

# Bayesian sample size re-estimation using power priors

TB Brakenhoff, KCB Roes and S Nikolakopoulos

Statistical Methods in Medical Research  
2019, Vol. 28(6) 1664–1675

© The Author(s) 2018



Article reuse guidelines:

[sagepub.com/journals-permissions](http://sagepub.com/journals-permissions)

DOI: 10.1177/0962280218772315

[journals.sagepub.com/home/smm](http://journals.sagepub.com/home/smm)



## Abstract

The sample size of a randomized controlled trial is typically chosen in order for frequentist operational characteristics to be retained. For normally distributed outcomes, an assumption for the variance needs to be made which is usually based on limited prior information. Especially in the case of small populations, the prior information might consist of only one small pilot study. A Bayesian approach formalizes the aggregation of prior information on the variance with newly collected data. The uncertainty surrounding prior estimates can be appropriately modelled by means of prior distributions. Furthermore, within the Bayesian paradigm, quantities such as the probability of a conclusive trial are directly calculated. However, if the postulated prior is not in accordance with the true variance, such calculations are not trustworthy. In this work we adapt previously suggested methodology to facilitate sample size re-estimation. In addition, we suggest the employment of power priors in order for operational characteristics to be controlled.

## Keywords

Sample size, re-estimation, power prior, Bayesian, randomized controlled trial, monitoring, variance, borrowing

## 1 Introduction

A frequentist approach is typically employed for the design and analysis of a randomized controlled trial (RCT). The sample size is thus chosen in order for frequentist operational characteristics to be retained. This is done by specifying the power  $(1 - \beta)$  with which to detect a clinically relevant treatment effect ( $\delta^*$ ), given a type I error ( $\alpha$ ). For an RCT with two groups of equal size being compared on a normally distributed outcome with common unknown variance ( $\sigma^2$ ),  $\delta$  is commonly measured as the difference between the two groups' means. If we are interested in testing  $H_0: \delta = 0$  versus  $H_1: \delta > 0$  with  $\delta^* > 0$ , the sample size is determined by finding the first even integer solution to satisfy the following inequality

$$N \geq 4\sigma_0^2 \left( \frac{t_{N-2, 1-\alpha} + t_{N-2, 1-\beta}}{\delta^*} \right)^2 \quad (1)$$

where  $N$  represents the required sample size,  $t_{N-2, 1-\alpha}$  represents the  $(1 - \alpha)$  point of the  $t$ -distribution with  $N - 2$  degrees of freedom, and  $\sigma_0^2$  represents an initial assumption for the variance.

The assumption for the variance is usually based on limited prior information. Especially in the case of small or sensitive populations such as the ones defined by rare diseases or pediatric patients, the prior information might consist of only one small pilot study. Calculating the sample size can therefore be subject to considerable uncertainty.<sup>1</sup> Overestimation of the variance can result in committing to more resources than necessary, while underestimation can lead to inconclusive results. Both situations are undesirable when available research participants are limited.

A vast amount of methods have been developed to deal with such situations.<sup>2–5</sup> These methods have in common that they monitor the interim estimates of parameters within a trial, and respond to these estimates by recalculating the required sample size to meet the design characteristics. Methods that only monitor nuisance

Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands

### Corresponding author:

TB Brakenhoff, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, PO Box 85500, Utrecht 3508, GA, the Netherlands.

Email: [T.B.Brakenhoff-2@umcutrecht.nl](mailto:T.B.Brakenhoff-2@umcutrecht.nl)

parameters, such as the variance, are generally well accepted, but methods responding to interim estimates of the treatment effect can introduce bias.<sup>6</sup> However, in the frequentist framework, quantification of the uncertainty about the estimate of the variance remains an obstacle. The variability of this (interim) estimate is dependent on the amount of data collected and is substantial if only a small amount of subjects have been recruited.<sup>7</sup> In addition, if the variance is monitored only once, its estimator will be negatively biased by the end of the trial.<sup>1</sup> This is because an underestimation of the true variance at interim causes the required sample size to be re-estimated downwards. It is thus increasingly difficult to correct this erroneous estimate in the remaining sample size.<sup>2</sup> On the other hand, when the true variance is overestimated at interim, the sample size is re-estimated upwards, allowing enough time to adjust the estimate by the end of the trial. Friede and Miller<sup>1</sup> suggest continuous monitoring and re-estimation as a preferred solution for these issues. However, continually altering the original design based on an unstable estimate can come at great costs. Repeated sample size re-estimation (SSR) limits the amount of times the sample size can be re-estimated,<sup>1</sup> but still fails to clearly recommend when it is appropriate to do so. This is especially important when dealing with RCTs with a small available study population.

A Bayesian approach formalizes the aggregation of prior information on the variance with newly collected data, potentially alleviating some of the issues mentioned above.<sup>8</sup> Calculating the sample size necessary for a Bayesian RCT depends on the decision scheme that is to be followed after completion of the trial. Several different methods have been proposed, including hybrid frequentist-Bayesian,<sup>9–12</sup> fully decision theoretic<sup>13–16</sup> and interval-length based approaches.<sup>17–19</sup> Whitehead et al.<sup>8</sup> have advocated a variant of the latter which is comparable in simplicity to the frequentist sample size calculation (equation (1)) and includes an analogy to frequentist type I and II errors. This design requires the formulation of two hypotheses: (1)  $\delta > 0$ , indicating that the experimental group performs better than the control, and (2)  $\delta < \delta^*$ , concluding that the experimental treatment fails to improve upon control by a defined clinically relevant difference.<sup>8</sup> The sample size needs to be large enough to provide convincing evidence that either (1) or (2) is the case. Even though the same notation is used for  $\delta^*$  to point out the similarity in this approach and the standard frequentist one, the two effect sizes are not necessarily equivalent conceptually.

When the variance is unknown, the sample size is calculated using a belief about the variance in the form of a prior distribution. If this belief is in agreement with the actual data-generating mechanism, the calculated sample size ensures that the design characteristics will be fulfilled by the end of the trial. If this is not the case, recruiting the original sample size might not be enough to satisfy either of the hypotheses, leading to an inconclusive trial. Just as in the frequentist context, monitoring the variance during the trial can facilitate interim SSR. Several approaches have been proposed using external information for sample size adjustment in a Bayesian framework.<sup>20–23</sup> In particular, Zhong et al.<sup>24</sup> discuss SSR for RCTs with a binary outcome, but a similar approach has not been considered for RCTs with a continuous outcome.

Moreover, when there is conflict between the prior and the data, Bayesian procedures can have unpredictable, and likely undesirable, frequentist characteristics. The *power prior* approach<sup>25</sup> can be employed in order to discard the influence of prior information on posterior inference; this is achieved by employing a power parameter  $\gamma \in [0, 1]$  which usually translates as a deflating factor of the precision of the prior distribution. The application of power priors in sample size determination has been considered before.<sup>26</sup> The most challenging aspect in employing power priors is the specification of  $\gamma$ . Adaptive formulations of the power prior<sup>27–29</sup> allow for  $\gamma$  to be estimated based on the similarity between the prior and current data and thus, with appropriate calibration, achieve desired operational characteristics.

In light of the above, the goal of the present research is to explore the operational characteristics of the sample size determination method proposed in Whitehead et al.<sup>8</sup> in the case of misspecified variance and to demonstrate the effects of SSR by interim variance monitoring. We employ the power prior approach introduced in Nikolakopoulos et al.<sup>29</sup> to synthesize prior and new data in order for operational characteristics (in this case the probability of having a conclusive trial) to be calibrated.

The paper is organized as follows: in the following section, the sample size determination procedure described in Whitehead et al.<sup>8</sup> is outlined and adapted to allow for SSR. The adaptive power prior, based on predictive distributions and termed *Prior-Data conflict calibrating power prior (PDCCPP)* in Nikolakopoulos et al.,<sup>29</sup> is then briefly described and applied in the variance re-estimation problem. Subsequently, the proposed approach is demonstrated for a clinical trial in the field of pediatrics. The paper ends with a discussion.

## 2 Bayesian sample size determination

We consider the case where an RCT is designed to evaluate an experimental treatment (E) against a control (C) on a normally distributed outcome ( $Y_j \sim N(\mu_j, \sigma^2)$ , where  $j = E, C$ ) with unknown variance ( $\sigma^2$ ).  $Y_{ji}$  is the outcome

value of subject  $i$  in group  $j$ . In the Bayesian framework the precision ( $\nu = 1/\sigma^2$ ) is often used for modeling purposes. The required sample size is denoted by  $N$ , where  $N = N_E + N_C$ . A positive value for  $\delta = \mu_E - \mu_C$  indicates that E is better than C. A gamma prior distribution is assumed for  $\nu$  with parameters  $\alpha_0$  and  $\beta_0$ . This corresponds to  $2\alpha_0$  observations with a sample precision of  $\alpha_0/\beta_0$ . The conditional prior for  $\mu_j$ , given  $\nu$ , is normal with mean  $\mu_{0j}$  and precision  $q_{0j}\nu$ . This information corresponds to  $q_{0j}$  virtual patients with an average of  $\mu_{0j}$  on the outcome variable.<sup>30</sup> The posterior of  $\mu_j$ , given  $\nu$ , is also normal with mean  $\mu_{1j}$  and precision  $q_{1j}\nu$ , where  $\mu_{1j} = \mu_{0j}(q_{0j}/q_{1j}) + N_j\bar{y}_j/q_{1j}$  and  $q_{1j} = q_{0j} + N_j$ .  $\nu$  has a gamma posterior distribution with  $\alpha_1 = \alpha_0 + \frac{N}{2}$  and  $\beta_1 = \beta_0 + \frac{H}{2}$  where

$$H = \sum_{i=1}^{N_E} (y_{Ei} - \bar{y}_E)^2 + \sum_{i=1}^{N_C} (y_{Ci} - \bar{y}_C)^2 + \frac{N_C q_{0C} (\bar{y}_C - \mu_{0C})^2}{q_{0C} + N_C} + \frac{N_E q_{0E} (\bar{y}_E - \mu_{0E})^2}{q_{0E} + N_E}$$

The posterior of  $\delta|\nu$  is then normal with mean  $\delta_1 = \mu_{1E} - \mu_{1C}$  and precision  $D\nu$ , with  $D = (q_{1E}q_{1C})/(q_{1E} + q_{1C})$ .

The sample size should be large enough to either provide convincing posterior evidence that E is better than C (a successful result), or that E is not better than C by some clinically relevant treatment effect ( $\delta^*$ ) (a futile result), as shown by the following criteria

$$\begin{aligned} Pr(\delta > 0|\mathbf{y}) &\geq \eta && \text{Success criterion} \\ Pr(\delta < \delta^*|\mathbf{y}) &\geq \zeta && \text{Futility criterion} \end{aligned}$$

where  $\eta$  and  $\zeta$  are probability thresholds for the success and futility criteria, respectively. As shown in Whitehead et al.,<sup>8</sup> the occurrence of at least one of these alternatives is guaranteed if

$$\frac{D\alpha_1}{\beta_1} \geq \left( \frac{t_{2\alpha_1, \zeta} + t_{2\alpha_1, \eta}}{\delta^*} \right)^2 \quad (2)$$

However,  $\beta_1$  is dependent on the data and therefore a random variable. Thus, equation (2) is required to be true with high probability ( $\xi$ ).

Furthermore, given  $\nu$ ,  $W = \nu H$  has a chi-squared distribution with  $N$  degrees of freedom. If  $\nu$  has a prior gamma distribution with parameters  $\alpha_0$  and  $\beta_0$ ,  $J_0 = 2\beta_0\nu$  also has a prior gamma distribution with parameters  $\alpha_0$  and  $\frac{1}{2}$ . If  $M$  (referred to as  $F$  in Whitehead et al.<sup>8</sup>) is defined as

$$M = \frac{W/N}{J_0/2\alpha_0} = \frac{H\alpha_0}{N\beta_0} \quad (3)$$

then the prior predictive distribution of  $M$  is an F-distribution with  $N$  and  $2\alpha_0$  degrees of freedom. Making use of the relationship between the F-distribution and the Beta distribution, it is shown that equation (2) will be satisfied with probability at least  $\xi$  if

$$D \frac{\alpha_1}{\beta_0} \left( 1 - \text{Beta}_{\frac{N}{2}, \alpha_0, \xi} \right) \geq \left( \frac{t_{2\alpha_1, \zeta} + t_{2\alpha_1, \eta}}{\delta^*} \right)^2 \quad (4)$$

where  $\text{Beta}_{a,b,\xi}$  denotes the  $\xi$  point of a beta distribution function with parameters  $a$  and  $b$ . Using a search procedure, the smallest even sample size that satisfies equation (4) can be determined.

## 2.1 Sample size re-estimation

To facilitate interim SSR, the design described in the above section can be adapted. The required sample size ( $N$ ) is now gathered in  $K$  stages. Let  $n_{(k)j}$  represent the sample size recruited in group  $j$  (where  $j = E, C$  and  $n_k = \sum_j n_{(k)j}$ ) in the  $k$ th stage, with  $k = 1, \dots, K$ . Equal allocation is assumed and interims are not required to be equally spaced.

At each interim, distributions of the precision ( $\nu$ ) and the means ( $\mu_j$ ) are updated with the collected data. The prior value of a parameter at the  $k$ th interim will now be referred to with subscript  $k-1$ . Consequently, the subscript for the posterior, updated value can be denoted by  $k$ . This value is equal to the prior for the  $k+1$ th interim. Note that  $k=0$  corresponds to the design phase and subscript  $K$  refers to the posterior value of the parameter if all the required sample size is recruited.

The posterior of  $\mu_j | \nu$  has parameters  $\mu_{(k)j} = \mu_{(k-1)j}(q_{(k-1)j}/q_{(k)j}) + n_{(k)j}\bar{\mathbf{y}}_{(k)j}/q_{(k)j}$  and  $q_{(k)j} = q_{(k-1)j} + n_{(k)j}$ , where  $\bar{\mathbf{y}}_{(k)j}$  is the mean of the data collected in group  $j$  in the  $k$ th stage. The gamma posterior of  $\nu$  has parameters  $\alpha_k = \alpha_{k-1} + n_k/2$  and  $\beta_k = \beta_{k-1} + H_k/2$  where

$$H_k = \sum_{i=1}^{n_{(k)E}} (\mathbf{y}_{(k)Ei} - \bar{\mathbf{y}}_{(k)E})^2 + \sum_{i=1}^{n_{(k)C}} (\mathbf{y}_{(k)Ci} - \bar{\mathbf{y}}_{(k)C})^2 \\ + \frac{n_{(k)C}q_{(k-1)C}(\bar{\mathbf{y}}_{(k)C} - \mu_{(k-1)C})^2}{q_{(k-1)C} + n_{(k)C}} + \frac{n_{(k)E}q_{(k-1)E}(\bar{\mathbf{y}}_{(k)E} - \mu_{(k-1)E})^2}{q_{(k-1)E} + n_{(k)E}}$$

The posterior of  $\delta | \nu$  has mean  $\delta_k = \mu_{(k)E} - \mu_{(k)C}$  and precision  $D_k \nu$ , with  $D_k = (q_{(k)E}q_{(k)C})/(q_{(k)E} + q_{(k)C})$ .

As the trial progresses, the relative influence of the trial data increases, and that of the initial prior belief decreases. This reflects the inherent updating nature of the Bayesian methodology. At interim  $k$ , the additional sample size required to obtain the design characteristics ( $N_k$ ) now depends on the last posterior value of  $\alpha$  and  $D$

$$D_K \frac{\alpha_K}{\beta_K} \left( 1 - \text{Beta}_{\frac{N_k}{2}, \alpha_K, \xi} \right) \geq \left( \frac{t_{2\alpha_K, \zeta} + t_{2\alpha_K, \eta}}{\delta^*} \right)^2 \quad (5)$$

where  $D_K = (q_{(K)E}q_{(K)C})/(q_{(K)E} + q_{(K)C})$  and  $\alpha_K = \alpha_k + N_k/2$ . The total required sample size (as estimated at stage  $k$ , including those already measured) is equal to  $N_k + \sum_{p=1}^k n_p$ .

## 2.2 $\xi$ Calculation

As mentioned earlier,  $\xi$  represents the probability of a conclusive decision by the end of the trial. By solving equations (4) and (5) for  $\xi$ , we can calculate this probability, given the prior, the data so far, and the remaining sample size.

When equation (4) is solved for  $\xi$ , we obtain

$$\xi = F \left[ 1 - \left( \frac{t_{2\alpha_1, \zeta} + t_{2\alpha_1, \eta}}{\delta^*} \right)^2 \frac{\beta_0}{\alpha_1 D}; \frac{N}{2}, \alpha_0 \right]$$

where  $F(x; a, b)$  is the c.d.f. of a Beta distribution with parameters  $a$  and  $b$ .

By applying the same steps to equation (5), we can also find an expression for  $\xi_k$  in a setting with multiple interims

$$\xi_k = F \left[ 1 - \left( \frac{t_{2\alpha_K, \zeta} + t_{2\alpha_K, \eta}}{\delta^*} \right)^2 \frac{\beta_k}{\alpha_K D_K}; \frac{N_k}{2}, \alpha_k \right] \quad (6)$$

In the case of limited available sample units, if the required sample size cannot be recruited,  $\xi_k$  can be used to evaluate the consequences of continuing the trial with the maximum available subjects. Moreover, the benefit of putting in the extra effort to recruit more subjects can now be quantified. In the following section, the operational characteristics of  $\xi$  are evaluated and the impact of a misspecified prior for the variance is assessed and shown to be substantial. The application of *PDCCPP*'s is demonstrated, as well as how they can be used as a remedy.

The importance of SSR for such a Bayesian approach is stressed. In addition to the reasons sketched for the frequentist case (i.e. the variance being poorly described by the prior distribution due to systematic differences in the two populations), the uncertainty about the variance is now, unlike in the frequentist case, directly incorporated in the sample size calculation. This results in larger sample sizes required for similar decision thresholds (as can be seen when comparing equation 4 with equation 1 for  $\sigma_0^2 = \beta_1/\alpha_1$ ). By incorporating interim data in the variance estimation, this uncertainty is reduced resulting in a re-estimated sample size smaller than the initial one, also when the expected value for the precision is the same.

This is shown in Table 1. For the situation where  $\delta^* = 0.6$ ,  $\eta = 1 - \alpha = 0.95$ ,  $\zeta = 1 - \beta = 0.8$ ,  $\xi = 0.9$  and only a prior on the variance is used (thus the priors for the groups' means are non-informative), the sample sizes calculated are compared with the frequentist one. In the frequentist case, a  $\sigma_0^2 = 1$  is assumed while in the Bayesian case a gamma prior distribution around  $\nu$  is used with  $E(\nu) = 1$ ,  $\alpha_0 = 5$ , and  $\beta_0 = 5$ . It is shown how SSR in a Bayesian RCT of this kind can decrease the, initially considerably larger, sample size if the variance is as

**Table 1.** Sample size required for frequentist and Bayesian procedures for  $\delta^* = 0.6$ ,  $\eta = 1 - \alpha = 0.95$ ,  $\zeta = 1 - \beta = 0.8$ ,  $\xi = 0.9$ .

$N_I$	$N_{freq}$	$N_{Bayes}$
0		140
10		108
20		96
30	72	90
40		86
50		82
100		80
500		72

Note:  $N_I$  denotes the sample size at which the interim analysis takes place (or the size of the prior if  $N_I > N_{Bayes}$  – see text for details) while the mean of  $\nu|N_I = \sigma_0 = 1$ . The initial prior (a gamma distribution with  $\alpha_0 = 5$  and  $\beta_0 = 5$ ), corresponding to  $N_I = 0$ , is based on a historical dataset of 10 patients. The priors for the groups' means are taken to be non informative ( $q_{0E} = q_{0C} = 0$ ).

expected. Here  $N_I$  denotes the sample size at which the interim takes place, but it is equivalent to the situation where the prior for  $\nu$  is based on  $N_I$ . As such  $N_I$  may be larger than  $N_{Bayes}$ .

Here we mention that comparison of Bayesian and frequentist sample sizes is by no means straightforward. Nevertheless the mathematical resemblance of equation (1) with equation (2) allows us to make such a comparison and note that the frequentist paradigm is similar to the Bayesian approach described here, if it were to assume that the mean of the posterior of  $\nu$  is known and equal to  $\alpha_1/\beta_1$ .

### 3 Frequentist properties of Bayesian SSR

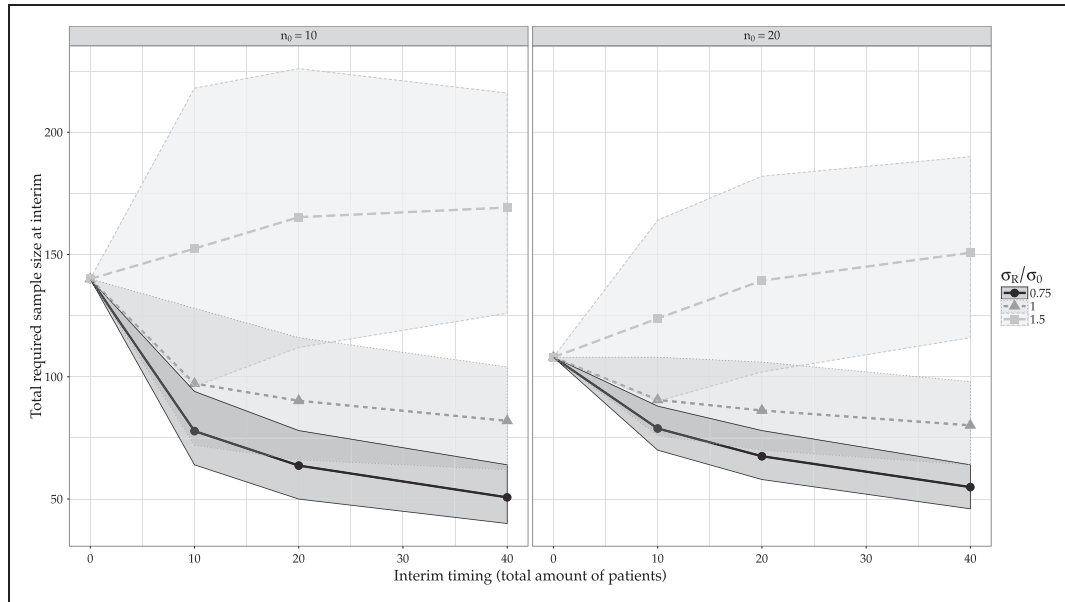
#### 3.1 Variance misspecification

From now on, even though modelling takes place in terms of precision ( $\nu$ ), we describe dispersion by the standard deviation  $\sigma$  for purposes of standardization and clarity. As shown in the previous section, Bayesian SSR can aid to mitigate the inflation on the initial sample size calculation imposed by modeling the uncertainty about  $\sigma$ . However, if the true  $\sigma$ ,  $\sigma_R$  is different than the one observed in the historical data ( $\sigma_0 = \sqrt{\frac{\beta_0}{\alpha_0}}$ ), the Bayesian procedure can have unpredictable operational characteristics.

For our illustrative case, when design parameters are as introduced in the previous section, Figure 1 shows the sample size estimated when the mean of the prior is  $E(\nu) = 1$ , so  $E(\sigma) \approx 1$ , but  $\sigma_R$  is not as expected, for different sizes of the prior and interim location. Especially when the true variance is larger than expected, the prior distribution can cause considerable discrepancies in the estimated sample size, even when SSR is employed.

These issues become more apparent when the frequentist properties of  $\xi$  are studied. The top two panes of Figure 2 shows the empirical  $\xi$  ( $\xi_{emp}$ ), that is the empirical probability of equation (2) being satisfied (calculated using equation (6)), as a function of the ratio of the (re-)estimated sample size (i.e. the collected sample size divided by the sample size (re-)estimated at interim that is required to obtain the design characteristics ( $\xi_{emp} = 0.9$ )). Plotted for different interim sizes and  $\sigma_R$ , such a metric explores how reliably equation (6) can estimate the frequentist probability of making a decision with the sample size estimated by the Bayesian approach. It is evident that such calculations are significantly unstable and heavily influenced by both the location and scale of the prior distribution.

The problem is only partially remedied by re-estimation and/or increasing the interim size and even then, when the true variance is larger than expected by the prior,  $\xi_{emp}$  is deviating considerably from its 90% assumed value for the sample size (re-)estimated. When  $\sigma_R = \sigma_0$  we see that equation (2) is true roughly 90% of the times at the sample size re-estimated, in accordance with the design requirements. This holds irrespective of the size of the prior distribution and the interim look location. But when  $\sigma_R = 1.5\sigma_0$ , the sample size required to make a decision with probability as by design ( $\xi = 90\%$ ) can be considerably larger than the one (re-)estimated.



**Figure 1.** Sample sizes estimated with Bayesian SSR, with their 95% confidence intervals, for different true  $\sigma$ 's ( $\sigma_R$ ), for assumed  $\sigma = 1$  and  $\delta^* = 0.6$ ,  $\eta = 1 - \alpha = 0.95$ ,  $\zeta = 1 - \beta = 0.8$ ,  $\xi = 0.9$ ,  $q_{0E} = q_{0C} = 0$ . The prior distribution is based on either 10 (left) or 20 (right) patients.

Clearly, deviations from assumptions imposed by the prior distribution can cause calculations which are very relevant for the planning of such a Bayesian RCT, to be untrustworthy. A remedy is suggested by using adaptive power priors which calibrate the prior distribution in light of the new data, thus circumventing the problem.

#### 4 Prior data conflict calibrated power priors

If the data of a current study is denoted by  $D_1$  with respective likelihood function  $L(\theta|D_1)$  where  $\theta$  is a vector of parameters and  $D_0$  denotes the data from a similar historical study, with  $L(\theta|D_0)$  the corresponding likelihood, the basic definition of the power prior as described in Ibrahim and Chen<sup>25</sup> is

$$\pi(\theta|D_0, \gamma) \propto L(\theta|D_0)^\gamma \pi_0(\theta)$$

where  $\pi_0(\theta)$  is the initial prior before the historical data are observed, usually assumed flat. Using this formulation, the posterior of  $\theta$  after observing  $D_1$  is then

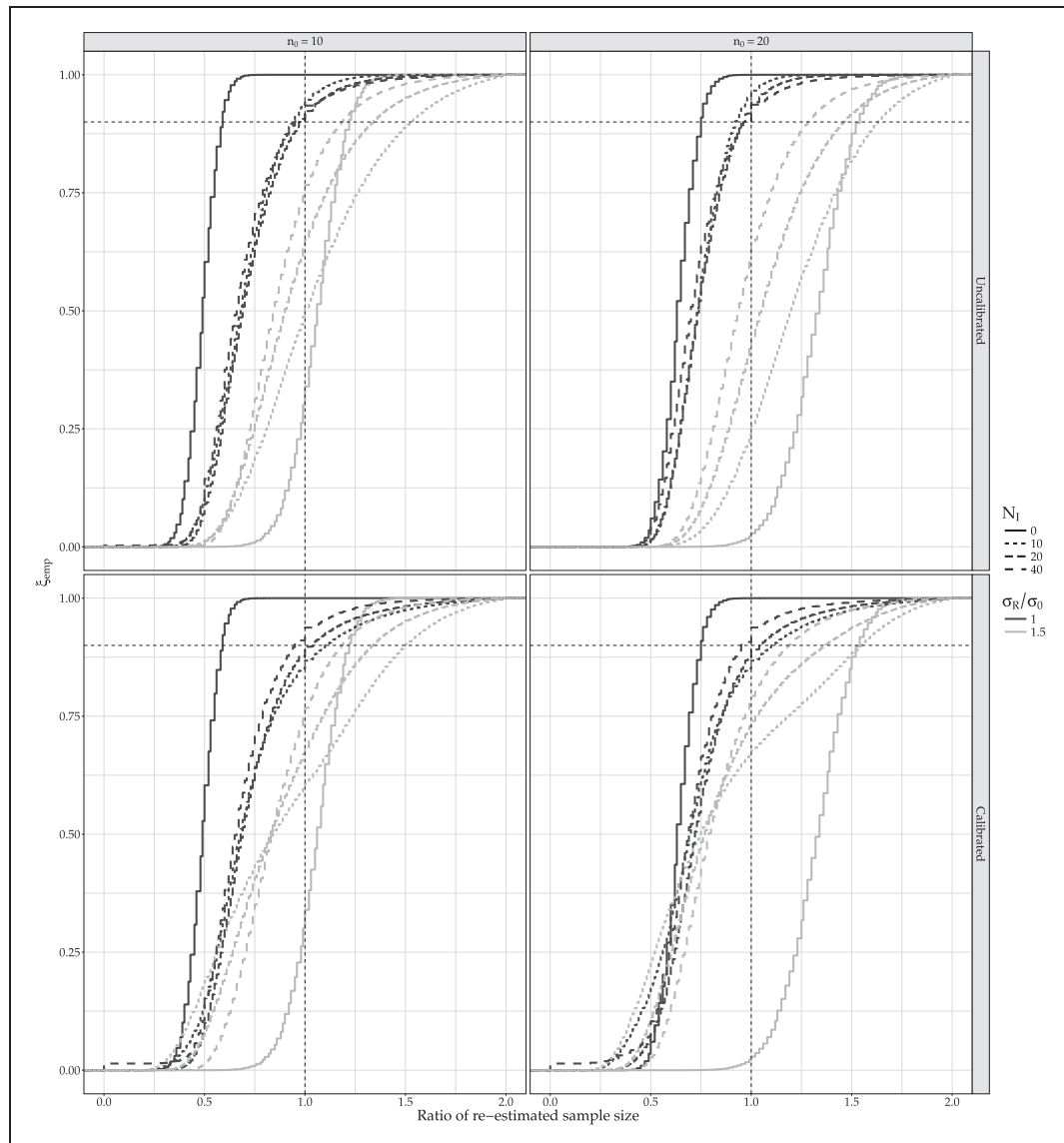
$$\pi(\theta|D_1, D_0, \gamma) \propto L(\theta|D_1)L(\theta|D_0)^\gamma \pi_0(\theta) \quad (7)$$

The  $\gamma$  parameter,  $\in [0, 1]$ , plays the role of a discounting factor, translating to the proportion of the sample size of the historical study at which the prior is finally based. Several extensions have been discussed in the literature and we refer the interested reader to Ibrahim et al.<sup>31</sup> Here we employ the one suggested by Nikolakopoulos et al.,<sup>29</sup> for its simplicity and adaptive nature. An additional attractive feature of *PDCCPP*'s is that in conjugate models the posterior in equation (7) is still tractable as  $\gamma$  is replaced by  $\hat{\gamma}$ . The approach can be described as follows: if  $T$  is a (sufficient) statistic for  $\theta$  and  $[l_{1-c/2}^{pr}, l_{c/2}^{pr}]$  is a  $100(1-c)\%$  credible interval from the prior predictive distribution for  $\theta$  then

$$\hat{\gamma}_{PDCCPP}(c) = \min \left[ \max_{\hat{\gamma}} \left( \left\{ \hat{\gamma} : T^{obs} \in (l_{c/2}^{pr}, l_{1-c/2}^{pr}) | \hat{\gamma} \right\}, 1 \right) \right] \quad (8)$$

Thus, the prior is calibrated in such a way so that the  $100(1-c)\%$  predictive credible interval for  $T$  includes the observed value  $T^{obs}$ . Or, in other words the  $\pi(\theta|D_0, \hat{\gamma})$  is such that the two-sided prior predictive  $p$ -value for  $T$  is at least  $c$ . By choosing  $c$ , as shown by Nikolakopoulos et al.,<sup>29</sup> one can calibrate the procedure in order for desirable





**Figure 2.** Empirical probabilities of making a decision ( $\xi_{emp}$ ) when sample size is (re-)estimated without (top) or with (bottom) calibration using PDSCPP's for different ratios of true  $\sigma$  ( $\sigma_R$ ) over assumed at design stage ( $\sigma_0$ ), for assumed  $\sigma = 1$  and  $\delta^* = 0.6$ ,  $\eta = 1 - \alpha = 0.95$ ,  $\zeta = 1 - \beta = 0.8$ ,  $\xi = 0.9$ ,  $q_{0E} = q_{0C} = 0$ . The prior distribution is based on either 10 (left) or 20 (right) patients.

frequentist characteristics to be achieved. Since the above formulation is the only one discussed here, we use the terms  $\hat{\gamma}_{PDCCPP}$  and  $\hat{\gamma}$  interchangeably.

#### 4.1 Application of PDCCPP in Bayesian SSR

In order to apply the PDCCPP methodology in the Bayesian SSR problem, we use the predictive distribution for  $M$  (see equation (3)). It can be shown that, if the initial priors before the historical study are assumed flat and only information for the variance is used at the design stage, such an empirical power prior formulation is equivalent to using a prior  $\text{Gamma}(\hat{\gamma}\alpha_0, \hat{\gamma}\beta_0)$  for  $\nu$  and hence  $\text{Var}(\nu|\hat{\gamma}) = \frac{1}{\hat{\gamma}} \text{Var}(\nu|\gamma = 1)$  and  $E(\nu|\hat{\gamma}) = E(\nu) = E(\nu|\gamma = 1)$  where the prior  $\pi(\nu|\gamma = 1)$  is equivalent to full borrowing of the prior data and not implementing PDCCPP.

Analytical derivation of  $\hat{\gamma}_{PDCCPP}$  is not straightforward due to the complex form of the cumulative distribution and quantile functions of the  $F$  distribution. Nevertheless, estimation of the power parameter by simulation is an easy task. After  $c$  is chosen, a simple search procedure with reasonable precision can be employed in order to find  $\hat{\gamma}_{PDCCPP}$ . If  $M^{obs} > I_{1-c/2}^{F_{pr}}$ , where  $I_{1-c/2}^{F_{pr}}$  is the  $1 - c/2$  quantile of the  $F_{pr}(N_1, 2\alpha_0)$

predictive distribution for  $M$  when  $N_1$  patients' responses are observed (such that  $Pr_{F_{pr}}(M < l_{1-c/2}^{F_{pr}}) = 1 - c/2$ ),  $\hat{\gamma}$  will be such that  $F_{pr}$  will be wide enough (by decreasing the second degrees of freedom parameter to  $2\hat{\gamma}\alpha_0$ ) so that  $M^{obs} = l_{1-c/2}^{F_{pr}}$ . The counterpart adjustment takes place when  $M^{obs} < l_{c/2}^{F_{pr}}$ . Note that the latter (adjusting  $F_{pr}$  so that  $M^{obs} = l_{c/2}^{F_{pr}}$  when  $M^{obs} < l_{c/2}^{F_{pr}}$ ) might not be possible if  $M^{obs}$  is close to 0. In these cases, in our simulations, we set  $\hat{\gamma}_{PDCCPP} = \{\hat{\gamma} : \hat{\gamma}\alpha_0 = 1\}$ . If  $l_{c/2}^{F_{pr}} \leq M \leq l_{1-c/2}^{F_{pr}}$ ,  $\hat{\gamma} = 1$ .

The choice of  $c$  should be such that the frequentist characteristics of interest are controlled. In this case, a predefined probability of making a decision with the estimated sample size is the key characteristic to satisfy. Note that the larger  $c$  is, the narrower the credible interval in equation (8) and consequently, the less probable to use the prior in full. In Nikolakopoulos et al.,<sup>29</sup> it is discussed how  $c$  has to be larger, the smaller  $N_1$  is relative to the historical sample size ( $2\alpha_0$  in this case), for a procedure based on *PDCCPP* to preserve the same operational characteristics. Thus, all else being equal, for larger historical sample sizes, larger  $c$ 's should be employed in order for the same operational characteristics to be met. We make this choice heuristically here, based on this principle, and elaborate further in section 5.

Figure 2 shows how the operational characteristics of Bayesian SSR turn out to be when *PDCCPP* are employed. The sample sizes (re-)estimated are now less sensitive to the prior distribution. Calculation of  $\xi$  is also more robust as the lines move closer together, with a higher  $\xi$  reached in most cases when all of the re-estimated sample size is collected. For example, for the case of  $N_1 = 40$ ,  $n_0 = 20$ , and  $\sigma_R/\sigma_0 = 1.5$ , empirical  $\xi$  goes from 59% without calibration to approximately 80% when calibrated through *PDCCPP*. Here  $1 - c/2$  was set to 0.2, 0.4, 0.6 and 0.8 for a ratio of the interim location over the prior of 0.5, 1, 2 and 4, respectively, a heuristic choice as discussed above. In general,  $c$  must be such that the empirical  $\xi$  does not depend heavily on the true value of the variance and is relatively close to the intended value (here 90%). The implemented value of  $c$  will depend on the location and precision of the prior, the true value of the variance and the required robustness of  $\xi$ . Since these dependencies are not straightforward to quantify, a simulation-based approach must be implemented. Currently, simulation-based choices for parameters are increasingly applied when designing clinical trials. In the following section, by means of an example, we show how such choices can be more refined.

SSR with or without *PDCCPP* did not affect operational characteristics such as the probability of showing efficacy or futility (see supplementary material). R code used for the procedure and simulations described in this section as well as the analyses presented in section 5 can be found at <https://github.com/timobrakenhoff/BayesianSSRwithPDCCPP>.

## 5 Example

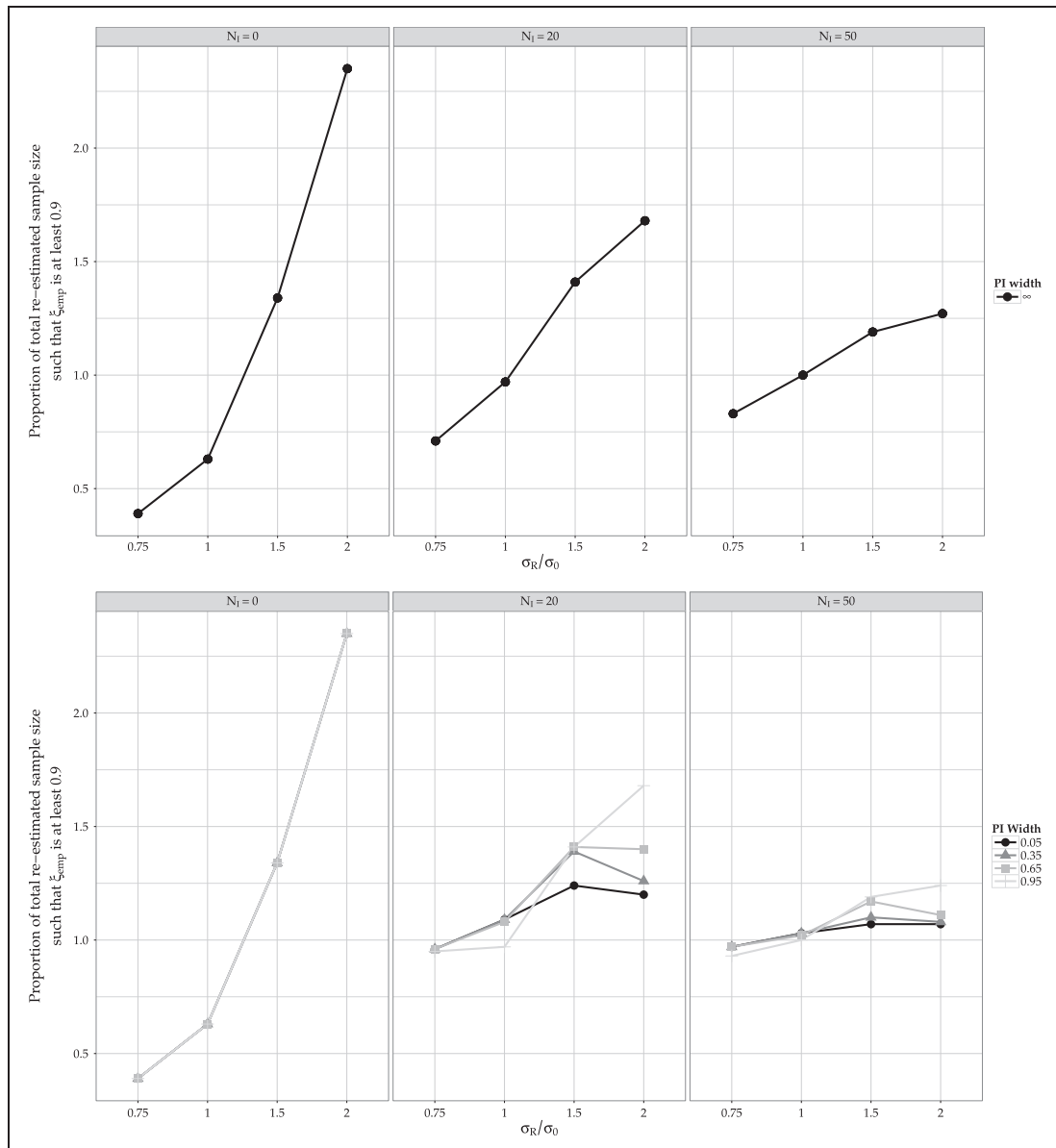
The following example is from a multicentre, double-blind, prospective, randomized, placebo-controlled trial that evaluated the efficacy of dexamethasone in very young patients mechanically ventilated for lower respiratory infection caused by respiratory syncytial virus (RSV-LRTI).<sup>32</sup> Eighty-five children younger than 24 months on mechanical ventilation were randomized to receive either dexamethasone (E) or placebo (C). The primary outcome measure was the duration of mechanical ventilation in days which was assumed normally distributed in the original trial.

Even though no adequate treatment has yet been identified for severe RSV-LRTI, a previous RCT by van Woensel et al.<sup>33</sup> found a potential beneficial effect of corticosteroids. Treatment with prednisolone as compared to placebo reduced the duration of mechanical ventilation in a small subgroup of patients on mechanical ventilation by 1.6 days. This result was based on seven patients in the prednisolone group and seven patients in the placebo group, for which the estimate for the standard deviation was  $\sigma_0 = 4.23$ .

In order to illustrate the design approach suggested by combining SSR and *PDCCPP*'s, we discuss the situation where the RCT at hand was to be designed in the Bayesian manner described in Whitehead et al.,<sup>8</sup> when only prior information on the variance were to be used. Thus, by assuming  $\alpha_0 = 7$ ,  $\beta_0 = 125.3$ ,  $\delta^* = 1.5$ ,  $\eta = 0.95$ ,  $\zeta = 0.8$  and  $\xi = 0.9$ , a sample size of 352 is deemed necessary for a Bayesian RCT that will declare efficacy based on posterior probabilities. This sample size is considerably larger than the 198 patients required by the frequentist approach since the uncertainty about the variance is also modelled. However, as discussed, SSR can reduce the sample size required (if the variance is as assumed). But, if the variance is not as expected by the prior, calculations are not to be trusted. Thus *PDCCPP* is employed as a remedy.

Figures 3 and 4 show the type of exploration that could be of help in deciding the value of  $c$  and the interim timing. The ratio of the (re-)estimated sample size to have a  $\xi = 90\%$  probability of making a decision as dictated by design, is explored. This ratio can alternatively be expressed as the total number of subjects *actually* required to reach  $\xi = 90\%$  divided by the total (re-)estimated number of subjects that were *expected* to be required at interim.

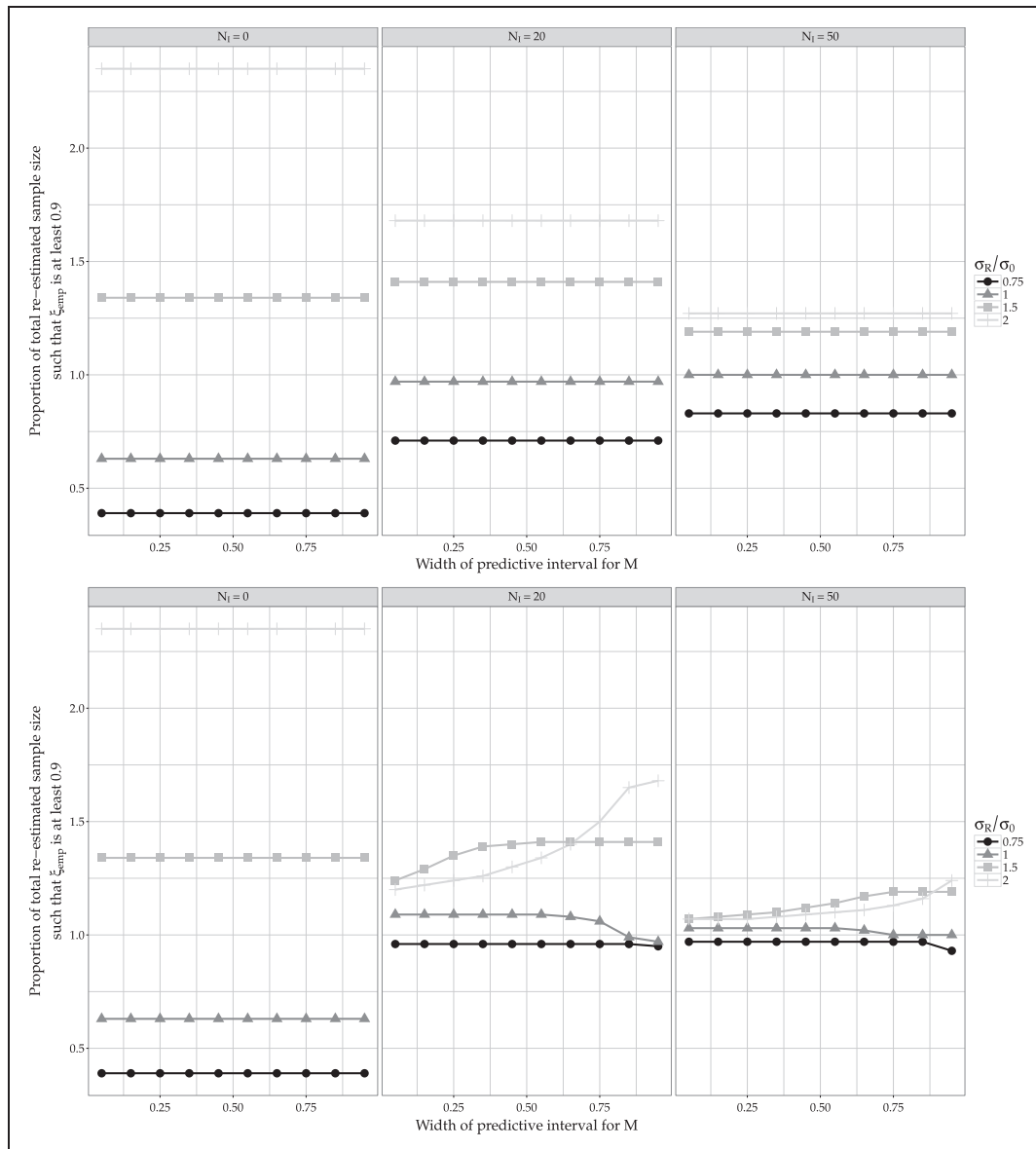




**Figure 3.** Ratio of (re-)estimated sample size required to reach  $\xi = 0.9$  without (upper) and with (lower) use of PDCCPP. Shown as a function of the true variance for different widths of the prior predictive intervals  $(1 - c/2)$  and different sample sizes at which the (re-)estimation takes place  $N_I = \{0, 20, 50\}$ .

In Figure 3 it is plotted as a function of the ratio of  $\sigma_R/\sigma_0$ , for different widths of the predictive distribution's credible interval (PI) and sample sizes at which the interim analysis takes place. In Figure 4 it is plotted as a function of the width of the PI, for different  $\sigma_R/\sigma_0$  and interim sample sizes.

In both figures it is shown how these factors affect the frequentist performance of  $\xi$ . The wider the PI is, the less the weight of the prior is adapted. This leads to more biased calculation of  $\xi$  (sample size required for 90% probability of making a decision is considerably different). The narrower the PI is, the less the frequentist performance of  $\xi$  depends on the true value of the variance (lines in both graphs are closer to each other). Note that at the lower middle plane of Figure 4, when  $\sigma_R = 2\sigma_0$ , smaller PI widths lead to better  $\xi$  estimation than when  $\sigma_R = 1.5\sigma_0$ . This happens because the discrepancy between the prior and  $\sigma_R$  (when  $\sigma_R = 2\sigma_0$ ) is such that there is small overlap between the sampling and predictive distributions, leading to higher chances of considerably down-weighting the prior. When the width of the PI becomes large enough, the effect of the large discrepancy takes over, leading to more biased estimation of  $\xi$  than when  $\sigma_R = 1.5\sigma_0$ .



**Figure 4.** Ratio of (re-)estimated sample size required to reach  $\xi = 0.9$  without (upper) and with (lower) use of PDCCPP. Shown as a function of the width of the prior predictive interval ( $l - c/2$ ), for different values of the true variance and different sample sizes at which the (re-)estimation takes place  $N_I = \{0, 20, 50\}$ .

Such explorations could lead to a choice of PI width and interim timing should result in a required balance between variability in sample size calculation and robustness of the calculation of  $\xi$ . This can be done alongside considerations about the maximum sample size available in the case of an RCT in small populations.

## 6 Discussion

In this paper, the sample size determination procedure described in Whitehead et al.<sup>8</sup> has been adapted to facilitate interim SSR based on the variance of the observed data. Furthermore, the frequentist properties of such a procedure were shown to be heavily dependent on the prior-data disagreement. A power prior method that calibrates the prior in case of conflict is suggested as a solution. It is also discussed how the interplay between the desired similarity and the ratio of the sample sizes of the prior and new data affect those frequentist properties.

As frequentist properties we considered the probabilities of making a decision within the Bayesian decision scheme suggested in Whitehead et al.<sup>8</sup> Robustness of calculations can be of importance in research conducted in

small populations where recruitment difficulties can result to very long clinical trials. Feasibility of such an ongoing trial is of interest.

We did not dwell into probabilities of correct or wrong decisions (the analogues of type I and II errors). However, as shown in the supplementary material, such probabilities were only marginally affected by the SSR suggested here. This does not come as a surprise since our method is monitoring only the variance and not the treatment effect. Methods that facilitate SSR by monitoring the variance are generally more well accepted by regulators.<sup>6</sup> It should be noted that the current method requires unblinding, at least of the statistician.

We also encourage further research on the performance of this method when multiple interim looks are taken sequentially and the allocation of sample size is not 1:1. While the sample size re-estimation method proposed here is explored for multiple interims, we would advise the detailed attention of a statistician at the end of each interim (which can be considered best practice). If this is not possible, we propose to perform no more than 1 interim when interested in performing Bayesian sample size re-estimation with *PDCCPP*. Limited increase in efficiency, risk of unintended bias, and other logistical concerns may also be reasons to restrict the number of interims. In addition, the allocation of sample size in this paper is assumed to be 1:1. However, as the original Bayesian sample size estimation approach proposed by Whitehead et al.<sup>8</sup> does not require equal allocation, this assumption can likely be relaxed in the re-estimation approach proposed here.

The Bayesian method explored here results in considerably larger sample sizes than the frequentist ones for seemingly similar decision criteria. We show how SSR can partly remedy this. Assuming that the uncertainty in any variance estimate (prior distribution or fixed assumption) is acknowledged, and SSR is part of the design of a new trial, the Bayesian and frequentist approaches suggest two different strategies: Let us consider the case where a two-stage SSR is planned (thus one interim analysis for SSR) for a trial with very limited prior information, for example one small pilot study. A frequentist would start small, bearing considerable uncertainty concerning the sample size estimate of the second stage. The amount of this uncertainty depends on the sample size of the pilot study. Furthermore this uncertainty is not incorporated in the assumed value for the variance thus the sample sizes calculated might be deemed as unrealistic in small populations RCTs.

A Bayesian would start large, thus being prepared in terms of commitment of resources, and then reduce the sample size if the true variance was indeed equal to the point estimate of the pilot study – the same estimate the frequentist would use. The *PDCCPP* approach is fairly robust against variance misspecification; a robustness all more important in RCTs in populations where repetition of a trial that was subject to misspecification is rather unlikely. The Bayesian would also have a different decision scheme as posterior inference as described in Whitehead et al.<sup>8</sup> could also conclude futility whereas “acceptance of  $H_0$ ” is not very popular amongst frequentists.

In both the frequentist and Bayesian approaches, one can think of intuitively awkward issues. In the frequentist approach, the sample size is calculated under the assumption of a value for the unknown variance. In the Bayesian one, empirical  $\xi$  is far from the one calculated. This is not surprising as a single value for  $\nu$  is not the data generating mechanism assumed in the Bayesian model. This discrepancy reduces the larger the new data is relatively to the old. It is hard to imagine a data-generating mechanism that depends on how much knowledge one has for its parameters.

We try to bridge these gaps by an application of the *PDCCPP*. Essentially an Empirical Bayes methodology, it facilitates both the Bayesian belief and the frequentist operational characteristics in the design and analysis of a clinical trial. We argue that these are both features that could be of interest in conducting research in small or sensitive populations.

### Declaration of conflicting interests

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

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the European Union's seventh framework programme (FP7-HEALTH-2013-INNOVATION-1, Grant-Agreement No. 603160, ASTERIX).

### ORCID iD

TB Brakenhoff  <http://orcid.org/0000-0003-3543-6296>

## Supplemental material

Supplemental material for this article is available online.

## References

1. Friede T and Miller F. Blinded continuous monitoring of nuisance parameters in clinical trials. *J Royal Stat Soc: Series C (Appl Stat)* 2012; **61**: 601–618.
2. Gould A. Sample size re-estimation: recent developments and practical considerations. *Stat Med* 2001; **20**: 2625–2643.
3. Jennison C and Turnbull B. *Group sequential methods with applications to clinical trials*. Boca Raton, USA: Chapman & Hall/CRC, 2000.
4. Denne J. Sample size recalculation using conditional power. *Stat Med* 2001; **20**: 2645–2660.
5. Mehta C and Tsiatis A. Flexible sample size considerations using information-based interim monitoring. *Drug Inform J* 2001; **35**: 1095–1112.
6. U.S. Department of Health and Human Services, Food and Drug Administration, et al. *Guidance for industry: adaptive design clinical trials for drugs and biologics (Draft guidance)* 2010. <https://www.fda.gov/downloads/drugs/guidances/ucm201790.pdf>
7. Gould A. Planning and revising the sample size for a trial. *Stat Med* 1995; **14**: 1039–1051.
8. Whitehead J, Valdés-Márquez E, Johnson P, et al. Bayesian sample size for exploratory clinical trials incorporating historical data. *Stat Med* 2008; **27**: 2307–2327.
9. Brown B, Herson J, Atkinson E, et al. Projection from previous studies: a Bayesian and frequentist compromise. *Control Clin Trials* 1987; **8**: 29–44.
10. Lecoutre B. Two useful distributions for Bayesian predictive procedures under normal models. *J Stat Plan Inference* 1999; **79**: 93–105.
11. Lee S and Zelen M. Clinical trials and sample size considerations: another perspective. *Stat Sci* 2000; **15**: 95–110.
12. Spiegelhalter D, Freedman L and Parmar M. Applying Bayesian ideas in drug development and clinical trials. *Stat Med* 1993; **12**: 1501–1511.
13. Berry S, Carlin B, Lee J, et al. *Bayesian adaptive methods for clinical trials*. Boca Raton, USA: CRC Press, 2010.
14. Claxton K, Lacey L and Walker S. Selecting treatments: a decision theoretic approach. *J Royal Stat Soc: Ser A (Stat Soc)* 2000; **163**: 211–225.
15. Sahu S and Smith T. A Bayesian method of sample size determination with practical applications. *J Royal Stat Soc: Ser A (Stat Soc)* 2006; **169**: 235–253.
16. Stallard N. Sample size determination for phase II clinical trials based on Bayesian decision theory. *Biometrics* 1998; **54**: 279–294.
17. Joseph L, Wolfson D and Du Berger R. Some comments on Bayesian sample size determination. *Statistician* 1995; **44**: 167–171.
18. Pezeshk H. Bayesian techniques for sample size determination in clinical trials: a short review. *Stat Meth Med Res* 2003; **12**: 489–504.
19. Pham-Gia T and Turkkan N. Sample size determination in Bayesian analysis. *Statistician* 1992; **41**: 389–397.
20. Gould A. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Stat Med* 1992; **11**: 55–66.
21. Wang M. Sample size reestimation by Bayesian prediction. *Biometric J* 2007; **49**: 365–377.
22. Hartley A. Adaptive blinded sample size adjustment for comparing two normal means a mostly Bayesian approach. *Pharmaceut Stat* 2012; **11**: 230–240.
23. Hartley A. A Bayesian adaptive blinded sample size adjustment method for risk differences. *Pharmaceut Stat* 2015; **14**: 488–514.
24. Zhong W, Koopmeiners J and Carlin B. A two-stage Bayesian design with sample size reestimation and subgroup analysis for phase II binary response trials. *Contemporary Clin Trials* 2013; **36**: 587–596.
25. Ibrahim J and Chen M. Power prior distributions for regression models. *Stat Sci* 2000; **15**: 46–60.
26. De Santis F. Using historical data for Bayesian sample size determination. *J Royal Stat Soc, Ser A* 2007; **170**: 95–113.
27. Hobbs B, Carlin B, Mandrekar S, et al. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* 2011; **67**: 1047–1056.
28. Gravestock I and Held L. Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceut Stat* 2017; **16**: 349–360.
29. Nikolakopoulos S, van der Tweel I and Roes K. Dynamic borrowing through empirical power priors that control type I error. *Biometrics* Epub ahead of print 11 December 2017. DOI: 10.1111/biom.12835
30. Gsponer T, Gerber F, Bornkamp B, et al. A practical guide to Bayesian group sequential designs. *Pharmaceut Stat* 2014; **13**: 71–80.
31. Ibrahim J, Chen M, Gwon Y, et al. The power prior: theory and applications. *Stat Med* 2015; **34**: 3724–3749.
32. van Woensel J, Van Aalderen W, De Weerd W, et al. Dexamethasone for treatment of patients mechanically ventilated for lower respiratory tract infection caused by respiratory syncytial virus. *Thorax* 2003; **58**: 383–387.
33. van Woensel J, Wolfs T, Van Aalderen W, et al. Randomised double blind placebo controlled trial of prednisolone in children admitted to hospital with respiratory syncytial virus bronchiolitis. *Thorax* 1997; **52**: 634–637.