Received: 8 March 2021 | Accepted: 23 February 2022

DOI: 10.1111/biom.13649

BIOMETRIC METHODOLOGY

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Bayesian sample size determination using commensurate priors to leverage preexperimental data

Haiyan Zheng[1,2] | **Thomas Jaki**[1,3] | **James M.S. Wason**[2]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[2]Population Health Sciences Institute, Newcastle University, Newcastle upon Tyne, UK

[3]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

Correspondence
Haiyan Zheng, MRC Biostatistics Unit, University of Cambridge, Cambridge, CB2 0SR, UK.
Email: haiyan.zheng@mrc-bsu.cam.ac.uk

Funding information
Medical Research Council, Grant/Award Numbers: MC_UU_00002/6, MC_UU_00002/14; Cancer Research UK, Grant/Award Number: RCCPDF\100008; National Institute for Health Research, Grant/Award Number: NIHR-SRF-2015-08-001

Abstract

This paper develops Bayesian sample size formulae for experiments comparing two groups, where relevant preexperimental information from multiple sources can be incorporated in a robust prior to support both the design and analysis. We use commensurate predictive priors for borrowing of information and further place Gamma mixture priors on the precisions to account for preliminary belief about the pairwise (in)commensurability between parameters that underpin the historical and new experiments. Averaged over the probability space of the new experimental data, appropriate sample sizes are found according to criteria that control certain aspects of the posterior distribution, such as the coverage probability or length of a defined density region. Our Bayesian methodology can be applied to circumstances that compare two normal means, proportions, or event times. When nuisance parameters (such as variance) in the new experiment are unknown, a prior distribution can further be specified based on preexperimental data. Exact solutions are available based on most of the criteria considered for Bayesian sample size determination, while a search procedure is described in cases for which there are no closed-form expressions. We illustrate the application of our sample size formulae in the design of clinical trials, where pretrial information is available to be leveraged. Hypothetical data examples, motivated by a rare-disease trial with an elicited expert prior opinion, and a comprehensive performance evaluation of the proposed methodology are presented.

KEYWORDS

Bayesian experimental designs, historical data, rare-disease trials, robustness, sample size

1 | INTRODUCTION

Conventionally, sample size has often been determined to control certain aspects of the sampling distribution of a test statistic (Desu and Raghavarao, 1990). This is typically considered from a frequentist perspective that operating characteristics, for example, type I error rate and power, should be maintained for detecting a meaningful magnitude of the difference. For data that are assumed to be independently and identically distributed normal, sample size may be a function also of nuisance parameters such as unknown variances. Fixing such parameters to certain values may leave the determination inaccurate, or only locally optimal, as an arbitrary guess could deviate far from the true value. The Bayesian framework has been argued to be more advantageous to sample size determination (SSD), since it allows uncertainty to be described in a prior for the parameters (O'Hagan and

Forster, 2004). Moreover, it brings about the possibility of incorporating preexperimental information, if available, in a prior for the parameter of interest and/or nuisance parameters. Considerable attention has thus been given to Bayesian SSD; see, for example, Clarke and Yuan (2006).

Two main kinds of methodology written in the literature are "hybrid classical and Bayesian" and "proper Bayesian" SSD (Spiegelhalter *et al.*, 2004). With the former, sample size would be chosen to ensure that the predictive power, obtained by averaging the frequentist power function over a prior distribution for the unknown parameter(s), reaches a desired target level. By contrast, "proper Bayesian" SSD approaches refer to those for which the final analysis of data would also be Bayesian. Joseph *et al.* (1995) derive formulae for binomial experiments comparing two proportions; specifically, sample sizes are sought to ensure, for example, an adequate coverage probability or width of a defined interval of the posterior "success" rate. Joseph and Bélisle (1997) concentrate on normal distributions and use normal-gamma conjugate priors for experiments that estimate either single normal means or the difference between two normal means. In the context of clinical trials, Whitehead *et al.* (2008) develop Bayesian methods resembling frequentist formulations of the SSD problem in exploratory trials, to demonstrate the treatment effect based on posterior interval probabilities. These fully Bayesian approaches shed light on the option of incorporating preexperimental data into a prior for both the design and analysis consistently.

This paper is focused on fully Bayesian SSD permitting the use of preexperimental data from multiple sources. Our research is partly motivated by the efficient design and analysis of clinical trials that evaluate a new treatment for rare diseases (EMA, 2006), where asking for a sample size to achieve the frequentist power is often infeasible. Pretrial information, collected from historical studies which had been conducted under similar circumstances, or elicited from expert opinion, could play an essential role. The proposed methodology would nonetheless be generic: it can be applied to areas where there is a need to use preexperimental data formally through the mechanism of specifying priors. For instance, the sample size for environmental water quality evaluation could be limited: borrowing strength from historical water monitoring data has been considered helpful (Duan *et al.*, 2006).

2 | METHODS

2.1 | Borrowing of historical information from multiple sources

Suppose there are K relevant sets of data, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K$, to specify a prior for the parameter, denoted by μ_Δ, underpinning a new experiment. Let $\theta_1, \ldots, \theta_K$ denote the counterparts of μ_Δ, specific to each historical experiment $k = 1, \ldots, K$. Following Zheng and Wason (2022), we specify K commensurate predictive distributions by the source of information, which are formulated as conditional normal distributions with an unknown mean θ_k and precision ν_k (the variance would thus be ν_k^{-1}):

$$\tilde{\theta}_k \mid \theta_k, \nu_k \sim N(\theta_k, \nu_k^{-1}), \qquad \text{for } k = 1, \ldots, K, \tag{1}$$

where each $\tilde{\theta}_k$ is regarded as *equivalent* to μ_Δ in terms of the parameter space. More precisely, it means that the parameter space for $\tilde{\theta}_k$, as projected from a preexperimental parameter θ_k, would be defined with the same or comparable set of parameter values to that of μ_Δ. The precision ν_k is sometimes referred to as a commensurate parameter. Different from the original proposal with a spike-and-slab prior on each ν_k, we consider a mixture of conjugate priors for analytical derivations; that is, for the predictive precision:

$$\nu_k \sim w_k \text{Gamma}(a_{01}, b_{01}) + (1 - w_k)\text{Gamma}(a_{02}, b_{02}), \tag{2}$$

where w_k is the prior mixture weight, on the scale of [0, 1], to represent preliminary skepticism about how commensurate θ_k and μ_Δ would be. The hyperparameters are chosen so that the first mixture component with a_{01}, b_{01} has the density concentrated on small values of ν_k, while the second mixture component with a_{02}, b_{02} has density covering larger values of ν_k. A large prior mixture weight allocated to either component distribution would thus result in sufficient down-weighting (with no borrowing at all as one extreme) or strong borrowing of historical information (with fully pooling as the other extreme), respectively. Stipulating $0 < w_k < 1$ in (2) produces a compromise between the two extreme cases. The strength of this Gamma mixture prior is then tuned by w_k, which can be interpreted as the prior probability of incommensurability.

Then, $f(\tilde{\theta}_k, \nu_k \mid \theta_k)$ has a Normal-Gamma mixture distribution. By integrating out the nuisance parameter ν_k, we further obtain

$$f(\tilde{\theta}_k \mid \theta_k) \propto w_k \left(\frac{(\tilde{\theta}_k - \theta_k)^2}{2b_{01}} + 1 \right)^{-\frac{2a_{01}+1}{2}} + (1 - w_k) \left(\frac{(\tilde{\theta}_k - \theta_k)^2}{2b_{02}} + 1 \right)^{-\frac{2a_{02}+1}{2}}, \tag{3}$$

which is a two-component mixture of nonstandardized (shifted and scaled) t distributions. In particular, the

component t distributions have their location parameters identically as θ_k yet scale parameters as $\frac{b_{01}}{a_{01}}$ and $\frac{b_{02}}{a_{02}}$, respectively. Detailed derivation of (3) and the demonstration of it being a nonstandardized t mixture distribution are given in Section A of the Supporting Information. For easing the synthesis of K predictive priors later on, we approximate this unimodal t mixture distribution by a normal distribution that

$$\tilde{\theta}_k \mid \theta_k \dot{\sim} N\left(\theta_k, \frac{w_k b_{01}}{a_{01}-1} + \frac{(1-w_k)b_{02}}{a_{02}-1}\right), \quad \text{with } a_{01}, a_{02} > 1. \tag{4}$$

This approximation is based on the first two moments of the nonstandardized t mixture distribution, which are analytically available; see Section B of the Supporting Information for details. The variance of the normal approximation takes account of the dispersion of both t mixture components. The goodness of such normal approximation to the original t mixture distribution depends on the degrees of freedom, $2a_{01}$ and $2a_{02}$, and the scale parameters, $\frac{b_{01}}{a_{01}}$ and $\frac{b_{02}}{a_{02}}$, which are of the investigators' choice. We show the numerical accuracy of this approximation in Section C of the Supporting Information.

With the normal approximation given by (4), we stipulate μ_Δ as a linear combination of $K \geq 2$ hypothetical random variables, $\tilde{\theta}_k$, projected from the preexperimental parameters. That is, $\mu_\Delta = \sum_k p_k \tilde{\theta}_k$, for $k = 1, \ldots, K$. These synthesis weights $p_1, \ldots, p_K$ sum to 1, with each reflecting the relative importance of a corresponding preexperimental dataset to constitute the collective predictive prior for μ_Δ. Pragmatically, one may associate these synthesis weights with the prior probabilities of commensurability, that is, $1 - w_k$, so that a preexperimental dataset thought of as more commensurate would be assigned a larger p_k to derive the collective prior for μ_Δ. Applying the convolution operator for the sum of normal random variables, μ_Δ has a normal prior distribution. Suppose each preexperimental dataset leads to an estimate of $\theta_k \mid \boldsymbol{y}_k \sim N(m_k, s_k^2)$, $k = 1, \ldots, K$. We thus obtain a normal collective prior that

$$\mu_\Delta \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K \dot{\sim} N\left(\sum_k p_k \lambda_k, \sum_k p_k^2 \xi_k^2\right), \tag{5}$$

with

$$\lambda_k = m_k \quad \text{and} \quad \xi_k^2 = s_k^2 + \frac{w_k b_{01}}{a_{01}-1} + \frac{(1-w_k)b_{02}}{a_{02}-1}, \quad (a_{01}, a_{02} > 1) \tag{6}$$

being the marginal prior means and variances. It accounts for both the variability in a preexperimental dataset $\boldsymbol{y}_k$ and the postulated level of incommensurability, w_k, through the Gamma mixture prior placed on the predictive precision, ν_k. We give more details in Section D of the Supporting Information for this derivation. Using Bayes' theorem, this collective prior will be updated by the new experimental data, denoted by $\boldsymbol{y}_{K+1}$, to a robust posterior.

2.2 | Criteria for the Bayesian SSD

Most Bayesian SSD criteria aim to control certain property of the posterior, denoted by $f_p(\mu_\Delta \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1})$, wherein $\boldsymbol{y}_{K+1}$ are unobserved at the design stage. It is important to state that uncertainty of sampling a set of data as $\boldsymbol{y}_{K+1}$ from the entire *probability space* needs to be accounted for. Thus, strictly speaking, the Bayesian SSD criteria can only maintain the average properties of the posterior.

Joseph and Bélisle (1997) propose specifying a density region, $R(\boldsymbol{y}_{K+1})$, bounded by r and $r + \ell_0$, to contain possible parameter values. Here, ℓ_0 is the desired interval length and r chosen so that $R(\boldsymbol{y}_{K+1})$ is the highest posterior density (HPD) interval; the so-called HPD because this interval includes the mode of the posterior distribution. This specification can ensure the coverage probability of $R(\boldsymbol{y}_{K+1})$ to be at least $1 - \alpha$, when averaged over all possible samples. Formally, it requires that

$$\int_{\mathcal{Y}} \left\{ \int_r^{r+\ell_0} f_p(\mu_\Delta \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1}) \mathrm{d}\mu_\Delta \right\} f_d(\boldsymbol{y}_{K+1}) \mathrm{d}\boldsymbol{y}_{K+1} \geq 1 - \alpha, \tag{7}$$

where $\mathcal{Y}$ denotes the probability space and $f_d(\boldsymbol{y}_{K+1})$ the marginal distribution of the sample, that is, the new experimental data. For controlling the coverage probability, it is often referred to as the average coverage criterion (ACC). The posterior distribution in our context would be unimodal and symmetric about the posterior mean, as we can envisage from the collective prior given by (5). We would then simply stipulate the HPD interval as

$$R(\boldsymbol{y}_{K+1}) = \mathbb{E}(\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1}) \pm \frac{\ell_0}{2}, \tag{8}$$

which coincides with the alpha-expectation tolerance region by Fraser and Guttman (1956).

An alternative to the ACC is the average length criterion (ALC), which limits the interval length to be at most ℓ for a posterior interval that has a coverage probability of $1-\alpha_0$ (Joseph and Bélisle, 1997). Let $\ell'(\boldsymbol{y}_{K+1})$ be the random interval length of the posterior credible interval dependent on the unobserved new experimental data. Targeting a fixed coverage probability of $1-\alpha_0$, one may solve $\ell'(\boldsymbol{y}_{K+1})$ to meet

$$\int_r^{r+\ell'(\boldsymbol{y}_{K+1})} f_p(\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1}) \mathrm{d}\mu_\Delta = 1-\alpha_0, \tag{9}$$

where r would be specified to give the HPD interval as that for the ACC above. Averaged over all possible samples, the ALC requires that

$$\int_{\mathcal{Y}} \ell'(\boldsymbol{y}_{K+1}) f_d(\boldsymbol{y}_{K+1}) \mathrm{d}\boldsymbol{y}_{K+1} \le \ell. \tag{10}$$

The ALC could be more favored than the ACC, since Bayesian practitioners are keen to report, for example, a 95% credible interval for the posterior mean, in the analysis.

As we can see from (7) and (10), sample sizes chosen to meet the ACC or ALC rely on the marginal, predictive distribution of $\boldsymbol{y}_{K+1}$; that is, $f_d(\boldsymbol{y}_{K+1}) = \int f(\boldsymbol{y}_{K+1} \mid \mu_\Delta)\pi(\mu_\Delta)\mathrm{d}\mu_\Delta$. When $f_d(\boldsymbol{y}_{K+1})$ also depends on nuisance parameters, say the variance σ_0^2 being unknown, it becomes $f_d(\boldsymbol{y}_{K+1}) = \int\int f(\boldsymbol{y}_{K+1} \mid \mu_\Delta, \sigma_0^2)\pi(\mu_\Delta)g(\sigma_0^2)\mathrm{d}\mu_\Delta\mathrm{d}\sigma_0^2$. In our context, priors for unknown μ_Δ and σ_0^2 would be specified based on preexperimental information. The predictive distribution $f(\boldsymbol{y}_{K+1})$ would thus formally be $f_d(\boldsymbol{y}_{K+1} \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K)$, given our $\pi(\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K)$ and $g(\sigma_0^2 \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K)$.

We consider one additional criterion relating to the moments of posterior distribution. For practical reasons, we focus on the second central moment only, so the criterion would be referred to as the average posterior variance criterion (APVC) hereafter. Given a fixed level of dispersion ϵ_0, a suitable sample size is chosen to ensure that

$$\mathbb{E}_{\mathcal{Y}}[\mathrm{Var}(\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1})] \le \epsilon_0. \tag{11}$$

As Adcock (1997) commented, this criterion is equivalent to using the L_2-norm loss function for inferences: $L_2(\mu_\Delta) = (\mu_\Delta - \mathbb{E}(\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1}))^2$. It is also worth noting that the literature also documented many other Bayesian approaches to SSD, for example, based on the use of utility theory (Lindley, 1997) and Bayes factors (Weiss, 1997); the latter is relevant to pursuing the control of type I error rate and power in hypothesis testing problems.

We note that the fixed values of ℓ_0, α_0, and ϵ_0 are all positive real numbers. Unlike the frequentist statistical significance levels, there is no convention to set these thresholds. It is most likely to be backed up by supporting details from the field of application; questions such as what is the meaningful range of μ_Δ that can provide compelling evidence for the inference may be discussed with a subject-matter expert.

2.3 | Sample size required for comparing two normal means

Consider the comparison of two normal means, denoted by μ_j, $j = A, B$, in a new experiment. The difference $\mu_\Delta = \mu_A - \mu_B > 0$ would indicate A is superior to B. Let X_{ij} be the measured outcome from experimental unit i assigned to group j. We assume these measurements are independent, random samples drawn from populations with overall mean μ_j and a common variance σ_0^2. Letting n_j be the groupwise sample sizes, the sample means $\bar{X}_j = (X_{1j} + \cdots + X_{n_j j})/n_j$ follow an asymptotic normal distribution by the central limit theorem; that is, $\bar{X}_j \sim N(\mu_j, \frac{\sigma_0^2}{n_j})$, for $j = A, B$. This further leads to $\bar{X}_\Delta \sim N(\mu_\Delta, \frac{\sigma_0^2}{n_A} + \frac{\sigma_0^2}{n_B})$.

2.3.1 | For cases of known variance

When the common variance σ_0^2 is known, $\mu_\Delta = \mu_A - \mu_B$ has a normal prior based on preexperimental datasets $\boldsymbol{y}_1, \dots, \boldsymbol{y}_K$, as was given in (5). Since the joint likelihood of the $n_A + n_B$ measurements in the new experiment $\mathcal{L}(\boldsymbol{y}_{K+1} | \mu_A, \mu_B) \propto \mathcal{L}(\bar{X}_\Delta | \mu_A, \mu_B)$, we formulate the data likelihood in terms of $\bar{X}_\Delta$, which is regarded as a random variable. We further derive the posterior distribution as

$$\mu_\Delta \mid \boldsymbol{y}_1, \dots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1} \dot{\sim} N\left(\eta, \left(\frac{1}{\sum p_k^2 \xi_k^2} + \frac{1}{\left(\frac{1}{n_A} + \frac{1}{n_B}\right)\sigma_0^2}\right)^{-1}\right), \tag{12}$$

with

$$\eta = \frac{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}{\sum p_k^2\xi_k^2+\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}\sum p_k\lambda_k + \frac{\sum p_k^2\xi_k^2}{\sum p_k^2\xi_k^2+\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}\bar{x}_\Delta, \tag{13}$$

where $\bar{x}_\Delta$ is the realization of $\bar{X}_\Delta$. The marginal distribution (unconditional on μ_Δ) for the difference in sample means is

$$\bar{X}_\Delta \mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_K \dot{\sim} N\left(\sum_k p_k\lambda_k, \left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2+\sum_k p_k^2\xi_k^2\right), \tag{14}$$

which corresponds to $f_d(\boldsymbol{y}_{K+1})$ in Section 2.2.

As Joseph and Bélisle (1997) noted, the ACC and ALC result in the same outcome for cases where the variance is known. Hence, we illustrate using the ACC in the following. Letting the HPD interval $(r, r+\ell_0)$ stretch symmetrically around the posterior mean η, the coverage can be computed by

$$\mathbb{P}\left[|\mu_\Delta-\eta|\le\frac{\ell_0}{2}\mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_K,\boldsymbol{y}_{K+1}\right] = \Phi\left(\sqrt{\frac{1}{\sum p_k^2\xi_k^2}+\frac{1}{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}}\frac{\ell_0}{2}\right), \tag{15}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. We thus have $\Phi(\sqrt{\frac{1}{\sum p_k^2\xi_k^2}+\frac{1}{(\frac{1}{n_A}+\frac{1}{n_B})\sigma_0^2}}\frac{\ell_0}{2}) \ge 1-\alpha$, which is rearranged as

$$\frac{n_An_B}{n_A+n_B} \ge \left(\frac{4z_{\alpha/2}^2}{\ell_0^2}-\frac{1}{\sum p_k^2\xi_k^2}\right)\sigma_0^2, \tag{16}$$

where $z_{\alpha/2}$ is the upper $(\alpha/2)$-th quantile of the standard normal distribution, that is, $\Phi^{-1}(1-\alpha/2)$. Similarly, averaging over the entire data space, the APVC gives

$$\frac{n_An_B}{n_A+n_B} \ge \left(\frac{1}{\epsilon_0}-\frac{1}{\sum p_k^2\xi_k^2}\right)\sigma_0^2. \tag{17}$$

When $\sum p_k^2\xi_k^2$, the prior variance for μ_Δ based on pre-experimental data is so small that the right-hand side of the inequalities above becomes zero or negative, little or no information would be required to accrue from a new experiment.

2.3.2 | For cases of unknown variance

When σ_0^2 is unknown, we assume that the quantity $c\sum p_k^2\xi_k^2/\sigma_0^2 \sim \chi^2(c)$, where $\chi^2(c)$ refers to a chi-square distribution with c degrees of freedom (Gelman *et al.*, 2013). This is equivalent to specifying that $\sigma_0^2 \sim$ Inv-Gamma$(\frac{c}{2}, \frac{c\sum p_k^2\xi_k^2}{2})$; hence, the larger value c takes, the more σ_0^2 converges to the prior variance for $\mu_\Delta \mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_K$. The marginal posterior for μ_Δ will then be obtained by integrating out the nuisance parameter σ_0^2:

$$\begin{aligned} f_p(\mu_\Delta \mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_K,\boldsymbol{y}_{K+1}) &= \int \pi_p(\mu_\Delta,\sigma_0^2\mid \boldsymbol{y}_1,\dots,\boldsymbol{y}_K,\boldsymbol{y}_{K+1})g(\sigma_0^2)\mathrm{d}\sigma_0^2 \\ &\propto \exp\left(-\frac{(\mu_\Delta-\sum p_k\lambda_k)^2}{2\sum p_k^2\xi_k^2}\right)\left[1+\frac{1}{c}\cdot\frac{(\mu_\Delta-\bar{x}_\Delta)^2}{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sum p_k^2\xi_k^2}\right]^{-\frac{c+1}{2}}, \end{aligned} \tag{18}$$

that is, the posterior is proportional to the product of normal and nonstandardized t kernels (Ahsanullah *et al.*, 2014). Detailed steps for deriving (18) are given in Section E of the Supporting Information. In particular, the t density kernel (with the location and scale parameters being $\bar{x}_\Delta$ and $(\frac{1}{n_A}+\frac{1}{n_B})\sum p_k^2\xi_k^2$, respectively) can be related to a normal kernel with the same location parameter and the variance as $(\frac{1}{n_A}+\frac{1}{n_B})\sigma_0^2$, conditional on $c\sum p_k^2\xi_k^2/\sigma_0^2 \sim \chi^2(c)$. The posterior (18) can thus be further developed as

$$\begin{aligned} f_p(\mu_\Delta \mid \sigma_0^2,\boldsymbol{y}_1,\dots,\boldsymbol{y}_K,\boldsymbol{y}_{K+1}) &\propto \exp\left(-\frac{(\mu_\Delta-\sum p_k\lambda_k)^2}{2\sum p_k^2\xi_k^2}\right)\exp\left(-\frac{(\mu_\Delta-\bar{x}_\Delta)^2}{2\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}\right) \\ &\overset{\text{def}}{=} \exp\left(-\frac{(\mu_\Delta-\mu_N)^2}{2\sigma_N^2}\right), \end{aligned} \tag{19}$$

with

$$\sigma_N^2 = \left(\frac{1}{\sum p_k^2\xi_k^2}+\frac{1}{\left(\frac{1}{n_A}+\frac{1}{n_B}\right)\sigma_0^2}\right)^{-1}, \tag{20}$$

which is consistent with (12) but here with unknown $\sigma_0^2 \sim$ Inv-Gamma$(\frac{c}{2}, \frac{c\sum p_k^2\xi_k^2}{2})$. We can also find the distribution

for $\bar{X}_\Delta$ unconditional on μ_Δ as

$$\bar{X}_\Delta \mid \sigma_0^2, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K \dot{\sim} N\left(\sum p_k \lambda_k, \left(\frac{1}{n_A} + \frac{1}{n_B} \right) \sigma_0^2 + \sum_k p_k^2 \xi_k^2 \right); \tag{21}$$

see the derivation also in Section E of the Supporting Information. Apparently, this marginal distribution for $\bar{X}_\Delta$ relies on prior distribution for the unknown σ_0^2, which may yield different solutions of n_A and n_B across the Bayesian SSD criteria considered in this paper.

Let the interval $(a, a + \ell_0)$ be symmetric about μ_N given the marginal posterior for μ_Δ in (19). The sample size is found requiring $\mathbb{P}[|\mu_\Delta - \mu_N| \leq \frac{\ell_0}{2} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_K, \boldsymbol{y}_{K+1}] \geq 1 - \alpha$, based on the ACC; thus

$$\frac{\ell_0}{2\sigma_N} = \sqrt{\frac{1}{\sum p_k^2 \xi_k^2} + \frac{1}{\left(\frac{1}{n_A} + \frac{1}{n_B} \right) \sigma_0^2}} \cdot \frac{\ell_0}{2} \geq z_{\alpha/2}, \tag{22}$$

where $z_{\alpha/2}$ denotes the upper $(\alpha/2)$-th quantile of the standard normal distribution. We rewrite the expression and obtain

$$\frac{n_A n_B}{n_A + n_B} \geq \left(\frac{4 z_{\alpha/2}^2}{\ell_0^2} - \frac{1}{\sum p_k^2 \xi_k^2} \right) \int_0^\infty \sigma_0^2 g(\sigma_0^2) \mathrm{d}\sigma_0^2, \tag{23}$$

where $g(\sigma_0^2)$ is the probability density function of an Inv-Gamma$(\frac{c}{2}, \frac{c \sum p_k^2 \xi_k^2}{2})$ distribution. The reader may compare this inequality with what was obtained for cases where σ_0^2 is known in (16).

Applying the ALC, we need to average the random credible interval length $\ell'(\bar{x}_\Delta) = 2 z_{\alpha_0/2} \sigma_N$ over the marginal distribution for $\bar{X}_\Delta$ which varies with σ_0^2. According to the definition of ALC, we obtain that

$$2 z_{\alpha_0/2} \int_0^\infty \left(\frac{1}{\sum p_k^2 \xi_k^2} + \frac{1}{\left(\frac{1}{n_A} + \frac{1}{n_B} \right) \sigma_0^2} \right)^{-\frac{1}{2}} g(\sigma_0^2) \mathrm{d}\sigma_0^2 \leq \ell, \tag{24}$$

which does not have a closed-form solution. This requires a search over the integers for n_A and n_B to find the smallest sum that satisfies the inequality. With the APVC, we would likewise remove the dependence on σ_0^2 by integration. The formula thus becomes

$$\frac{n_A n_B}{n_A + n_B} \geq \left(\frac{1}{\epsilon_0} - \frac{1}{\sum p_k^2 \xi_k^2} \right) \int_0^\infty \sigma_0^2 g(\sigma_0^2) \mathrm{d}\sigma_0^2. \tag{25}$$

3 | APPLICATION

Hampson *et al.* (2014) present a Bayesian approach for elicitation of expert opinion on model parameters for enhanced design and analysis of rare-disease trials. An elicitation meeting (Hampson *et al.*, 2015) was held for the MYPAN trial, which compares the efficacy of a new treatment (labeled A) relative to the standard of care (labeled B) for polyarteritis nodosa, a rare and severe inflammatory blood vessel disease. Priors were elicited from the input of 15 experts individually. Specifically, opinion was sought on (i) the probability that a patient given B would achieve disease remission within 6 months (a dichotomous event) and (ii) the log-odds ratio of remission rates. Consensus distributions for the remission rates were obtained, with the mode at 71% for A and 74% for B.

In line with the original assumptions for the MYPAN trial, we suppose the Bernoulli probability is not close to 0 or 1, so the log-odds ratio of treatment benefit, that is, $\theta_k = \log[(\rho_{Ak}(1 - \rho_{Bk}))/((1 - \rho_{Ak})\rho_{Bk})]$, would be approximately normally distributed (Agresti, 2003). Here, ρ_{jk} denotes the probability of remission for patients receiving treatment $j = A, B$. We regard the expert opinion as a type of pretrial information and further assume it had been summarized in the form of $\theta_k \mid \boldsymbol{y}_k \sim N(m_k, s_k^2)$, $k = 1, \ldots, K$. Eliciting such expert opinion is a nontrivial problem; we refer the interested reader to the literature such as Dias *et al.* (2017). Furthermore, Hampson *et al.* (2014) detailed the elicitation process for reaching a probabilistic summary for the log-odds ratio. For illustrative purposes, we assume five sets of expert opinion had been summarized as $N(-0.26, 0.25)$, $N(-0.24, 0.23)$, $N(-0.37, 0.22)$, $N(-0.34, 0.36)$, and $N(-0.32, 0.26)$ to inform θ_k. The opinion would also be sought on w_k, $k = 1, \ldots, 5$, to represent the experts' skepticism about the predictability of each pretrial parameter θ_k towards the parameter μ_Δ, measured on the continuous scale of 0 to 1. In this example, we suppose such pretrial information is valued about equally, with $w_1 = 0.15$, $w_2 = 0.20$, $w_3 = 0.17$, $w_4 = 0.13$, $w_5 = 0.20$. In practice, the trial statistician could look into the levels of pairwise commensurability between the $N(m_k, s_k^2)$ distributions through a discrepancy measure, such as the Hellinger distance (Dey and Birmiwal, 1994), to reconcile the choices of value for w_k.

For reaching a collective prior for $\mu_\Delta \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_5$, synthesis weights $p_1, \ldots, p_5$ need to be specified. We apply a decreasing function:

$$p_k = \frac{\exp(-w_k^2/s_0)}{\sum_k \exp(-w_k^2/s_0)}, \tag{26}$$

with a concentration parameter s_0 to transform these weights from $w_1, \ldots, w_K$. Specifically, for $s_0 \gg w_k$, all p_k

will be close to $1/K$ irrespective of the values of w_k. Whereas, with $s_0 \to 0^+$, the smallest w_k would have $p_k \to 1$, meaning that the corresponding $\theta_k \mid \boldsymbol{y}_k$ tends to dominate the collective prior. The rationale behind this approach is that both w_k and p_k might be determined by some distance measure between parameters θ_k and μ_Δ. It is an objective-directed approach, since we hope to discount preexperimental information to a larger extent via small values of p_k, when it is believed a priori to be less commensurate (thus, large values of w_k) with the new experimental data. Figures S2 and S3 (in the Supporting Information) visualize the impact of w_k, $k = 1, \ldots, 5$, and s_0 on the informativeness of the collective prior. A thorough evaluation by Zheng and Wason (2022) shows this objective-directed approach has desirable properties. We generally recommend choosing a small value (relative to the magnitudes of w_k) for s_0, particularly because this can discern the degree of relevance and can further lead to a heavy-tailed collective prior for cases of divergent pretrial information. Here, we set $s_0 = 0.05$ for illustration; consequently, $p_1 = 0.23$, $p_2 = 0.16$, $p_3 = 0.20$, $p_4 = 0.25$, $p_5 = 0.16$. This gives a collective prior $\mu_\Delta \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_5 \sim N(-0.309, 0.154)$, when specifying $\nu_k \sim w_k\text{Gamma}(2,2) + (1-w_k)\text{Gamma}(18,3)$ for our model.

Assuming known variance of $\sigma_0^2 = 0.35$ and that $n_A = n_B$, the total sample sizes (i.e., $n_A + n_B$) found based on the ACC and ALC criteria are both 41.8 for 95% posterior coverage probability and the credible interval length as 0.65 on average. For cases of unknown σ_0^2, we let $\sigma_0^2 \sim$ Inv-Gamma$(2.500, 0.385)$ (i.e., setting $c = 5$). The ACC and ALC sample sizes become 30.7 and 24 for attaining the same posterior behaviors, respectively. Targeting $\epsilon_0 = 0.03$, the APVC sample sizes are 32.2 and 27.6 for known and unknown σ_0^2, respectively.

It may be counterintuitive to find the sample size for cases of unknown variance is smaller than those for known variance here, especially if the latter is perceived as a version of the former with infinite precision. We would reiterate that the prior specification for σ_0^2 in our methodology uses pretrial information, via an Inv-Gamma$(\frac{c}{2}, \frac{c\sum p_k^2 \xi_k^2}{2})$ distribution. Taking the mode for illustration, the sample size would be proportional to the quantity $\frac{c}{c+2}\sum p_k^2 \xi_k^2$, that is, the magnitude of the collective prior variance (i.e., 0.154 in this illustration) scaled by the constant relying on c. This is smaller than the fixed $\sigma_0^2 = 0.35$; so not surprisingly, a smaller sample size would be yielded by the same criterion. We also caution that the distribution is not necessarily symmetric about the mode, and the uncertainty in σ_0^2 needs to be integrated out for the formal SSD.

4 | PERFORMANCE EVALUATION

4.1 | Basic settings

Motivated by the MYPAN trial, we generate four base scenarios of historical data, which are configured with different levels of pairwise (in)commensurability and informativeness. Such preexperimental information from K sources is supposed to have been summarized as $\theta_k \mid \boldsymbol{y}_k \sim N(m_k, s_k^2)$, $k = 1, \ldots, K$. For each base scenario, two distinct sets of prior mixture weights I and II for robust borrowing are considered to implement the proposed approach for borrowing of information, as listed in Table 1. These fractions are chosen to (a) reflect high and low levels of prior confidence in the historical data when they are consistent between themselves or (b) designate a certain source of historical data to be more influential.

We compute the Hellinger distances of any two $N(m_k, s_k^2)$ distributions to describe their pairwise (in)commensurability, as visualized in Figure S4 of the Supporting Information. This is used to justify the values of w_k in Table 1 for our numerical study being no greater than 0.500, as the largest Hellinger distance in Figure S4 is below 0.500. Both the Gamma mixture prior for ν_k, and derivation of the weights, p_k, for prioritizing certain historical data to form a collective prior, follow our specification in Section 3. Nonetheless, we note at the outset the Gamma component distributions can be equally essential, as choices have an impact on the effective sample size of the collective prior (Neuenschwander *et al.*, 2020).

We compare the sample sizes computed using the proposed Bayesian SSD formulae with those computed (a) without robustification, that is, setting each $w_k = 0$ for $k = 1, \ldots, 5$, (b) without leveraging historical information for μ_Δ, that is, setting each $w_k = 1$, (c) from the proper Bayesian SSD approach driven by a single prior, here specified as the most informative $N(m_k, s_k^2)$, for example, $N(-0.37, 0.22)$ for configuration 1, and (d) from an optimal approach as the benchmark. Specifically, the optimal approach is coupled with a perfectly commensurate prior, by equating σ_0^2 to the collective prior variance $\sum p_k^2 \xi_k^2$. In this way, the corresponding result would serve as the benchmark referring to the scenario of perfect consistency between the collective prior and the new data, so the largest saving in sample size could be attained by the proposed methodology. For cases of unknown σ_0^2, the optimal sample sizes could be approached by setting c to a sufficiently large value.

TABLE 1 Configurations of hypothetical historical data, each accompanied by two sets of weights for robust borrowing of information. preexperimental information about $\theta_k \mid \boldsymbol{y}_k$ is assumed to have been summarized by a $N(m_k, s_k^2)$ prior for $k = 1, \ldots, 5$

			Hypothetical historical data						
			$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$\sum p_k \lambda_k$	$\sum p_k^2 \xi_k^2$
Configuration 1		m_k	−0.260	−0.240	−0.370	−0.340	−0.320		
		s_k^2	0.250	0.230	0.220	0.360	0.260		
	Robust weights I	w_k	0.103	0.175	0.081	0.143	0.077	−0.311	0.129
		p_k	0.214	0.143	0.232	0.176	0.235		
	Robust weights II	w_k	0.252	0.319	0.140	0.306	0.149	−0.325	0.198
		p_k	0.149	0.069	0.359	0.082	0.341		
Configuration 2		m_k	−0.260	−0.240	−0.370	−0.340	−0.320		
		s_k^2	0.100	0.100	0.100	0.100	0.100		
	Robust weights I	w_k	0.103	0.175	0.081	0.143	0.077	−0.311	0.096
		p_k	0.214	0.143	0.232	0.176	0.235		
	Robust weights II	w_k	0.252	0.319	0.140	0.306	0.149	−0.325	0.158
		p_k	0.149	0.069	0.359	0.082	0.341		
Configuration 3		m_k	−0.260	−0.170	−0.440	−0.150	0.120		
		s_k^2	0.250	0.640	0.970	1.540	0.590		
	Robust weights I	w_k	0.101	0.219	0.385	0.385	0.304	−0.198	0.295
		p_k	0.559	0.263	0.035	0.035	0.108		
	Robust weights II	w_k	0.325	0.203	0.171	0.180	0.272	−0.215	0.379
		p_k	0.065	0.235	0.298	0.280	0.122		
Configuration 4		m_k	−0.260	−0.170	−0.440	−0.150	0.120		
		s_k^2	0.250	0.150	0.400	0.890	0.220		
	Robust weights I	w_k	0.066	0.303	0.459	0.355	0.115	−0.099	0.226
		p_k	0.473	0.082	0.008	0.041	0.396		
	Robust weights II	w_k	0.537	0.306	0.054	0.220	0.350	−0.312	0.343
		p_k	0.002	0.098	0.602	0.243	0.055		

4.2 | Results

Figure 1 visualizes a subset of the results, which compare the proposed Bayesian SSD formulae using robust weights I and II with the alternative approaches for cases of known and unknown σ_0^2, respectively. Here, we assume $\sigma_0^2 = 0.35$ and, if unknown, $\sigma_0^2 \sim$ Inv-Gamma(1.5, $1.5\times \sum p_k^2 \xi_k^2$) for illustration. We fix the posterior credible interval length $\ell_0 = 0.65$ to find the ACC sample sizes, so that the average coverage probability would be 95%, that is, targeting $\alpha = 0.05$ in (16). Likewise, for computing the ALC sample sizes, we fix $\alpha_0 = 0.05$ and constrain the average length of the posterior credible interval below 0.65. When applying the APVC, sample sizes are found with the average posterior variance retained to level $\epsilon = 0.03$.

In all configurations 1–4, we see that the sample sizes computed according to the same criterion, using robust weights I, are smaller than those using robust weights II. This is because following our setting the collective prior, produced by robust weights I, has a smaller variance than its counterpart by robust weights II, for each configuration. Moreover, sample sizes yielded using either robust weights I or II are always bounded by those using no robustification ($w_k = 0$) and no borrowing ($w_k = 1$). We may think that no robustification leads to the least conservative result by the proposed SSD formulae, for the given historical information fully used. These, however, are not necessarily identical to the optimal situations, where σ_0^2 is equated to the collective prior variance, or largely determined by the latter if unknown. In Figure 1, we omit the benchmark optimal sample sizes that may be obtained by using the proposed formulae with robust weights I and II for each configuration. Yet we will comment on the maximal saving that the proposed SSD approach can achieve in the following along with other figures.

The height difference across bars of sample sizes, computed using our approach with robust weights I or II and no borrowing ($w_k = 1$), quantifies the benefit from leveraging preexperimental information for μ_Δ. Looking across subfigures (i) and (ii), such height differences between methods are far greater for the unknown variance case than the known variance case. Comparison of SSD

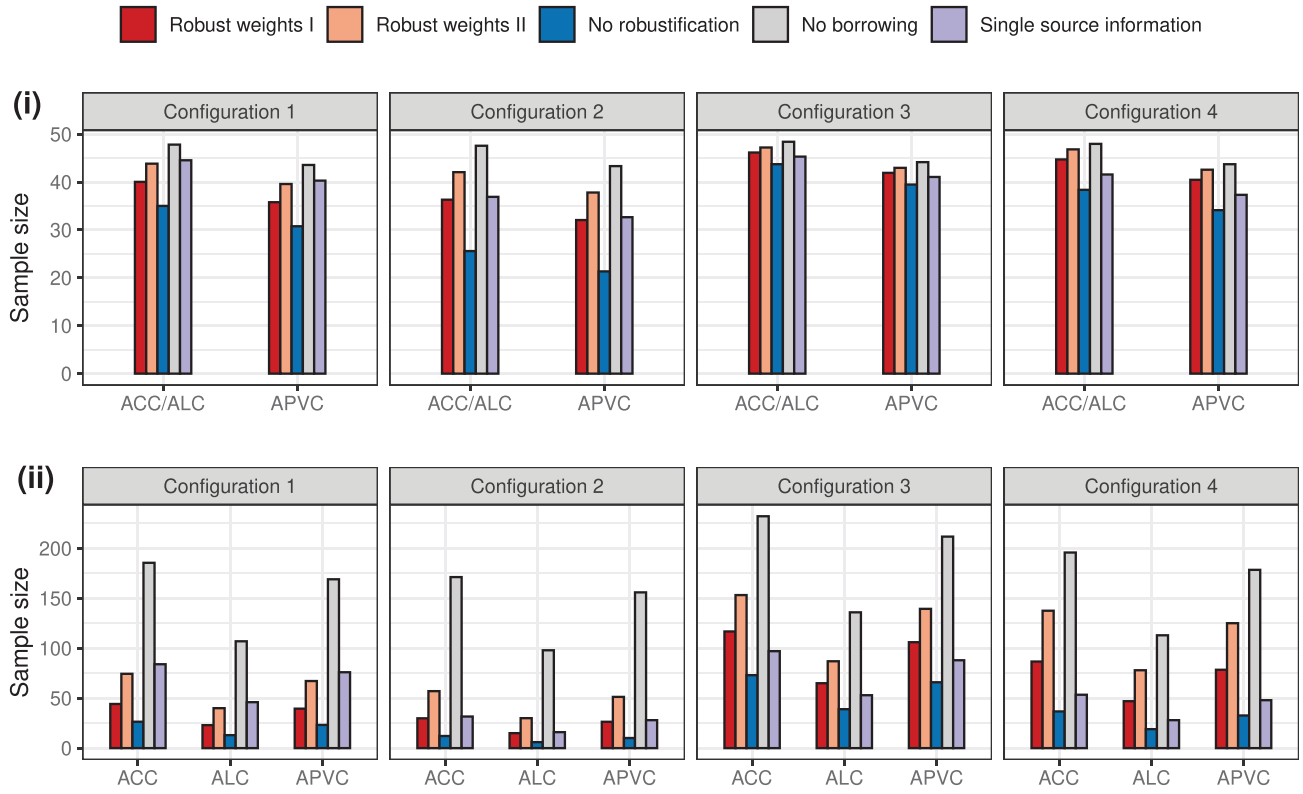

FIGURE 1 Comparison of the Bayesian SSD approaches in terms of the sample size obtained according to the ACC, ALC, and APVC criteria for cases of (i) known $\sigma_0^2 = 0.35$ and (ii) unknown σ_0^2. Sample sizes in subfigure (ii) for unknown σ_0^2 are computed setting $c = 3$, that is, assuming that $\sigma_0^2 \sim \text{Inv-Gamma}(1.5, 1.5 \times \sum p_k^2 \xi_k^2)$, for fairly limited use of preexperimental information to inform the variance σ_0^2. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

approaches with borrowing versus no borrowing, as visualized in subfigure (ii) of Figure 1, would be more objective for illustrating the benefit. As mentioned, choosing $c = 3$ means σ_0^2 would be related with $\sum p_k^2 \xi_k^2$ to a very limited extent, as if a diffuse prior had been placed on σ_0^2. Thereby, implementing no borrowing by setting $w_k = 1$, preexperimental information would neither be leveraged through the robust prior for μ_Δ, nor through the prior for the unknown $\sigma_0^2 \sim \text{Inv-Gamma}(\frac{c}{2}, \frac{c \sum p_k^2 \xi_k^2}{2})$. Consequently, larger sample sizes would be found for no borrowing SSD for the unknown σ_0^2 than the known cases assuming $\sigma_0^2 = 0.35$, to retain similar properties of the posterior distribution. Focusing on the bars for robust weights I and II against no borrowing within subfigure (ii), saving in all the ACC, ALC, and APVC sample sizes could be as much as two-thirds for configurations 1 and 2. Such saving is attenuated in configurations 3 and 4 when historical information is divergent. In configurations 3, the ACC (ALC) sample size obtained from the no borrowing approach is about twice the size from the proposed approach with robust weights I, specifically, 232.2 versus 116.8 (136 vs. 65), respectively. We observe a small increase in sample size by using robust weights II instead of I, because slightly higher prior probabilities of incommensurability had been allocated to certain informative $N(m_k, s_k^2)$ for greater down-weighting. The trend is similar for results in configuration 4.

We then compare the proposed approach with an alternative strategy, that is, restricting the use of preexperimental information from a single source. When the historical data are consistent (divergent) between themselves, the proposed SSD formulae lead to smaller (larger) sample sizes, as presented obviously in configuration 1 (configurations 3 and 4) for both cases of known and unknown σ_0^2. As one may perceive, such selection of a single source could be less robust than averaging over all available preexperimental information. Another noteworthy finding is concerned with the comparison of the ACC and ALC sample sizes, particularly when σ_0^2 is unknown and we place a minimally informative prior on it (setting $c = 3$). As shown in Figure 1, the ALC sample size is universally smaller than the ACC sample size for all these investigated configurations.

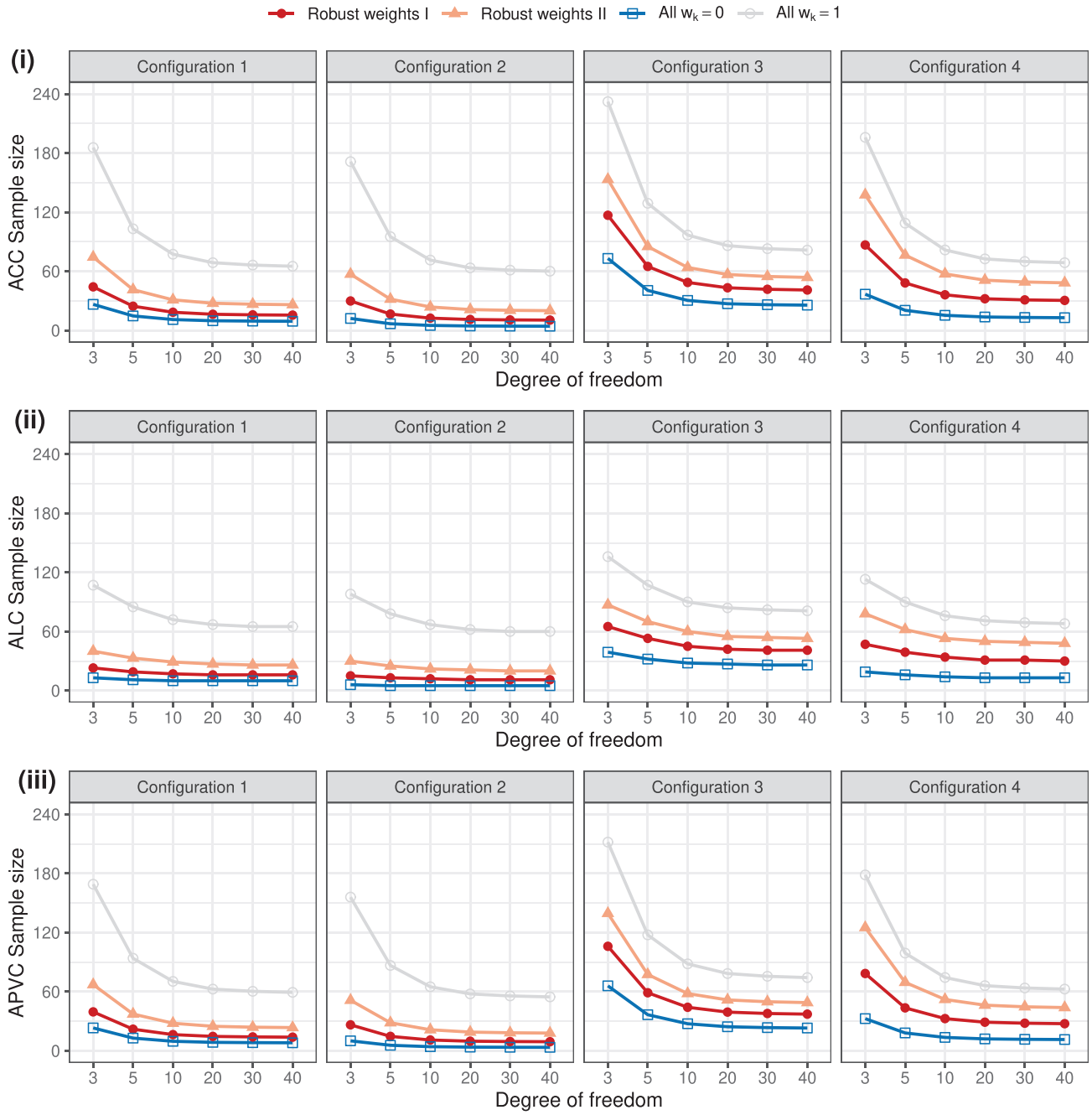

FIGURE 2 The ACC, ALC, and APVC sample sizes for the new trial, where the unknown σ_0^2 could be related to the collective prior variance by assuming the quantity $c\sum p_k^2\xi_k^2/\sigma_0^2 \sim \chi^2(c)$. The extent of borrowing for better knowledge about σ_0^2 depends on the number of degrees of freedom, c. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

We move on to quantify how the sample sizes would vary as c changes. Focusing on approaches using preexperimental information from multiple sources, Figure 2 displays the sample sizes exclusively for cases of unknown $\sigma_0^2 \sim$ Inv-Gamma$(\frac{c}{2}, \frac{c\sum p_k^2\xi_k^2}{2})$. We set $c = 3, 5, 10, 20, 30, 40$, and keep the target level of each SSD criterion unchanged from what we have used for Figure 1. As c gets larger, the sample sizes for all approaches investigated here decrease and tend to stabilize at their own lowest levels possible. This could be explained from the perspective of prior effective sample size, to which variance is a key determining factor. Consider the prior placed on the inverse of the unknown variance that $\frac{1}{\sigma_0^2} \sim$ Gamma$(\frac{c}{2}, \frac{c\sum p_k^2\xi_k^2}{2})$, of which the mean and variance are $\frac{1}{\sum p_k^2\xi_k^2}$ and $\frac{2}{c}\cdot\frac{1}{(\sum p_k^2\xi_k^2)^2}$, respectively. As c

increases, the prior variance diminishes, meaning that possible values of $\frac{1}{\sigma_0^2}$ are more concentrated around the prior mean obtained based on historical data. For $c \geq 20$, the ACC and ALC sample sizes are nearly identical. Whereas, the ACC sample size is more sensitive than the ALC to small values of c, for example, when $c = 3, 5$. We note that the so-called "no borrowing" (by setting $w_k = 1$) should be clarified as no borrowing in terms of the parameter μ_Δ. When c gets larger, it means the unknown variance σ_0^2 would be more closely tied to the prior variance based on the historical data. That is, borrowing is enabled through the variance, although not directly the parameter of inferential interest. By fixing $w_k = 1$, historical data would not be leveraged through the robust prior for μ_Δ, but nevertheless could be used to inform the unknown σ_0^2, particularly when c is sufficiently large.

Figure 3 illustrates how the sample size varies, for cases of unknown σ_0^2, when targeting the average coverage probability, posterior credible length, and posterior variance at different levels. Like in Figure 1, these results are obtained by setting $c = 3$ for the very limited use of preexperimental information to inform σ_0^2. The optimal sample sizes are also plotted to show the maximal saving the proposed SSD formulae may achieve. Specifically, Optim I and II should be taken as the benchmark for formulae using robust weights I and II, respectively. As expected, sample sizes by robust weights I and II would always be bounded by the extremes of no robustification (all $w_k = 0$) and no borrowing (all $w_k = 1$). Given a fixed length $\ell_0 = 0.65$ of the HPD interval, more ACC sample sizes would be required if increasing the desired coverage probability on average, $1 - \alpha$. For example, the ACC sample size computed using robust weights I (II) rises from 78.7 to 156.5 (104.4 to 204.2) for configuration 3 had the level of $1 - \alpha$ been lifted from 90% to 97.5%. The displayed ALC sample sizes in subfigure (ii) ensure the coverage probability as 95%; by relaxing the target average HPD interval length, fewer sample sizes would be needed. Likewise, the APVC sample sizes in subfigure (iii) share this commonality of decreasing as we relax the target posterior variance. Generating these plots would be helpful in practice for balancing between obtaining an economic sample size planning and a posterior sufficiently informative for inferences on a case-by-case basis. For example, targeting the average length of the HPD interval with 95% coverage probability as $\ell = 0.60$ requires the ALC sample size to be 28 for configuration 1 using robust weights I, which may not be much different from 23 yielded by the level $\ell = 0.65$.

We further investigate the impact of s_k^2, the associated levels of uncertainty inherent to historical data $k = 1, \ldots, K$, on the respective sample sizes. Configurations 1 and 2, with the robust weights kept the same, have been constructed for this purpose. From Figures 1–3, it is clear that Configuration 1 requires a larger sample size than Configuration 2 under the same criterion. The explanation is that Configuration 2, with smaller sample variation, leads to a more informative collective prior for μ_Δ, so less information (sample size) would be required from the new experiment for the inference.

We also examine how sensitive the proposed Bayesian SSD formulae are to the Gamma mixture components. Since a suitable yet least informative Gamma(a_{01}, b_{01}) has been chosen for down-weighting, the other component of the mixture prior, Gamma(a_{02}, b_{02}), determines the maximum borrowing possible. Assuming unknown σ_0^2 and setting $c = 3$, Figure 4 shows the Bayesian SSD under different choices of the hyperparameters, a_{02} and b_{02}, for each criterion. As expected, a more informative Gamma(a_{02}, b_{02}) yields a smaller sample size given the same set of w_k, $k = 1, \ldots, K$. The ALC sample sizes appear to have least decreasing, compared with the ACC and APVC, in this sensitivity evaluation. We also observe that the reduction in Bayesian sample sizes is not proportional to the improving of informativeness of Gamma(a_{02}, b_{02}): setting the informative component as Gamma(18, 3) is not much different from Gamma(54, 3) for our illustrative examples. For practical implementation, we recommend the component Gamma distributions to be chosen for representing two extremes of very limited borrowing and complete pooling of information, when given a full prior mixture weight $w_k = 1$ and $w_k = 0$, respectively.

Finally, comprehensive simulation studies have been performed in Section H–J of the Supporting Information to investigate (i) the average properties of the posterior for μ_Δ as updated by the new experimental data, (ii) the sensitivity to nonnormal data, and (iii) the performance if original priors (without normal approximation) are used for the analysis.

5 | DISCUSSION

Planning a new experiment with a sufficient sample size necessitates the use of relevant information. Bayesian methods allow for the inherent uncertainty in the estimate of model parameters, as well as a formal incorporation of any expert opinion or historical data. In this paper, we have developed Bayesian sample size formulae that use commensurate priors to leverage preexperimental data, available from multiple sources, for the model parameter(s) of interest. While we note proposals based on the "two-prior" approach (De Santis, 2007; Brutti *et al.*, 2009), the proposed method specifies a singular prior for both the design and analysis of the new experiment.

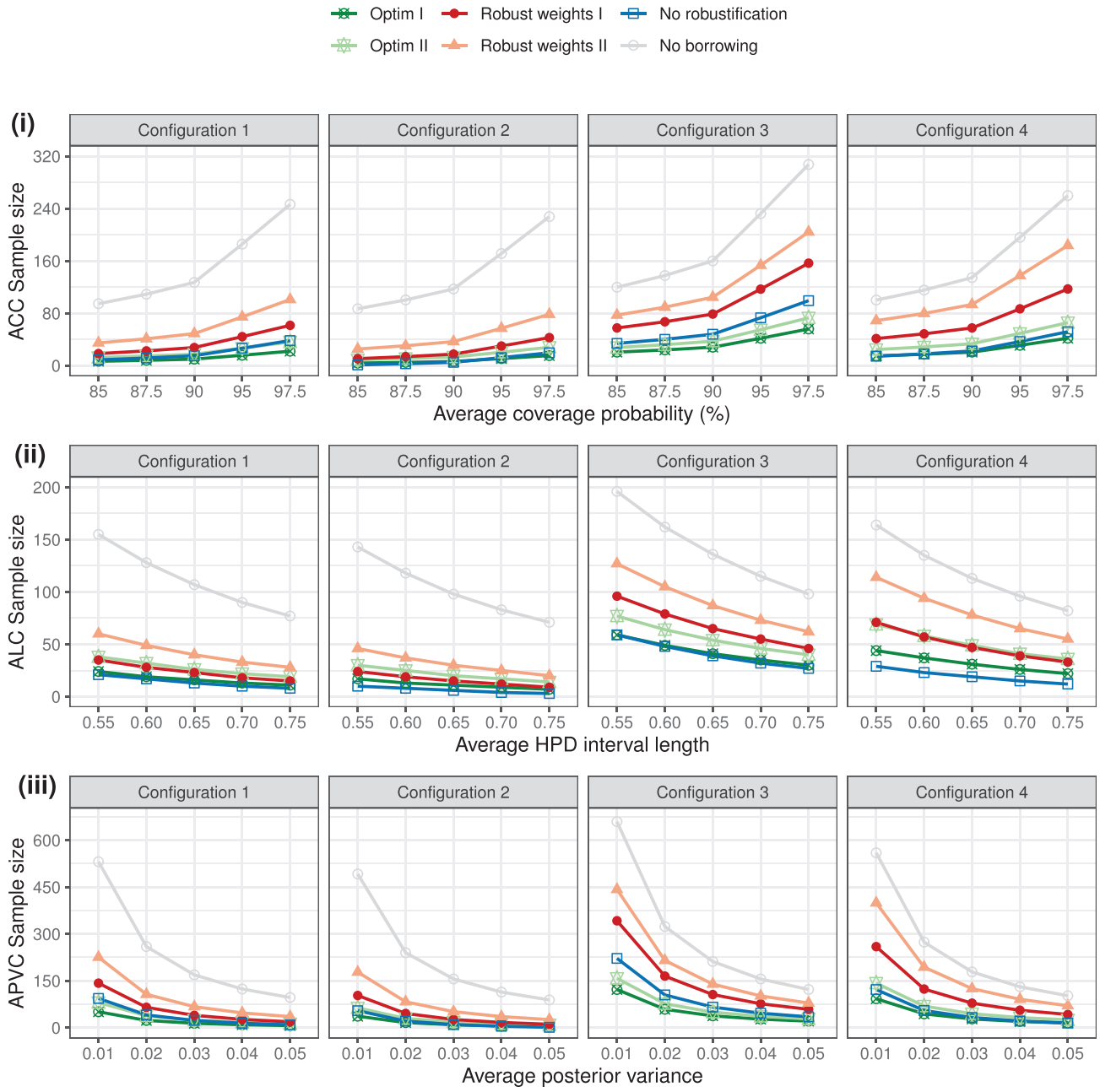

FIGURE 3 Sample sizes required when σ_0^2 is unknown to retain the desired average property of the posterior distribution. The ACC and ALC sample sizes are computed by fixing the credible interval length $\ell_0 = 0.65$ and coverage probability $1 - \alpha_0 = 95\%$, respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

One area that deserves more investigation is surrounding w_k. Following Zheng and Wason (2022), we recommend these are based on some measures of distributional discrepancy, such as the Hellinger distance between any two $N(m_k, s_k^2)$ distributions. The underlying logic is that the new experiment, at the planning stage, is regarded as compatible with the historical experiments, then their data would also be. The levels of the (in)commensurability between a preexperimental parameter and the new experimental parameter would thus be comparable to those between the preexperimental parameters themselves. Nevertheless, we recognize that these prior mixture weights w_k cannot be correctly specified when the new experimental data are yet to be generated. Pragmatically, the new experiment could be embedded with interim analyses to enable midcourse modifications towards w_k. Each update in terms of w_k tends to better reflect the genuine incommensurability (Zheng and Hampson, 2020).

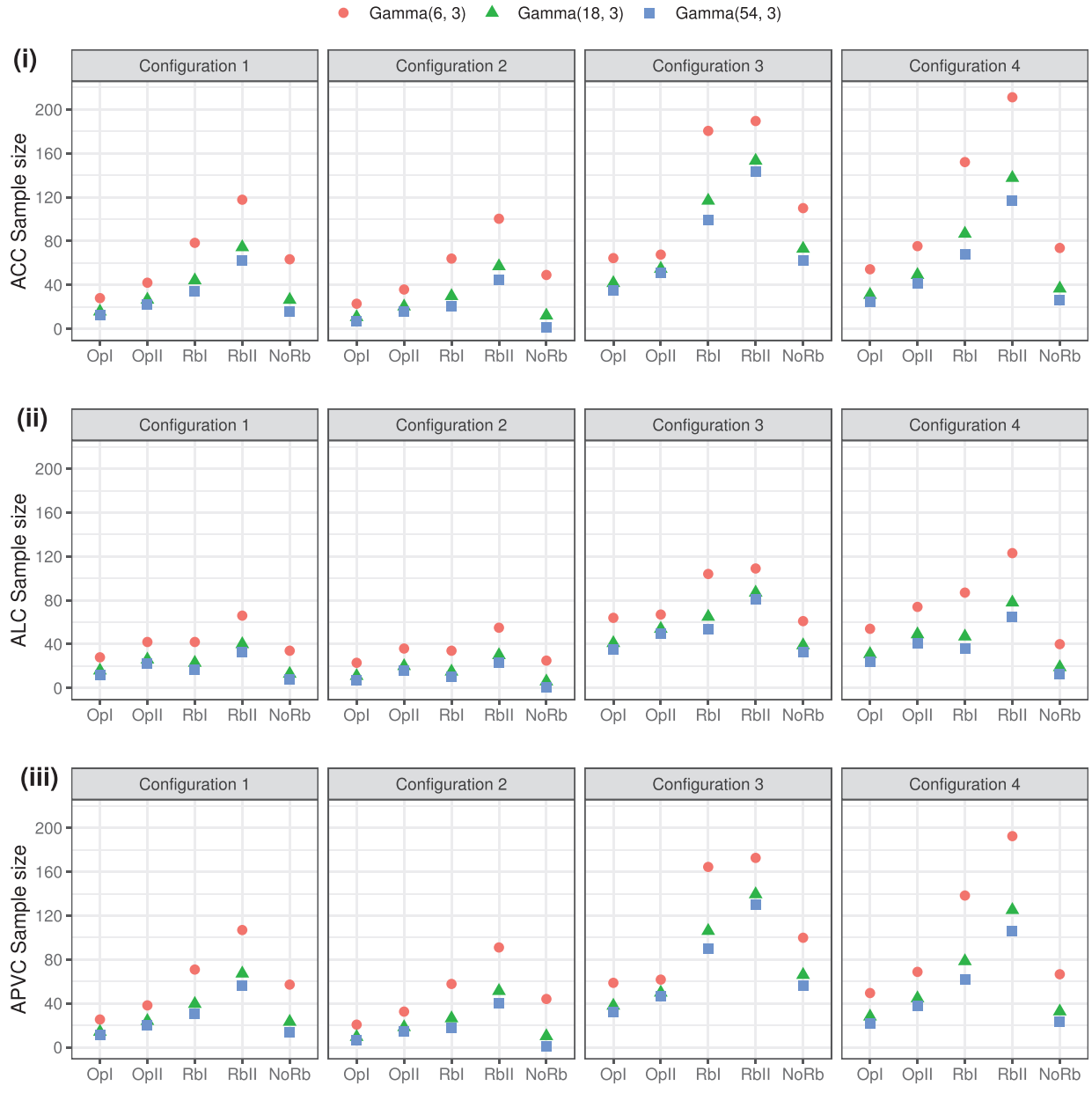

FIGURE 4 The proposed Bayesian SSD is dependent on the choice of the informative Gamma component distribution for strong borrowing. The labels at the x-axis are short for Optim I, Optim II, Robust weights I, Robust weights II, and no robustification, respectively. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

As noted by one reviewer, there are circumstances that preexperimental information may be available for a single arm (say, to inform μ_A or μ_B) only. The proposed Bayesian methodology can still be useful in that the information may be represented into a commensurate predictive prior to the arm-based statistic(s). Analytical derivation of a posterior for the mean difference can follow our one presented in Section 2. This would be particularly relevant to the special topic of using historical control in clinical trials to supplement or replace a concurrent control. However, we caution that the selection of relevant pretrial data on one arm needs to be done carefully, since the model may introduce systematic difference between arms that would affect the inference of the difference in means.

For comparing two Bernoulli probabilities in Section 3, we used a logit transformation to consider the log-odds ratio, which is generally adequately modeled by a normal distribution. The approach of constructing a normal statistic can also be used for time-to-event data, which is elaborated upon in Section K of the Supporting

Information with new formulae presented. We are aware of the limitations. For example, accurate estimation of the Bernoulli probabilities is not straightforward and the censoring assumptions in the time-to-event data are simplified. We hope this work motivates further research for SSD in both binomial and time-to-event data within this Bayesian context.

Throughout this paper, we supposed preexperimental information had been available with regard to the parameter of influential interest. Situations may be more complex in practice. For instance, historical data may have been recorded on a different measurement scale (Zheng *et al.*, 2020) from what might be for the new experiment under planning. This is an area where our future research would look towards. To promote the uptake of our methodology, we have summarized the necessary actions, along with the specification of key parameters, at different stages of the planning of a new experiment in Section L of the Supporting Information. As a separate note, we applied quite general criteria such as ACC and ALC to control the average coverage probability or length of the HPD interval of the posterior distribution for the parameter of influential interest throughout. In such decision frameworks, the sample size largely depends on the informativeness of a prior distribution for μ_Δ, as well as for σ_0^2 when using preexperimental data to inform the variance. With each criterion concerning an average property of the posterior distribution, permitting borrowing (with $0 < w_k < 1$) yields a smaller sample size than the approach of no borrowing (which can be a limiting case of the proposed model with $w_k = 1$). However, when alternative decision criteria are applied, it is not necessarily true that enabling borrowing always leads to a sample size reduction. An example is research for overcoming prior-data conflict, where the prior mismatches the data accrued from the new experiment. There is relevant literature addressing the issue in clinical trials, where maintaining strong control of error rates is required by regulatory agencies (EMA, 1998). Our sample size formulae according to the ACC can be closely relevant for giving a solution analogous to the frequentist hypothesis testing; for example, rejection of the null hypothesis could be defined based on posterior interval probabilities with respect to a magnitude of effect deemed clinically meaningful (Whitehead *et al.*, 2008).

ACKNOWLEDGMENTS

This work was supported by Cancer Research UK through Dr. Zheng's Population Research Postdoctoral Fellowship (RCCPDF\100008). JW and TJ received funding from the UK Medical Research Council (MC_UU_00002/6, MC_UU_00002/14). This report is independent research arising in part from Prof. Jaki's Senior Research Fellowship (NIHR-SRF-2015-08-001) supported by the National Institute for Health Research. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health and Social Care (DHSC).

DATA AVAILABILITY STATEMENT

The authors confirm that the simulated data supporting the findings of this paper are reproducible with openly available R code in the Supporting Information.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and code are available at https://github.com/haiyanzheng/SSDcmspriors.

ORCID

Haiyan Zheng https://orcid.org/0000-0002-3385-2117
James M.S. Wason https://orcid.org/0000-0002-4691-126X

REFERENCES

Adcock, C.J. (1997) Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 261–283.

Agresti, A. (2003) *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Hoboken. NJ: Wiley.

Ahsanullah, M., Kibria, B. and Shakil, M. (2014) *Normal and Student's t Distributions and Their Applications*. Atlantis Studies in Probability and Statistics. Paris: Atlantis Press.

Brutti, P., De Santis, F. and Gubbiotti, S. (2009) Mixtures of prior distributions for predictive Bayesian sample size calculations in clinical trials. *Statistics in Medicine*, 28, 2185–2201.

Clarke, B. and Yuan, A. (2006) Closed form expressions for Bayesian sample size. *Annals of Statistics*, 34, 1293–1330.

De Santis, F. (2007) Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170, 95–113.

Desu, M.M. and Raghavarao, D. (1990) *Sample Size Methodology*. Statistical Modeling and Decision Science. San Diego, CA: Academic Press.

Dey, D.K. and Birmiwal, L.R. (1994) Robust Bayesian analysis using divergence measures. *Statistics & Probability Letters*, 20, 287–294.

Dias, L., Morton, A. and Quigley, J. (2017) *Elicitation: The Science and Art of Structuring Judgement*. Berlin: Springer.

Duan, Y., Ye, K. and Smith, E.P. (2006) Evaluating water quality using power priors to incorporate historical information. *Environmetrics*, 17, 95–106.

EMA (1998) *Statistical Principles for Clinical Trials*. European Medicine Agency: London UK. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. [Accessed 11 June 2020].

EMA (2006) *Guideline on clinical trials in small populations*. European Medicine Agency: London UK. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-clinical-trials-small-populations_en.pdf. [Accessed 11 June 2020].

Fraser, D.A.S. and Guttman, I. (1956) Tolerance regions. *Annals of Mathematical Statistics*, 27, 162–179.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. and Rubin, D. (2013) *Bayesian Data Analysis*, 3rd edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton, FL: CRC Press, Taylor & Francis.

Hampson, L.V., Whitehead, J., Eleftheriou, D. and Brogan, P. (2014) Bayesian methods for the design and interpretation of clinical trials in very rare diseases. *Statistics in Medicine*, 33, 4186–4201.

Hampson, L.V., Whitehead, J., Eleftheriou, D. and et al. (2015) Elicitation of expert prior opinion: application to the MYPAN trial in childhood polyarteritis nodosa. *PLOS ONE*, 10, 1–14.

Joseph, L. and Bélisle, P. (1997) Bayesian sample size determination for normal means and differences between normal means. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 209–226.

Joseph, L., Wolfson, D.B. and Berger, R.D. (1995) Sample size calculations for binomial proportions via highest posterior density intervals. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 44, 143–154.

Lindley, D.V. (1997) The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 129–138.

Neuenschwander, B., Weber, S., Schmidli, H. and O'Hagan, A. (2020) Predictively consistent prior effective sample sizes. *Biometrics*, 76, 578–587.

O'Hagan, A. and Forster, J.J. (2004) *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference*, 2nd edition. Kendall's Library of Statistics. London: Oxford University Press.

Spiegelhalter, D., Abrams, K. and Myles, J. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*. Statistics in Practice. New York: Wiley.

Weiss, R. (1997) Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46, 185–191.

Whitehead, J., Valdés-Márquez, E., Johnson, P. and Graham, G. (2008) Bayesian sample size for exploratory clinical trials incorporating historical data. *Statistics in Medicine*, 27, 2307–2327.

Zheng, H. and Hampson, L.V. (2020) A Bayesian decision-theoretic approach to incorporate preclinical information into phase I oncology trials. *Biometrical Journal*, 62, 1408–1427.

Zheng, H., Hampson, L.V. and Wandel, S. (2020) A robust Bayesian meta-analytic approach to incorporate animal data into phase I oncology trials. *Statistical Methods in Medical Research*, 29, 94–110.

Zheng, H. and Wason, J.M.S. (2022) Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. *Biostatistics*, 23, 120–135.

SUPPORTING INFORMATION

Web Appendices A–E referenced in Section 2, Figures S2–S4 in Sections 3 and 4, and Appendices H–L for additional simulations, extended application to time-to-event data, and a brief user-guide to apply the proposed methodology, are available at the Biometrics website on Wiley Online Library. Programming code for the sample size formulae and reproducing the numerical results, is posted online along with this paper, as well as available at GitHub: https://github.com/haiyanzheng/SSDcmspriors

How to cite this article: Zheng, H., Jaki, T., Wason, J.M.S. (2023) Bayesian sample size determination using commensurate priors to leverage preexperimental data. *Biometrics*, 79, 669–683. https://doi.org/10.1111/biom.13649