Received: 14 December 2018 | Revised: 26 April 2019 | Accepted: 14 June 2019

DOI: 10.1002/sim.8316

RESEARCH ARTICLE

WILEY Statistics in Medicine

Bayesian consensus-based sample size criteria for binomial proportions

Lawrence Joseph[1] | **Patrick Bélisle**[2]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada

[2]Division of Clinical Epidemiology, McGill University Health Centre, Montreal, QC, Canada

Correspondence
Lawrence Joseph, Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, QC H3A 1A2, Canada.
Email: lawrence.joseph@mcgill.ca

Many sample size criteria exist. These include power calculations and methods based on confidence interval widths from a frequentist viewpoint, and Bayesian methods based on credible interval widths or decision theory. Bayesian methods account for the inherent uncertainty of inputs to sample size calculations through the use of prior information rather than the point estimates typically used by frequentist methods. However, the choice of prior density can be problematic because there will almost always be different appreciations of the past evidence. Such differences can be accommodated a priori by robust methods for Bayesian design, for example, using mixtures or ϵ-contaminated priors. This would then ensure that the prior class includes divergent opinions. However, one may prefer to report several posterior densities arising from a "community of priors," which cover the range of plausible prior densities, rather than forming a single class of priors. To date, however, there are no corresponding sample size methods that specifically account for a community of prior densities in the sense of ensuring a large-enough sample size for the data to sufficiently overwhelm the priors to ensure consensus across widely divergent prior views. In this paper, we develop methods that account for the variability in prior opinions by providing the sample size required to induce posterior agreement to a prespecified degree. Prototypic examples to one- and two-sample binomial outcomes are included. We compare sample sizes from criteria that consider a family of priors to those that would result from previous interval-based Bayesian criteria.

KEYWORDS

Bayesian methods, binomial proportions, clinical trials, credible intervals, prior specification, sample size determination, study design

1 | INTRODUCTION

A wide variety of sample size criteria have been proposed. These range from power calculations and methods based on confidence interval widths from a frequentist viewpoint (reviewed by Lemeshow et al[1] and Desu and Raghavarao[2]) to Bayesian versions of these same criteria,[3-5] reviewed by Adcock[6] and Wang and Gelfand.[7] Bristol[8] showed that sample sizes based on interval widths are not directly related to those based on power, so that sample sizes guaranteeing high power may not be sufficient for accurate estimation. It is therefore important that the sample size methods match the eventual analysis. Because reporting interval estimates is preferable to hypothesis testing in most practical circumstances,[9] sample sizes should be based upon interval widths rather than power of hypothesis tests. Decision theoretic criteria have also been proposed,[10] but while interesting in theory, these methods are difficult to implement in practice in a large part

because realistic loss functions are difficult to derive and are highly specific to a given application. These criteria will therefore not be discussed further here.

Frequentist sample size methods depend on point estimates of the required inputs such as means, proportions, and standard deviations, but these are typically not accurately known at the design stage of any study. It is therefore advantageous to consider Bayesian methods, where prior densities not only allow for uncertainty in the inputs but also incorporate this uncertainty into the sample size calculations. However, the choice of prior density can be problematic because there will almost always be different assessments of the previous evidence about any parameter, leading to different posterior conclusions from the data collected in the study. This is extremely important to address to avoid study results that are nondefinitive, in the sense that interested parties with different prior views may not agree on the final conclusions. It is desirable, therefore, at the planning stage to know the sample size that is sufficiently large to bring initially divergent views together. From an analysis perspective, one can examine robustness to prior inputs by reporting a family of posterior densities corresponding to a "community of priors," which run from optimistic to pessimistic.[11] In this paper, we propose a design methodology that corresponds to this type of analysis.

This work falls into the general category of robust Bayesian design, and there have been notable methods in the past that are in a similar direction to the consensus sample size methods presented here. De Santis[12] considered a class of power priors, with the goal of resolving possible differences between historical data and data collected in the current study, for example, by down-weighting the historical data. Brutti and de Santis[13] select a class of priors and ensure a large-enough sample size such that the lower bound of the posterior credible interval for a treatment difference will be above a given threshold, regardless of which prior in the class is used to analyze the data. They apply their methods to normal sampling situations, using priors with the same mean but with varying standard deviations. Interestingly, they also consider optimistic and pessimistic prior scenarios, but they are used in separate sample size calculations, providing a sample size relating to each possibility, but not necessarily resolving differences between them a posteriori. Similarly, Brutti et al[14] propose ϵ-contamination priors, ensuring a "successful trial" in terms of avoiding the range of clinical equivalence regardless of which prior in that class is used. Brutti et al[15] extended their own earlier work[14] by using mixtures of priors rather than ϵ contamination priors. Gajewski and Mayo[16] employ a mixture prior similar to Brutti et al,[15] but they focus on designing phase II clinical trials. De Santis[17] proposes to select the sample size for normal sampling such that the upper and/or lower posterior quantities of a given statistic will be close, regardless of which prior in a class of priors is used. While they also consider using optimistic and pessimistic priors, these are used only as design but not analysis priors (the distinction between design and analysis priors is discussed in Section 2). DasGupta and Mukhopadhyay[18] proposed calculating the sample size to guarantee posterior robustness using a minimax approach.

The general idea behind these works[12-18] has been that there is uncertainty about the prior, not primarily because of widely differing opinions about the past work but because the choice of the exact prior to use within a class of priors is not clear, and concern about robustness to this choice of prior. For example, many of the methods have chosen a single prior mean, and varied the strength of the prior around this mean. Other methods have combined mixtures of different priors to form a single prior, and still, others have used ϵ-contamination priors. In all of these cases, the primary concern is to ensure robustness to a prior thought to be possibly imperfect or uncertain, and not to address widely divergent priors from different interested parties. Although some authors did directly address families of distributions including optimistic and pessimistic priors, they did so by not finding the sample size required to resolve these differences in comparing across posterior densities from these individual priors, but rather by combining these priors by a mixture or calculating distinct sample sizes for each of the optimistic and pessimistic priors. While all of these methods can be used to help design a wide variety of studies in the presence of prior uncertainty, none uses the identical consensus-based criteria proposed here and, therefore, will not necessarily produce the same sample sizes.

In this paper, we develop methods that account for variability in prior opinions by providing the sample size required to induce posterior agreement to a prespecified degree across a range of prior opinions. Specifically, we consider the most *optimistic* and *pessimistic* choices of priors and determine the sample size to induce posterior agreement between these extremes. We develop criteria using three different ways to handle sampling uncertainty with applications to both single- and two-population binomial sampling designs, as commonly found in clinical trials. The results will provide the sample size required to design a definitive study or trial, in the sense that the amount of information gathered will be sufficient to ensure posterior agreement regardless of prior opinion.

The outline of this paper is as follows. Section 2 reviews various Bayesian sample size criteria based on the highest posterior density (HPD) credible interval lengths and applies these criteria to deriving sample sizes for posterior agreement given two divergent prior densities. HPD intervals are optimal, in that they will lead to the smallest possible sample sizes for a given desired length and coverage probability. Specific methods for calculating the sample sizes defined by the

criteria in Section 2 in the case of single- and two-group binomial sampling are given in Section 3. Sample sizes from a series of prototypic examples are given in Section 4, comparing the change in sample size with and without consideration of posterior agreement. We end with a discussion in Section 5.

2 | BAYESIAN SAMPLE SIZE CRITERIA THAT ACCOUNT FOR DISPARITIES IN PRIOR OPINIONS

Let θ denote an unknown parameter to be estimated, let Θ denote the parameter space for θ, and let $f(\theta)$ summarize the prior information about θ. Suppose further that data $x \in \mathcal{X}$ with sample size N will be collected to inform about θ. The preposterior predictive distribution for x, the marginal density for the data, is given by

$$f(x) = \int_{\Theta} f(x|\theta) f(\theta) d\theta, \tag{1}$$

and the posterior distribution of θ given x is $f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}$, where $f(x|\theta)$ is the likelihood function of the data x.

If $f(\theta)$ is known or generally agreed upon by all parties of interest, then sample sizes can be based on ensuring a large-enough sample for accurate estimation, as measured by the width of an HPD interval. Because the data are not known at the design stage of the study, one can guarantee the desired accuracy on average, or as a percentage of all possible data sets, weighted by $f(x)$. Such methods have been developed for binomial parameters by Joseph et al[5] and extended to differences between binomial parameters by Joseph et al.[5] M'Lan et al[19] extended these methods to include curve fitting algorithms for efficient searching for the optimal sample size. These methods are implemented in an R package available from CRAN (cran.r-project.org/web/packages/SampleSizeProportions/index.html).

In practice, however, it is rare that there is consensus about the choice of $f(\theta)$, as different appreciations of the available information about θ will lead to different prior densities. Let $f_1(\theta)$ and $f_2(\theta)$ represent two prior distributions over θ. For example, θ might represent the difference in effectiveness between a standard and a novel therapy, with $f_1(\theta)$ and $f_2(\theta)$ representing two expert opinions, one being optimistic and the other being more pessimistic about the value of the new treatment. It may then be of interest to not only estimate θ to a given accuracy using either $f_1(\theta)$ or $f_2(\theta)$ but to ensure that the two distinct posterior densities arising from $f_1(\theta)$ and $f_2(\theta)$ agree to a prespecified degree. In doing so, one would ensure a definitive trial, in the sense that the data from the trial will be sufficiently informative to resolve a priori differences in opinion. We note that it would not be sufficient to simply employ a noninformative prior because not only does that ignore any existing prior opinions, it does not guarantee that the sample size will be large enough to ensure posterior agreement when different interested parties may draw conclusions from the study draw using their own prior densities.

One can define this posterior agreement in several ways. Given that final results are most often reported as posterior intervals, it is perhaps most natural to keep to HPD intervals, and ensure that the maximum distance between the lower and upper HPD interval limits from the two posterior densities associated with $f_1(\theta)$ and $f_2(\theta)$ are within a prespecified distance ϵ. Letting the HPD intervals derived from $f_1(\theta)$ and $f_2(\theta)$ given data x be represented by $(L_1(x), U_1(x))$ and $(L_2(x), U_2(x))$, respectively, we would seek a sample size that ensures

$$m(x) = \max(|L_1(x) - L_2(x)|, |U_1(x) - U_2(x)|) < \epsilon. \tag{2}$$

Alternative definitions of posterior agreement are of course possible. While our R software package includes criteria based on closeness of sets of posterior cumulative probabilities and sets of posterior quantiles, the methods leading to sample sizes from these criteria are very similar to those for the maximum distance between the lower and upper HPD interval limits. Indeed, all steps are identical except for the check of whether the criterion is satisfied at the current sample size, which of course depends on the particular criterion. Hence, without much loss of generality, these alternate possibilities are not further discussed here.

For any given sample size, some data sets x may satisfy (2) while others may not. Hence, one requires a method to handle the inherent data uncertainty in selecting the final sample size at the design stage. For example, one may be satisfied with ensuring that (2) holds on average over all possible x, the average being over the predictive distribution of the data given by (1). Alternatively, one can apply a more stringent criterion, such as ensuring that the sample size is sufficiently large such that (2) will hold across all possible data sets, or over a sufficiently large proportion of all possible data sets.

We term the most restrictive criterion, when (2) must hold over all possible data sets the *worst outcome criterion* (WOC). We use the term "modified WOC" (MWOC) when there is a prespecified proportion of all data sets over which (2) must hold. Thus, for example, MWOC(90) and MWOC(50) would indicate the criteria for sample sizes that fulfill (2) over 90% or 50%, respectively, of all data sets, according to the probabilities given by (1). Finally, the average coverage criterion is used when one wishes for (2) to be satisfied on average over all data sets, that is, ϵ is the average distance achieved over all data sets, again weighted by the probabilities given by (1).

One must also select a *design prior,* that is, the prior density to plug into (1) to generate the set of all possible data x, which is not necessarily the same as the "analysis priors" $f_1(x)$ or $f_2(x)$. For example, one can set the design prior to be a linear combination of the two analysis priors leading to $f(x) = af_1(x) + (1-a)f_2(x)$, where $0 \leq a \leq 1$, or one can choose an entirely different prior, say, $f_3(x)$, to use as the design prior. The density $f_3(x)$ could represent, for example, an opinion close to that of an average clinician, or the "clinical prior" as defined by Spiegelhalter et al.[11] One can also consider ensuring that the criterion (2) is satisfied both when $f_1(x)$ or $f_2(x)$ are plugged into (1) as the design prior.

Taking all possible combinations of the three choices of design prior density to use in (1) and degree of certainty of reaching the desired agreement accuracy (again three choices, on average, over all possible data sets or over a given proportion of the data sets) leads to seven possible sample size criteria. Note that we do not have nine criteria as might be expected (three prior choices times three degrees of certainty based on the data) because if we wish to ensure that (2) is satisfied over all data sets, the choice of prior in (1) does not matter.

In the next section, we will present methods for applying the above criteria in selecting sample sizes for studies involving binomial parameters and the difference between two-binomial parameters.

3 | SAMPLE SIZE METHODS FOR BINOMIAL PARAMETERS AND THE DIFFERENCE BETWEEN TWO-BINOMIAL PARAMETERS

In this section, we apply the above criteria to calculate sample size requirements for studies involving binomial parameters. We begin by specifying the likelihood function and prior distributions for binomial parameters, first for simple experiments that estimate a single-binomial parameter and, then, for experiments estimating the difference between two-binomial parameters. Once these are defined, all criteria of Section 2 are fully specified and, in theory, can be applied. Closed-form posterior densities are available for single-binomial parameters, and while in theory they are also available for the difference between two-binomial parameters,[20] the resulting calculations are cumbersome and it is thus preferable to use simulations from the posterior density or substitute an approximate density. Therefore, below, we will outline both exact and approximate algorithms that can be used to calculate the sample sizes in practice.

3.1 | Methods for single-binomial parameters

Let θ be the binomial probability to be estimated. We assume that two different analysis priors $f_1(\theta)$ and $f_2(\theta)$ can be expressed as beta densities, with

$$f_i(\theta) \sim \text{Beta}(\alpha_i, \beta_i), \quad i = 1, 2. \tag{3}$$

After observing data x of sample size n, the corresponding posterior densities will then also be from the beta family, with

$$f_i(\theta) \sim \text{Beta}(\alpha_i + x, \beta_i + n - x), \quad i = 1, 2. \tag{4}$$

We calculate HPD intervals from the beta posterior density using the algorithm given by M'Lan et al.[19]

In the case of estimating a binomial parameter with a beta prior density, the predictive density (1) is beta-binomial. We will denote this probability function by $\text{Bb}(\alpha, \beta, n)$, given by

$$\begin{aligned}
\text{Bb}(x; \alpha, \beta, n) &= \int \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\text{B}(\alpha, \beta)} d\theta \\
&= \binom{n}{x} \frac{1}{\text{B}(\alpha, \beta)} \int \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} d\theta \\
&= \binom{n}{x} \frac{\text{B}(x+\alpha, n-x+\beta)}{\text{B}(\alpha, \beta)} \\
&= \frac{n!}{x!(n-x)!} \cdot \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)\Gamma(\alpha+\beta)}{\Gamma(n+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)},
\end{aligned}$$

where $\Gamma(x)$ is the gamma function. If the design prior for θ is a linear combination of beta densities, then the predictive density is a linear combination of beta-binomial distributions.

To determine the sample size, let $m(x)$ be the criterion of interest as defined by (2), obtained when x successes are obtained from a sample size of n. Given the sample size n and prior values α and β, $\text{Bb}(\alpha, \beta, n)$ depends only on the observed data x, and so can be denoted as $f(x)$. Using this notation, the minimum sample size n satisfying

$$\sum_{x=0}^{n} f(x)m(x) \leq \epsilon \tag{5}$$

is the optimal sample size n. Given a bounded range that contains the correct sample size, it can be found through a bisectional search algorithm.[21] In practice, a starting sample size is selected, and the above criterion is checked. The sample size is then increased or decreased according to whether $\sum_{x=0}^{n} f(x)m(x)$ is below or above ϵ, continuing until the optimal sample size is found.

3.2 | Methods for the difference between two-binomial parameters

We next consider sample size determination for experiments aimed at drawing inferences about the difference between two-binomial proportions, when two researchers have different prior beliefs and consensus about their posterior inferences is desired. Let $i = 1, 2$ index the two populations, and let $k = 1, 2$ index the two researchers, each of whom models their prior knowledge about the binomial proportions θ_i, $i = 1, 2$, through the beta densities

$$\theta_i \sim \text{Beta}(\alpha_{ik}, \beta_{ik}),\ i = 1, 2,\ k = 1, 2.$$

We allow the sample sizes drawn from each group to differ. Let n_1 denote the sample size drawn to estimate θ_1, and let $n_2 = M \times n_1$, where the sample size ratio M is given by the study designer. If $M = 1$, then the two groups will have equal sample sizes.

As is typical in clinical trial design, we assume the two groups are independent. Therefore, the posterior density of each binomial proportion is given by

$$\theta_i | x_i \sim \text{Beta}(\alpha_{ik} + x_i, \beta_{ik} + n_i - x_i),$$

and the posterior density of the parameter of main interest is defined by the difference between these two parameters $\theta_2 - \theta_1$, the difference between two independent beta distributions. We approximate the exact posterior density with a generalized beta distribution, as previously discussed in the work of Joseph et al.[4] In particular, we approximate the exact density of $\theta_2 - \theta_1$ with a beta distributed variable $Z \sim \text{Beta}(\alpha^*, \beta^*)$, such that $Z^* = 2Z - 1$ is nonzero over the range $(-1, 1)$, and fit by the method of moments with $E(Z^*)$ and $V(Z^*)$ matching the posterior mean and variance of $\theta_2 - \theta_1$.

Having defined this approximation to the exact posterior density of $\theta_2 - \theta_1$, for any given sample size as defined by n_1 and the multiplier M, we can now sample from the predictive density of the as yet unobserved data to approximate the criterion as given by (2), and as in the single-binomial case, use a search algorithm to converge to the optimal sample size.

We next use our methods to determine sample sizes for various scenarios that may occur in the planning of studies with dichotomous outcomes. User-friendly R packages called "SampleSizeConsensusBinomialProportion" and "SampleSizeConsensusBinomialProportionsDiff" that implements all of the above methods (as well as some additional criteria not discussed in detail here) are available from www.medicine.mcgill.ca/epidemiology/Joseph/.

4 | SAMPLE SIZES FOR PROTOTYPIC SCENARIOS

We first look at a typical example when estimating a single-binomial proportion, followed by an application of our methods to the difference between two-binomial proportions. In each case, we calculate the required sample size to ensure a high degree of posterior agreement starting from prior densities from researchers whose opinions initially diverge. We also calculate the sample sizes required from non–consensus-based criteria because one may in practice wish to assess sample sizes from a variety of criteria before settling on a final sample size. For example, one might consider first running interval-based sample size methods, which ensure a sufficient sample size for estimation purposes when there is a high degree of prior agreement among researchers. One can then check how different the sample size might be if one, in addition, requires a high degree of posterior agreement among researchers whose opinions initially vary. Based on all of

the calculated sample sizes, one can decide on a reasonable final choice depending on the information one would obtain for each sample size and how much importance one places on posterior agreement.

4.1 | Example for a single-binomial proportion

Suppose one is designing a study of a new surgical technique, with the objective of estimating the probability of success for this technique through a single sample of size n. Suppose that one researcher (or interested party) is enthusiastic, and believes that the true proportion is highly likely to be near 90% and is a priori 95% certain that the true value is in the interval (0.85, 0.95). However, a second researcher (or interested party) is more pessimistic, and a priori believes the true rate to be closer to 80%, with 95% certainty that the true value is in the interval (0.75, 0.85). Following the criterion given by (2), suppose we define posterior agreement to be that the maximum difference between the upper and lower 95% posterior HPD intervals from the two researchers should be not larger than 0.005, ie, they will agree on both ends of the interval to within one-half of a percentage point. What should the sample size n be to ensure this degree of posterior agreement arising from these two distinct prior densities? Furthermore, apart from guaranteeing posterior agreement, is this sample size sufficient for accurate estimation of the probability of success, here specified as a 95% HPD interval with total width no larger than 0.04 (ie, roughly $\pm$ 2%)?

We first provide the relevant inputs to the sample size calculation. We assume analysis priors for the first and second researchers of beta(116.064, 12.045) and beta(194.0375, 47.79375), respectively, to match their prior intervals given above. Our software allows inputs of either a prior 95% HPD interval, or a prior mean and standard deviation, in each case calculating the parameters of a beta density that best fits the information given. Users can of course also directly input beta prior parameters. We next assume a design prior that covers the full range of the researchers prior beliefs, that is, a range of (0.75, 0.95), which implies a beta(36.596, 5.6483) design density. Referring to Equation (2), we assume $\epsilon = 0.005$.

A sample size of $n = 3,979$ is required to ensure a maximum distance between upper and lower HPD limits from the two researchers posterior densities of 0.005, or one-half of one percentage point, on average over all data sets. A very similar sample size of 4047 is required to attain the same posterior agreement over half of all data sets, whereas the size increases to 5423 to cover 90% of all data sets. These MWOC sample sizes are found by sampling 50% or 90% of all possible data sets according to the predictive density (1). To cover 100% of all data sets, one requires a much larger sample size of 16 386, in a large part because the extra 10% of data sets are likely to contain improbable but highly divergent data sets that require large sample sizes to reach posterior agreement.

In practice, therefore, ensuring a high degree of posterior agreement between initially divergent opinions requires a sample size roughly between 4000 and 5500, depending on the exact degree of certainty over which one wishes to ensure agreement. It is interesting to also calculate sample sizes that ensure accurate estimation width, regardless of posterior consensus, here defined as a total posterior HPD interval width of 0.04 using the same design prior as for the consensus-based calculations. Using the Bayesian criteria given by Joseph et al,[5] one requires sample sizes of 1070, 1639, or 2358 to attain the desired accuracy on average, over 90% and over 100% of all data sets, respectively. Therefore, one requires a much larger sample size to ensure very close posterior agreement between divergent researchers than to attain the desired estimation accuracy. In this case, one might choose to relax the degree of agreement. If 0.01 is used instead of 0.005, that is, 95% posterior HPD limits will agree to within 1%, then the sample size to achieve this on average is 1897, and to agree over 90% of all data sets requires a sample size of 2613. Similarly, requiring a 2% agreement difference results in sample sizes of 850 and 1194 on average and over 90% of all data sets, respectively.

Roughly speaking, one can see that a sample size close to 1100 will ensure accurate estimation in terms of a total HPD width of 0.04, while also ensuring posterior agreement from our a priori divergent researchers of 2%. Increasing this sample size to close to 2000 will of course ensure even better than the desired estimation accuracy while also attaining a posterior agreement for our two researchers of 1%. One can then select a final sample size based on all of this information, depending on practical considerations including budget and ability to recruit subjects, as well as the importance of getting the community of researchers to all agree on a final posterior estimate.

4.2 | Example for the difference between two-binomial proportions

We will now turn to the design of a prototypical comparative clinical trial. Suppose a new medication has been developed and is to be compared with the standard treatment. We again consider two researchers with different appreciations of the past literature on these two medications. With two researchers each providing their prior evaluations of both the standard and new treatments, we have four prior densities to input as analysis priors. Suppose the first researcher is more sure of their opinions, and strongly believes the new medication to be superior to the standard treatment, giving 95%

prior HPD ranges of (0.75, 0.85) for the standard therapy and shifting both upper and lower limits to (0.85, 0.95) for the new medication. The second researcher is rather pessimistic that the new medication is better than the standard but is also less certain of their opinion, so their two intervals overlap and are wider compared with the intervals from the first researcher. In particular, they provide 95% prior HPD intervals of (0.70, 0.90) for the standard treatment, and (0.70, 0.95) for the new, allowing for an upper tail for the new medication that exceeds that upper limit of the standard therapy. We will define posterior agreement by having both upper and lower 95% HPD limits agree to within 0.01, that is, to within 1%. We will also calculate the sample size required for non–consensus sample size methods, calculating the sample size requirements to estimate the difference between these two-binomial success probabilities to within a total HPD width of 0.04, or roughly ±2%.

We first convert the four analysis prior ranges to beta densities. This gives densities of beta(194.0375, 47.79375), beta(116.064, 12.045), beta(46.3288, 10.84949), and beta(25.22343, 4.56154) for the four rages of (0.75, 0.85), (0.85, 0.95), (0.7, 0.9), and (0.7, 0.95), respectively. Similar to our example with a single-binomial proportion, we will assume design priors that cover the entire range of both researchers. In this case, because the second researcher's intervals include those from the first researcher, the beta densities required here in fact match those from the analysis priors from the second researcher. We plug $\epsilon = 0.01$ into Equation (2).

A sample size of $n = 612$ per treatment group, giving a total sample size of 1,224, is required to ensure a maximum distance between upper and lower HPD limits from the two researchers posterior densities of 0.01, or one percentage point, on average over all data sets, weighted by the design prior given above. A slightly smaller sample size of 559 per group (total sample size of 1118) is required to attain the same posterior agreement over half of all data sets, whereas the size substantially increases to 1029 per group (total of 2058) to cover 90% of all data sets. To cover 100% of all data sets, one requires a much larger sample size of 6946 per group (total of 13892), again, because of highly improbably data sets that require large sample sizes to reach posterior agreement.

Total sample sizes close to 1100 or 1200 therefore will provide about 50% assurance of close posterior agreement between the two researchers, while doubling this to about to 2000 will ensure close agreement over 90% of all data sets, according to the design priors. We can compare these sample size to those required to estimate this difference to within a total credible interval HPD width of 0.04 ignoring the divergent priors from the two researchers and using the same design prior as for the consensus-based calculations. Using the Bayesian criteria given by Joseph et al,[4] one requires per group (total) sample sizes of 2639 (5078), 3650 (7300), and 4758 (9516) to attain the desired accuracy on average, over 90% and over 100% of all data sets, respectively. Unlike the example of Section 4.1.2, here, we find the consensus sample sizes are smaller than those required by the interval-based methods. In this case, one can investigate more strict degree of agreements. If 0.005 is used instead of 0.01, that is, 95% posterior HPD limits will agree to within one-half of a percentage point, the sample size required to achieve this on average is 1207 per group (2414) and to agree over 90% of all data sets requires a sample size of 2077 (4154). Thus, we are still ensured of our desired posterior agreement with sample sizes less than those given by interval-based methods.

Note that the main difference between this and the example of Section 4.1 is the stronger a priori agreement between researchers. In the first example, the researchers' prior intervals did not overlap, whereas in the example of this section, while the prior opinions were again quite different, there was a large degree of overlap, leading to smaller sample sizes required for agreement.

5 | DISCUSSION

Many sample size criteria have been proposed to aid the design of clinical research studies and clinical trials from both hypothesis testing and interval estimation viewpoints. Since the advent of fast desktop computers and the development of MCMC algorithms for Bayesian analysis, increasing numbers of statisticians are using Bayesian methods. Accompanying these advances have been less technical but no less important changes in how Bayesian statistics are practiced. Technical problems in computing posterior densities for complex models aside, a major roadblock to the use of Bayesian methods has been the choice of prior density. In particular, the nonuniqueness of prior densities as each researcher or other individual interested in the study question assesses the available evidence in different ways, has often been cited as a reason to avoid Bayesian methods. This problem was addressed for Bayesian analyses by reporting a family of posterior densities corresponding to a family of priors.[11] By ensuring that the family of posterior densities arises from a family of prior densities that span the range of reasonable prior opinions, all readers should be able to draw their own conclusions incorporating their view of past evidence together with the information in the new data set. If the range of posterior densities is

narrow enough, then the new data set should result in consensus even among researchers with previously divergent opinions. However, if the posterior densities remain sufficiently different, one must admit that more data must be collected to definitively resolve the issue. The next natural question is how much data needs to be collected to result in a definitive trial. It is this question that we and others[12-18] have aimed to resolve. In this paper, we provide sample size determination methods for ensuring closeness of the posterior densities starting from potentially highly divergent prior opinions.

Of course, having posterior agreement is not the only important outcome of a study; one also needs to ensure sufficient information for decision making. Therefore, as in both of the prototypic examples we presented in Section 4, we calculated sample sizes from our consensus-based approach and from an interval-based approach. As our examples illustrate, sometimes one sample size can be larger than the other, and vice versa. Taking the maximum over the two sample sizes guarantees both sufficiently small interval estimates and consensus among researchers.

While this paper has concentrated on binomial sampling, of course, our methods can be extended to other sampling situations. Software for both binomial and normal sampling situations, the two most common sampling designs used, are available from the first author's webpage at http://www.medicine.mcgill.ca/epidemiology/Joseph/.

DATA AVAILABILITY STATEMENT

There are no data used, so none need to be made available for this article.

ORCID

Lawrence Joseph https://orcid.org/0000-0002-5819-7999

REFERENCES

1. Lemeshow S, Hosmer D, Klar J, Lwanga S. *Adequacy of Sample Size in Health Studies*. Chichester, UK: John Wiley & Sons Ltd; 1990.
2. Desu M, Raghavarao D. *Sample Size Methodology*. New York, NY: Academic Press; 1990.
3. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial based on subjective clinical opinion. *Statist Med*. 1986;5:113.
4. Joseph L, du Berger R, Bélisle P. Bayesian and mixed Bayesian/likelihood criteria for sample size determination. *Statist Med*. 1997;16(7):769-781.
5. Joseph L, Wolfson D, du Berger R. Sample size calculations for binomial proportions via highest posterior density intervals. *J R Stat Soc Ser D*. 1995;44(2):143-154.
6. Adcock CJ. Sample size determination: a review. *J R Stat Soc Ser D*. 1997;46(2):261-283.
7. Wang F, Gelfand A. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*. 2002;17(2):193-208.
8. Bristol DR. Sample sizes for constructing confidence intervals and testing hypotheses. *Statist Med*. 1989;8:803-811.
9. Gardner MJ, Altman DG. Estimating with confidence. *Br Med J*. 1988;296(6631):1210-1211.
10. Pallay A. A decision analytic approach to determining sample sizes in a Phase III program. *Drug Inf J*. 2000;34:365-377.
11. Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials (with discussion). *J R Stat Soc Ser A*. 1994;157(3):357-416.
12. De Santis F. Using historical data for Bayesian sample size determination. *J R Stat Soc Ser A*. 2007;170(1):95-113.
13. Brutti P, de Santis F. Robust Bayesian sample size determination for avoiding the range of equivalence in clinical trials. *J Stat Planning Infer*. 2008;138:1577-1591.
14. Brutti P, de Santis F, Gubbiotti S. Robust Bayesian sample size determination in clinical trials. *Statist Med*. 2008;27:2290-2306.
15. Brutti P, de Santis F, Gubbiotti S. Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials. *Statist Med*. 2009;28:2185-2201.
16. Gajewski BJ, Mayo MS. Bayesian sample size calculations in phase II clinical trials using a mixture of informative priors. *Statist Med*. 2006;25:2554-2566.
17. De Santis F. Sample size determination for robust Bayesian analysis. *J Am Stat Soc*. 2006;101(473):278-291.
18. DasGupta A, Mukhopadhyay S. Uniform and subuniform posterior robustness: the sample size problem. *J Stat Planning Inf*. 1994;40:189-204.
19. M'Lan E, Joseph L, Wolfso D. Bayesian sample size determination for binomial proportions. *Bayesian Analysis*. 2008;3(2):269-296.
20. Pham-Gia T, Turkkan N. Bayesian analysis of the difference of two proportions. *Commun Statistics-Theory Methods*. 1993;22:1755-1771.
21. Thisted R. *Elements of Statistical Computing*. New York, NY: CRC Press; 1988.

How to cite this article: Joseph L, Bélisle P. Bayesian consensus-based sample size criteria for binomial proportions. *Statistics in Medicine*. 2019;1–8. https://doi.org/10.1002/sim.8316