for infectious diseases in global health research.

In this article, we propose a flexible Bayesian adaptive design where the primary composite outcome is monitored with respect to any number of available and relevant case definitions. The novelty of the proposed design is that it allows for a flexible stopping rule where early stopping decisions may be made with respect to different outcomes to optimize efficiency. Even when stopping decisions are made according to the same outcome, monitoring the other outcomes and performing secondary analyses can yield insight toward the mechanism of the effect. 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 a simulation study.

The present article contributes to the literature on composite outcomes for clinical trials which remains scarce. In general, caution is advised in use of composite outcomes in clinical trials and specific guidelines are provided by the Food and Drug Administration (FDA) for design of clinical trials with multiple outcomes. Work has been done on providing a set of conditions for well-defined composite outcomes, such as the assumptions of common relative risk (RR) reduction across components which is an underlying assumption of our data generating models. Another example of relevant work is exploration of adaptive sample size re-estimation and population enrichment strategies for trials with composite outcomes and low event rates in the context of cardiovascular diseases.

The remainder of the article is organized as follows. In the next section, we provide a motivating example. Specifically, we focus on neonatal sepsis with multiple case definitions as an example of a composite outcome with hierarchical structure. We describe the Bayesian adaptive trial design and provide a simulation study which explores a variety of design operating characteristics. We conclude with a summary of results and a discussion of potential deviations from or additions to the proposed design.

Methods

Motivating 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 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 diagnoses in clinical practice and often reported in the literature, but definitions for these syndromes have varied widely.

For neonatal sepsis and sepsis-related mortality in a low-resource community setting, clinical diagnostic data (e.g. vital signs, laboratory parameters) are not consistently recorded for infants admitted to hospital with suspected sepsis. Composite sepsis outcome definitions in this context rely primarily on microbiological test confirmation of an infection with a pathogenic organism, physician diagnosis of sepsis or sepsis-like illness and initiation of empirical antibiotics, and/or community health worker (CHW) ascertainment of at least one of a set of clinical criteria that indicate possible serious bacterial infection and therefore mandate urgent referral for medical care.

Table 1 contains three neonatal sepsis case definitions that range in terms of stringency with respect to the likelihood that a recorded episode represents true sepsis or death. More permissive case definitions are expected to encompass episodes of viral infections that are generally less severe than bacterial sepsis and non-infectious conditions that have overlapping clinical features with sepsis (e.g. poor feeding). While use of a more stringent definition for the primary outcome may be of interest for clinical precision, the corresponding low event risks 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.

As discussed above, a set of decision rules that utilize multiple case definitions in a Bayesian adaptive design are proposed as a flexible approach. We consider a clinical trial where the primary objective is to establish effectiveness of an intervention in reducing the RR of sepsis with respect to either of multiple available case definitions. In the following, we outline a trial design appropriate for a primary outcome, such as neonatal sepsis. The nested, and therefore correlated, nature of the three composite outcomes arising from the three selected case definitions for neonatal sepsis is incorporated into the simulations.

Design

Consider a two-arm clinical trial to assess the effect of an intervention on decreasing the risk of a relatively rare syndrome, such as neonatal sepsis with the three
case definitions described in Table 1 compared to placebo. In such a trial, newborns are enrolled in the immediate postnatal period and then both passive and active surveillance for neonatal sepsis including a schedule of home visits by first-level health workers or research personnel referred to as “community health workers”, and documentation of any infant hospitalization. Equally distanced interim analyses are planned according to the accumulating number of enrollees who have completed the follow-up time period and for whom the primary outcome can be recorded. For example, the first interim analysis is performed when 1000 participants have completed follow-up time from enrolment in the immediate postnatal period to 60 days postnatal age, the second interim analysis is performed when 2000 participants have completed follow-up time, and so on. The frequency of interim looks at the data is determined according to the design operating characteristics explored under various simulation scenarios.

The interim analyses are performed within the Bayesian framework. The quantity of interest is the $RR$ of the event among those who receive the intervention to those in the control group

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

where $\pi_1$ and $\pi_2$ are the probabilities of the event occurring in the control and intervention arms, respectively. Note that the event may be defined according to either of the case definitions discussed in the previous section.

The null and alternative hypotheses are formulated, respectively, as

$$H_0 : RR \geq 1, \quad H_A : RR < 1$$

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 $RR$, obtained from the posterior distributions of $\pi_1$ and $\pi_2$ in a beta-binomial model with a flat prior Beta(1, 1) over the risk parameters. At any interim analysis, efficacy is concluded if the posterior probability of the alternative hypothesis is higher than a prespecified threshold, $t_e$

$$P(RR < 1 | y_i) > t_s$$

where $y_i$ is the accumulated data (observed outcomes) up to interim analysis $i$.

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

$$P(RR > RR_f | y_i) > t_f$$

where $RR_f$ is referred to as the futility bound and $t_f$ as the futility probability threshold. In theory, superiority and futility events are not mutually exclusive. Therefore, to be rigorous, we define futility as

$$P(RR > RR_f | y_i) > t_f \quad \text{and} \quad P(RR < 1 | y_i) \leq t_s$$

Figure 1 shows the above posterior probabilities. Specifically, the posterior probability of effectiveness and a clinically unimportant effect are illustrated as the area under the posterior density curve shaded blue and red, respectively, under three different scenarios. The dashed vertical line identifies the $RR = 1$ and the dotted line is drawn at the futility-bound $RR = RR_f = 0.9$. Figure 1(a) represents a scenario where efficacy may be

| Table 1. Composite outcome definitions for “neonatal sepsis” at different levels of stringency. |
|------------------------------------------|-----------------------------------------------|----------------|
| **Outcome** | **Criteria (examples)** | **Stringency** |
| Non-injury-related death and/or blood culture-confirmed sepsis | • Death due to any cause except if directly caused by physical trauma; and/or | Stringent $(s)$ |
| | • Blood and/or cerebrospinal fluid (CSF) culture positive for a pathogenic bacterial or fungal organism | |
| | • Physician decision to admit to hospital with a sepsis or sepsis-related diagnosis, and administration or physician intention to administer parenteral antibiotics for at least 5 days; and/or | Moderately permissive $(p_1)$ |
| | • Non-injury-related death or blood culture-confirmed sepsis (as defined above) | |
| Physician-confirmed clinical sepsis | • Presence of at least one sign of possible serious bacterial infection (poor feeding, lethargy, seizures, severe lower chest wall indrawing, fever, or hypothermia) based on assessment by a CHW; and/or | Highly permissive $(p_2)$ |
| Community health worker (CHW)-ascertained or Physician-confirmed clinical sepsis | • Physician-confirmed clinical sepsis, non-injury related death or blood culture-confirmed sepsis (as defined above) | |
| | • Presence of at least one sign of possible serious bacterial infection (poor feeding, lethargy, seizures, severe lower chest wall indrawing, fever, or hypothermia) based on assessment by a CHW; and/or | |

Note that the event may be defined according to either of the case definitions discussed in the previous section.
concluded with the superiority threshold of $t_s = 0.975$ since posterior probability of $RR<1$ is above $t_s$. In the scenario illustrated in Figure 1(b), the results are indecisive with insufficient evidence in favor of either efficacy or futility; however, the results seem promising in favor of efficacy. In the final hypothetical scenario in Figure 1(c), futility may be concluded with a futility probability threshold of $t_f = 0.99$ since the probability of $RR > 0.9$ is 0.992 and the probability of efficacy (0.07) is less than $t_s = 0.975$.  

Stopping rules utilizing multiple outcomes. 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 $y_s$, $y_{p_1}$, and $y_{p_2}$, we may define one possible set of stopping criteria at the interim analyses as follows:

- Stop for superiority with respect to $y_s$, that is, if $P(RR < 1 | y_s) > t_s$;
- Stop for futility with respect to $y_{p_2}$, that is, if $P(RR > RR_f | y_{p_2}) > t_f$.

Note that the above decision rules rely on a marginal rather than joint analysis of the three outcomes resulting from the three case definitions. In other words, the posterior distribution of $RR$ is obtained from the marginal distribution of each outcome that is a binomial distribution with the corresponding risk parameter.

The advantage of such a decision rule is the ability 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 and much less common 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 the simulations section, we present results for stopping criteria which are defined according to the moderately permissive case definition $p_1$ only.

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 by providing 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 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 creates uncertainty regarding the timing of interim analyses since it depends on enrolment progress. Planning interim analyses according to calendar time may 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.

An alternative approach, which is especially popular in settings where the event rate varies with time, is to schedule the interim analyses according to the number of events rather than number of participants. In the present setting, however, low event rates create a large amount of uncertainty regarding the timing of interim analyses and the duration of the trial which is considered a major disadvantage from a feasibility and cost standpoint. Therefore, we take the equal interim sample size approach that has the advantage of simplicity in simulations.

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 while guaranteeing required power. If the stopping criteria are not met at any of the planned interim analyses, the trial is stopped when the pre-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 achieving sufficient power, and interim analyses are to be performed when the outcome is available in batches of size 3000, 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.

**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 probability statements. In other words, the test statistic is derived from the posterior distribution and its sampling distribution is not generally known. As a result, statistical power cannot be obtained analytically, and the assessment of operating characteristics for a variety of design parameters and assumptions requires simulation studies.

Power is defined as the probability of concluding efficacy, at any of the preplanned interim analyses or at the final analysis, for a given value of \( RR \) under the alternative hypothesis, for example, \( RR = 0.6 \). Analogously, the false-positive rate is the probability of concluding efficacy, at any of the preplanned interim analyses or at the final analysis, for the null value of \( RR \), that is, \( H_0 : RR = 1 \).

In addition to power and false-positive rate, other relevant operating characteristics in a Bayesian adaptive trial include: the probability of concluding futility under the null and alternative hypotheses, the probability of stopping early or stopping at a specific interim analysis, and the probability of reaching the maximum affordable sample size.

**Results**

**Simulation study**

A detailed description of the simulation study, including the underlying assumptions, design configurations and data-generating model, is provided in the Supplemental Material. In addition, the results of the simulation study, including all simulation scenarios and outputs, are provided in an R-Shiny application.

In the following, we provide select simulation results for a design that employs a set of stopping rules where superiority and futility decisions are made with respect to \( p_1 \). The interim analyses are performed after the collection of responses from every 2000 participants. Key conclusions that can guide specification of design components are discussed.

Figure 2 shows the estimated power for the three outcomes across different simulation scenarios arising from the combination of values for \( RR \), control risk and superiority thresholds. The futility bound is held fixed at 0.9, as it appeared not to impact the estimated power. As expected, power decreases as \( RR \) gets closer to 1. The rate of decline is higher under lower control risk, with little power under the stringent outcome \( s \). A larger superiority probability threshold is considered more conservative and results in lower power. However, the difference is negligible except for weak/negligible 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 0.6 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 conservative superiority stopping rule is employed.

The estimates of false-positive rate are presented in Figure 3. An immediate observation is that the probability of a false-positive result is higher for $p_1$, that is, the outcome with respect to which a superiority stopping decision is made. The inflation of false-positive rate 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-positive results are as likely for the other two outcomes but the trial is not stopped according to the superiority results for these outcomes. False-positive results for $s$ and $p_2$ are only taken into account at the final analysis. 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. We emphasize that precise estimation of false-positive rates through simulations is challenging. While the presented results are helpful in selecting a set of design parameters, once a design is determined, a final round of simulations with a larger number of iterations is recommended to obtain more precise estimates for type I error rates.

Figure 4 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 5 presents the estimated probability of concluding futility for relevant simulation scenarios. The results are filtered to include only small or zero-effect assumptions, that is, $RR = 0.8, 0.9, 1$, since the probability of concluding futility is negligible in other cases. Clearly, using a higher futility bound results in higher chance of concluding futility. The futility rule is sufficiently conservative that the chance of concluding
Figure 3. 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 control risk 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.

Figure 4. 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 control risk presented in Table 2 and the color of the bars refers to the two values for the futility bounds.
futility is only non-negligible when the trial is in fact futile, that is, $RR \geq RR_f$. The probability of stopping for futility is highest under the largest control event risks.

Considering the simulation results, we specify the design parameters as those that achieve acceptable operating characteristics across effect size assumptions. The superiority probability threshold is selected as $t_s = 0.99$, since it maintains the false-positive rate below 5%. The futility bound is chosen to be $RR_f = 0.8$, since it results in a smaller sample size when the true effects are negligible or zero. The probability of false futility results was estimated as negligible across all scenarios, as it never occurred throughout the 500 simulation iterations.

With the current specifications, Set 3 of the stopping rules—superiority and futility stopping with respect to $p_1$—is preferred. As discussed earlier, even if the stopping rules are defined according to the same case definition, performing secondary analyses for the other two outcomes and a joint analysis will provide additional insight regarding the effect of intervention on various components of the outcome definitions.

### 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 risks vary over these definitions. One of the challenges in selecting a final design is choosing between various components of the design and decision criteria according to their corresponding operating characteristics, each of which depends on at least a subset of design and decision parameters.

We showcase a simulation study that explores the 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 achieve acceptable operating characteristics while meeting feasibility requirements. Note that the decision of final design selection is informed by the design operating characteristics but does not take advantage of a formal decision theoretic framework. Ideally, a utility function should be defined based on the design operating characteristics with appropriate weights, and the optimal design may then be selected as that which optimizes this utility function. This is a topic of ongoing research by a subset of the authors.

The simulation study presented in this article assumes the risk of each event, defined by their specific case definitions, is reduced equally by the intervention. We argue that this assumption may be strong and

### Table 2. Control event risk parameters for the three binary (Bernoulli) outcomes resulting from the three case definitions.

| Scenario | $\pi_s$ | $\pi_{p_1}$ | $\pi_{p_2}$ |
|----------|---------|-------------|-------------|
| Worst    | 0.002   | 0.02        | 0.09        |
| Base     | 0.01    | 0.05        | 0.12        |
| Best     | 0.015   | 0.08        | 0.15        |

![Figure 5.](image-url) 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 control risk presented in Table 2.
unrealistic. For example, the intervention 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 $s$, can dilute the effect and result in loss of power. However, in the absence of granular information about the mechanism of effect, this assumption is required for the primary analysis and power calculations.

A secondary joint analysis should be performed to explore the validity of the assumption of equal risk reduction across event types. In Section 2.2 of the Supplemental Material, a joint analysis of the three outcomes under the equal risk reduction assumption is considered. Specifically, a key argument in favor of the marginal primary analysis is that due to the nested structure of the outcomes under the equal risk reduction assumption, a joint analysis is equivalent to the marginal analysis of the most permissive outcome $p_2$.

Another underlying assumption is that the control event risks remain constant over time. In reality, this assumption may not hold. Incorporating a time-varying control risk into the simulation study requires functional assumptions for the risk over time. Alternatively, the interim analyses may be scheduled with respect to the number of events rather than number of participants. As argued earlier, given the small event risks, this approach adds a significant level of uncertainty about the timing of interim analyses and the overall length of trial. Efficient incorporation of time-varying event rates in a Bayesian adaptive design is a direction for future work.

To conclude, Bayesian adaptive designs offer the potential for maximizing the information learned from the data collected through clinical trials. Modern clinical research is in increasing need of custom designs tailored to the specific characteristics of clinical questions rather than one-size-fits-all clinical trial designs with limited flexibility. Creative utilization of the flexibilities of Bayesian adaptive designs can lead to more efficient and effective studies in clinical research.

Code
The source code for the simulation study is available at the public repository: https://github.com/sgolchi/sepsis. The R-Shiny application mentioned in the article serves as an interactive plot repository for the simulation results.

Acknowledgements
The authors thank the SEPSiS project investigators for their collaboration.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Bill and Melinda Gates Foundation (grant number INV-007389). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Article version that might arise from this submission.

ORCID iD
Shirin Golchi https://orcid.org/0000-0003-3382-9563

Supplemental material
Supplemental material for this article is available online.

References
1. Cheng Y and Shen Y. Bayesian adaptive designs for clinical trials. Biometrika 2005; 92: 633–646.
2. Berry SM, Carlin BP, Lee JJ, et al. Bayesian adaptive methods for clinical trials. Boca Raton, FL: CRC Press, 2010.
3. Harrington D and Parmigiani G. I-SPY 2 — a glimpse of the future of phase 2 drug development. N Engl J Med 2016; 375(1): 7–9.
4. Warner P, Weir CJ, Hansen CH, et al. Low-dose dexamethasone as a treatment for women with heavy menstrual bleeding: protocol for response-adaptive randomised placebo-controlled dose-finding parallel group trial (DexFEM). BMJ Open 2015; 5(1): e006837.
5. Satlin A, Wang J, Logovsky V, et al. Design of a Bayesian adaptive phase 2 proof-of-concept trial for BAN2401, a putative disease-modifying monoclonal antibody for the treatment of Alzheimer’s disease. Alzheimer’s Dement 2016; 2(1): 1–12.
6. Food and Drug Administration (FDA). Multiple endpoints in clinical trials: guidance for industry, 2017, https://www.fda.gov/files/drugs/published/Multiple-Endpoints-in-Clinical-Trials-Guidance-for-Industry.pdf
7. McCoy CE. Understanding the use of composite endpoints in clinical trials. West J Emerg Med 2018; 19(4): 631–634.
8. Mehta C, Gao P, Bhatt DL, et al. Optimizing trial design: sequential, adaptive, and enrichment strategies. Circulation 2009; 119(4): 597–605.
9. Shane AL, Sanchez PJ and Stoll BJ. Neonatal sepsis. Lancet 2017; 390: 1770–1780.
10. Wynn JL. Defining neonatal sepsis. Curr Opin Pediatr 2016; 28(2): 135–140.
11. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA* 2016; 315: 801–810.
12. Vergnano S, Buttery J, Cailes B, et al. Neonatal infections: case definition and guidelines for data collection, analysis, and presentation of immunisation safety data. *Vaccine* 2016; 34: 6038–6046.
13. Vergnano S, Seale AC, Fitchett EJA, et al. Serious bacterial infections in neonates: improving reporting and case definitions. *Int Health* 2017; 9(3): 148–155.