Contemporary Clinical Trials 87 (2019) 105858

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Bayesian hypothesis testing with frequentist characteristics in clinical trials

Hui Quan*, Bingzhi Zhang, Yu Lan, Xiaodong Luo, Xun Chen

Biostatistics and Programming, Sanofi, 55 Corporate Drive, Bridgewater, NJ 08807, United States of America

ARTICLE INFO

Keywords:
Inferential prior
Design posterior
Historical data borrowing
Type I error rate
Power
Interim analysis

ABSTRACT

Through the use of an informative prior, Bayesian methodologies could potentially borrow the strength of historical information and become more and more popular for their applications to clinical trials. Nonetheless, even with tremendous effort, the reconciliation of the formulation of the hypotheses and the calculation of type I error between a Bayesian analysis and traditional frequentist analysis is still not very clear. In this research, we apply an inferential prior, null prior and design prior to the Bayesian data analysis, type I error control and sample size calculation. As demonstrated theoretically, the type I error control denies any borrowing of favorable prior information. Thus, the use of the calibrated critical value obtained through simulation for the commensurate or power prior for a Bayesian analysis has the effect of eliminating the borrowing of historical information. The validity of a Bayesian analysis with the borrowing of historical data should rest on the a priori assumption of consistency of data from the historical and current studies. Just in case the consistency assumption is not totally true, dynamic borrowing through the commensurate or power prior can regulate the level of borrowing based on the degree of consistency in the data. An example along with simulations are used to illustrate the applications and compare the characteristics of the methods.

1. Introduction

With the advent of analytic tools, Bayesian analyses have become more and more popular and feasible for a wide set of statistical models applied to clinical research and new drug developments [1–9]. A Bayesian analysis has certain advantages over the traditional frequentist analysis. It allows the borrowing of historical information in a more natural and formal way through the application of an informative inferential prior in data analysis. This is particularly important for rare disease, oncology and pediatric studies. In these studies, recruitment of patients is very challenging and the utilization of all available information becomes very crucial to the feasibility of new drug development. Another advantage of a Bayesian analysis is that it provides the flexibility for quantifying the treatment effect through the posterior distribution for making a probabilistic statement rather than just the *p*-value, point estimate and confidence interval. To facilitate the applications of Bayesian analyses, the US FDA has issued a Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials [10], which indicates that health authorities are gradually accepting results via Bayesian analyses.

Statistical literature is rich with a large volume of publications on theory and methods of Bayesian analyses. For the quantification of treatment effect, many authors discuss sample size calculations for achieving the specific length of the Bayesian credible interval [11] for the desired coverage level (average coverage criterion [12]) or the fixed coverage level with the desired length of the credible interval (average length criterion [13]). In terms of hypothesis testing, different considerations and approaches have been proposed. Goodman [14,15] introduces a term “The P Value Fallacy”, the mistaken idea that a single number can capture both the long-term outcomes of an experiment and the evidential meaning of the result. He promotes the use of a Bayesian Factor (BF) as an objective measure of the evidential strength. Berger and Sellke [16] further illustrate the conflict between the *p*-value and posterior probability of the null, concluding that a *p*-value can be a highly misleading measure of the evidence against the null hypothesis. Reyes and Ghosh [17] propose the control of the total weighted average of type I and type II error rates calculated under different prior distributions. Rather than controlling the maximum type I error rate calculated in the domain of the null hypothesis, Psioda and Ibrahim [18] advocate the control of the average type I error rate derived using the inferential prior conditioned under the null hypothesis domain. Berger [19] on the other hand accepts or rejects the hypotheses by minimizing the expected posterior loss after the specification of a loss function. Clearly, some of these methods discussed further in Section 2 are somewhat different from the traditional frequentist methods.

Frequentist properties for the Bayesian analyses may be requested

* Corresponding author.
E-mail address: hui.quan@sanofi.com (H. Quan).

https://doi.org/10.1016/j.cct.2019.105858
Received 18 April 2019; Received in revised form 18 September 2019; Accepted 21 September 2019
Available online 24 October 2019

by the health authorities to measure the frequency of making the correct decisions [20]. Upon all available research outcomes, we conduct further research with the primary objective to fully understand the basis for the validity of Bayesian hypothesis testing. As demonstrated in Section 3, the borrowing of favorable historical data in a Bayesian analysis potentially inflates the type I error. To ease the concern, many authors propose methods such as the power prior, commensurate power prior and other approaches for dynamic borrowing of the historical data based on the level of consistency between the historical and current data. They also use simulation to obtain the calibrated adjusted threshold or critical value for hypothesis testing and type I error rate control. Nonetheless, as shown in our simulation in Section 7, these measures may actually have the effect of no borrowing at all. The assumption that supports the historical data borrowing is the combinability of data from the historical and current studies assessed by the consistency criteria of Pocock [21]. These along with the sample size calculation using a design prior and considerations for interim analyses will be discussed in the following sections.

2. Some existing approaches for Bayesian hypothesis testing

Suppose θ is the quantity of interest for evaluating either a within or between treatment effect depending on the circumstances. The primary objective for most confirmatory clinical trials is to assess whether a treatment has a positive effect. Thus, we either test the one-sided null hypothesis $H_0: \theta \le \theta_0$ versus the alternative hypothesis $H_1: \theta > \theta_0$ or test the two-sided null hypothesis $H_0: \theta = \theta_0$ versus the alternative $H_1: \theta \ne \theta_0$ with a sign checking after the rejection of the null hypothesis. Approaches for frequentist hypothesis testing are straightforward. Nonetheless, different approaches and considerations have been proposed for Bayesian hypothesis testing. In this section, we discuss several of them which have somewhat different characteristics compared to those of a frequentist test.

Setting $\theta_0 = 0$, Guo and Heitjan [22] test a simple null hypothesis $H_0: \theta = 0$ versus a two-sided composite alternative $H_1: \theta \ne 0$. They start by letting the prior probability for H_0 (H_1) to be π_0 ($\pi_1 = 1\text{-}\pi_0$). The prior distribution of θ under the alternative is commonly specified as a normal prior centered at 0 with a standard deviation based on the prior relevant studies. With the marginal density $\Pr(y|H_i)$ of data y under H_i, the posterior probability of H_i is

$$\Pr(H_i \mid y) = \frac{\pi_i \Pr(y \mid H_i)}{\pi_0 \Pr(y \mid H_0) + \pi_1 \Pr(y \mid H_1)}. \quad (1)$$

The Bayes Factor (BF) is then

$$BF = \frac{\Pr(H_0 \mid y)/\Pr(H_1 \mid Y)}{\pi_0/\pi_1} = \frac{\Pr(y \mid H_0)}{\Pr(y \mid H_1)}. \quad (2)$$

A small BF indicates the support of the alternative hypothesis. A natural way to calibrate a Bayesian test is to select a threshold BF* [1] such that

$$\Pr(BF < BF^* \mid H_0) = \alpha. \quad (3)$$

With this approach, it is not very clear how we can connect hypothesis testing with the quantification of treatment effect.

Bogdan et al. [6], on the other hand, consider the scenario of Bayesian multiple tests where the data have either the null distribution $N(0, \sigma^2)$ or non-null distribution $N(\mu, \sigma^2)$ with μ following a prior distribution $N(0, \tau^2)$. Hence the unconditional non-null distribution of the data is essentially $N(0, \sigma^2 + \tau^2)$ which simply has just one more variance component compared to the one of the null distribution. Suppose the probability that data to be generated by the non-null distribution is p. Then the marginal distribution for the data will be

$$(1 - p)N(0, \sigma^2) + pN(0, \sigma^2 + \tau^2)$$

with mean 0. Since the goal is to test whether data have a null or non-null distribution, the use of a subjective and informative prior on p is recommended for the test. These and other additional methods (e.g., [16,23]) are for a two-sided test and do not directly connect with our goal of assessing positive treatment effect.

The BF approach can be extended to a one-sided test. Suppose $\pi_0(\theta)$ is a prior for $H_0: \theta \le 0$ and $\pi_1(\theta)$ a prior for $H_1: \theta > 0$. The marginal density of the data under these priors can be derived to form the BF in (2), then the corresponding threshold (see (3)) can be derived for the one-sided test. There is also another approach [19] for accepting or rejecting the hypotheses based on the Bayesian decision scheme. Let a_i ($i = 0, 1$) be the decision of accepting H_i. The ultimate goal is to minimize the expected posterior loss with a general form of the loss function

$$L(\theta, a_i) = \begin{cases} 0 & \text{if } \theta \in H_i \\ K_i & \text{if } \theta \in H_j, j \ne i \end{cases}.$$

The optimal decision a_1 is to reject H_0 if and only if

$$\frac{\Pr(H_0 \mid y)}{\Pr(H_1 \mid y)} < \frac{K_0}{K_1}$$

where $\Pr(H_i|y)$ is defined by (1). When $K_0 = K_1$, we choose the most probable hypothesis. Minimizing the expected posterior loss is different from the type I error probability control.

Reyes and Ghosh [17] apply a Bayesian average error (or marginal error) framework to hypothesis testing. Suppose we reject the null hypothesis H_0 if $T(y) > t$ for a statistic T and some value t. The average type I error probability is

$$AE_0(t) = \int \Pr(T(y) > t \mid \theta)\pi_0(\theta)d\theta$$

and the average type II error probability is

$$AE_1(t) = \int \Pr(T(y) \le t \mid \theta)\pi_1(\theta)d\theta.$$

The total weighted error probability for a pre-specified weight w is

$$TWE(t, w) = wAE_0(t) + (1 - w)AE_1(t).$$

The goal is to control the total weighted error probability with appropriate cutoff t and sample size. This is somewhat different from a traditional frequentist test where we first control the type I error rate through a valid test then control the type II error rate through the increase in sample size and we do not mix the two types of error together.

In the above methods, it is not very obvious how the historical data through a potentially favorable prior can be utilized in the Bayesian analysis or posterior distribution calculation.

3. Reconciliation between Bayesian and frequentist hypothesis testing

In this section, we discuss the reconciliation between the Bayesian and frequentist hypothesis testing, particularly the impact of the historical data borrowing on type I error.

Let $\pi(\theta)$ be the inferential prior distribution [24] of θ which will be used in statistical inference when data of the current study is analyzed. The $\pi(\theta)$ could relate to the posterior distribution derived from the historical data and therefore be informative. Given θ, let $f(y|\theta)$ be the likelihood of data y observed from the current study. The updated posterior distribution of θ is then

$$\pi(\theta \mid y) \propto f(y \mid \theta)\pi(\theta).$$

One intuitive and straightforward approach for Bayesian hypothesis testing is to use the posterior distribution or predictive probability. If

$$\Pr(\theta \le \theta_0 \mid y) \le \alpha$$

for a small α, it is reasonably justifiable to reject the null hypothesis $H_0: \theta \le \theta_0$.

Let's first consider the case of a one sample continuous endpoint which follows a normal distribution to get the insight. By specifying a distribution, we will be able to see the derivations and arguments much

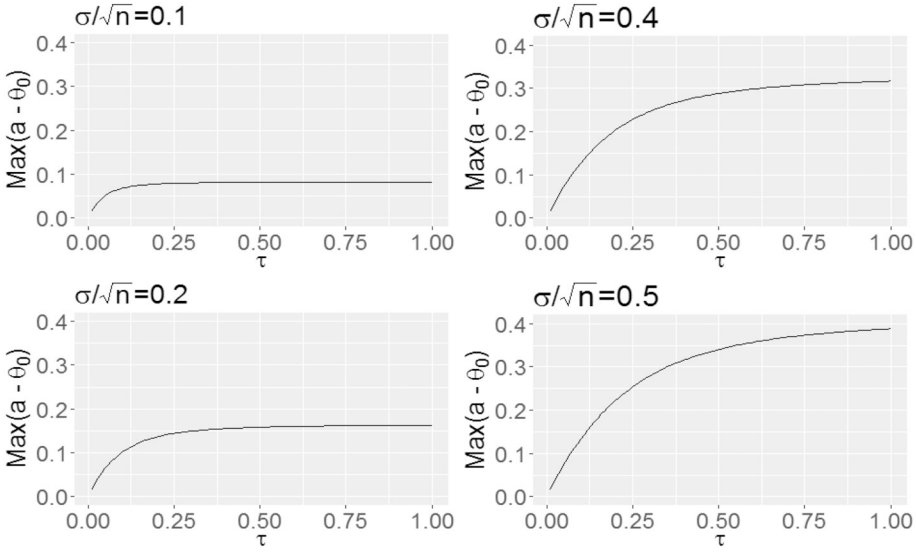

Fig. 1. Value of $\max(a-\theta_0)$ for (13) as a function of the other parameters.

clearer. Suppose iid $y_1, \ldots, y_n \sim N(\theta, \sigma^2)$ where σ^2 is known (if it is unknown, a consistent estimate can be used to replace it). For the classical frequentist testing of the null $H_0: \theta \le \theta_0$ versus the alternative $H_1: \theta > \theta_0$ hypothesis, we will reject H_0 with type I error rate controlled at α if $\bar{y} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$ where $\bar{y}$ is the sample mean and $z_{1-\alpha}$ is the $1-\alpha$ percentile of the standard normal distribution. To have $1-\beta$ power for detecting $\theta = \theta_1$ or

$$1-\beta = \Pr\left(\bar{y} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha} \mid \theta = \theta_1\right) = \Phi\left(\sqrt{n}\frac{\Delta}{\sigma} - z_{1-\alpha}\right) \tag{4}$$

where $\Delta = \theta_1 - \theta_0$, the required sample size for (4) is

$$n = (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma}{\Delta}\right)^2. \tag{5}$$

For a Bayesian approach, let the inferential prior or fitting prior to be used in the final statistical inference and to determine the historical data borrowing be $\pi(\theta) \sim N(a, \tau^2)$, where a and τ^2 are known fixed values rather than unknown parameters. The overall likelihood is then

$$\pi(\theta) \times N\left(\bar{y} \mid \theta, \frac{\sigma^2}{n}\right) = N(\theta \mid a, \tau^2) \times N\left(\bar{y} \mid \theta, \frac{\sigma^2}{n}\right)$$

and the corresponding posterior distribution for θ is

$$N\left(\frac{\frac{n\bar{y}}{\sigma^2} + \frac{a}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right). \tag{6}$$

This posterior distribution provides the updated information regarding θ and will be directly applied to Bayesian hypothesis testing. From (6), a Bayesian analysis can be simply treated as a combined analysis combining data from the historical and current studies. Specifically, the posterior probability for the null hypothesis H_0 is

$$\Pr(\theta \le \theta_0 \mid \bar{y}) = \Phi\left[\left(\theta_0 - \frac{\frac{n\bar{y}}{\sigma^2} + \frac{a}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}\right) \Big/ \sqrt{\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}}\right]. \tag{7}$$

Intuitively, one can directly reject H_0 once

$$\Pr(\theta \le \theta_0 \mid \bar{y}) \le \alpha \tag{8}$$

for a test at significance level α. However, this may not always be the case as will be seen later. The domain of $\bar{y}$ derived from (7) for (8) to hold is

$$\left(\frac{\bar{y}}{\frac{\sigma^2}{n}} + \frac{a}{\tau^2}\right) \Big/ \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) \ge \theta_0 + z_{1-\alpha} \Big/ \sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \tag{9}$$

(which has the form of the frequentist test for a combined analysis) or

$$\bar{y} \ge \frac{\sigma^2}{n}\left[\theta_0\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) + z_{1-\alpha}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a}{\tau^2}\right]. \tag{10}$$

Suppose a null prior distribution of θ in the domain of H_0 is $f_0(\theta)$. The marginal average type I error rate with (10) based on this distribution will be

$$\int Pr\left[\bar{y} \ge \frac{\sigma^2}{n}\left(\theta_0\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) + z_{1-\alpha}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a}{\tau^2}\right) \mid \theta\right] f_0(\theta) d\theta. \tag{11}$$

Nonetheless, to be consistent with the frequentist approach, the focus is on the maximum value of (11) for all possible $f_0(\theta)$ which is achieved when $f_0(\theta)$ puts all the mass on θ_0, the boundary of H_0. Then (11) becomes

$$1 - \Phi\left[\sqrt{n}\left(\frac{\sigma^2}{n\tau^2}\theta_0 + z_{1-\alpha}\frac{\sigma^2}{n}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a\sigma^2}{n\tau^2}\right) \Big/ \sigma\right]. \tag{12}$$

It is α when $\tau^2 = \infty$ or the non-informative inferential prior is applied regardless of the value of a. In such a case, (10) will become $\bar{y} \ge \theta_0 + z_{1-\alpha}\sigma\sqrt{\frac{1}{n}}$, the same rejection criterion for $\bar{y}$ as in the regular frequentist testing. For finite τ^2, as long as

$$a - \theta_0 < \frac{z_{1-\alpha}\left(\sqrt{1 + \frac{\sigma^2}{n\tau^2}} - 1\right)\sqrt{n}\tau^2}{\sigma}, \tag{13}$$

(12) will be less than α, or the use of α in (8) will control the type I error rate of testing H_0 to be no more than α. Note that some a values which are larger than θ_0 can still satisfy (13). That implies that the use of a slightly favorable inferential prior will still control the type I error probability even the nominal α is used in (8), which may be due to the increased variability from $\tau^2 > 0$.

Fig. 1 shows the value of $\max(a-\theta_0)$ for (13) as a function of the other parameters. It is an increasing function of the variance τ^2 of the inferential prior distribution and $\frac{\sigma^2}{n}$. In the figure, if $\sigma = 2$ and $\sigma/\sqrt{n} = 0.1$, the sample size $n = 400$.

For $\tau^2 \ne \infty$ and other a values which do not satisfy (13), it cannot be guaranteed that (12) will always be no more than α. To ensure the

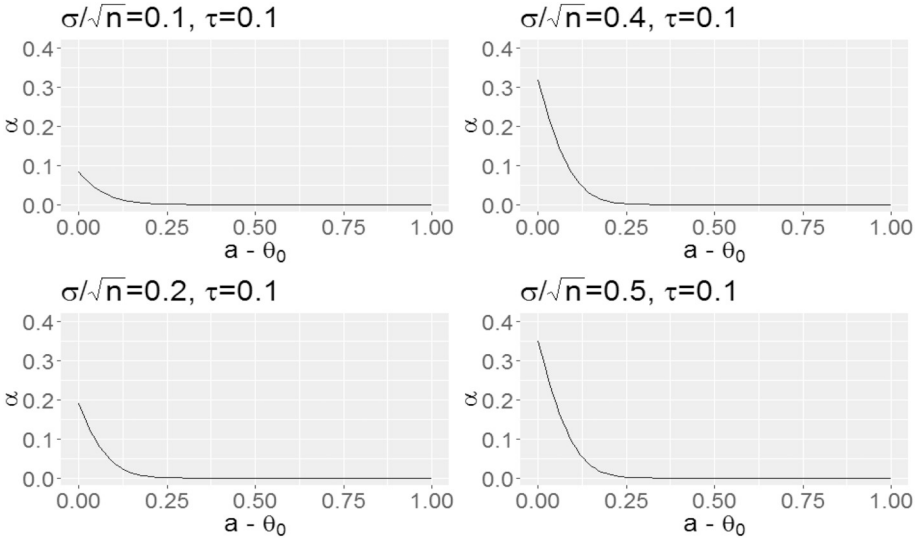

Fig. 2. Value of α for controlling the type I error probability to be $\alpha' = 0.025$.

control of the type I error probability at a desired level α' or for (12) to be

$$1 - \Phi\left[\sqrt{n}\left(\frac{\sigma^2}{n\tau^2}\theta_0 + z_{1-\alpha}\frac{\sigma^2}{n}\sqrt{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} - \frac{a\sigma^2}{n\tau^2}\right)/\sigma\right] = \alpha',$$

the corresponding calibrated α level in (8) should satisfy

$$z_{1-\alpha} = \frac{z_{1-\alpha'} + \frac{\sigma}{\sqrt{n}\tau^2}(a-\theta_0)}{\sqrt{1+\frac{\sigma^2}{n\tau^2}}} \tag{14}$$

or

$$\alpha = 1 - \Phi\left(\frac{z_{1-\alpha'} + \frac{\sigma}{\sqrt{n}\tau^2}(a-\theta_0)}{\sqrt{1+\frac{\sigma^2}{n\tau^2}}}\right) \tag{15}$$

which depends on the assumed values of the parameters and can be larger or smaller than α'. Basically, the larger the treatment effect reflected in the inferential prior, the higher the level of the calibration. Fig. 2 shows the value of α in relation to the other parameters and sample size of the current study for controlling the type I error rate to be $\alpha' = 0.025$. It is a decreasing function of $a - \theta_0$.

Combining (10) and (14), we have

$$\bar{y} \geq \frac{\sigma^2}{n}\left[\theta_0\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) + \frac{z_{1-\alpha'} + \frac{\sigma}{\sqrt{n}\tau^2}(a-\theta_0)}{\sqrt{1+\frac{\sigma^2}{n\tau^2}}}\sqrt{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} - \frac{a}{\tau^2}\right]$$

$$= \theta_0 + \frac{\sigma}{\sqrt{n}}z_{1-\alpha'}. \tag{16}$$

The right side of (16) is exactly the critical value at significance level α' via the non-informative prior. In other words, to control the type I error probability at the exact α' level for testing H_0, the calibrated α in (8) should be derived based on (15), which results in the critical value of a frequentist test for $\bar{y}$. This is consistent with the conclusion reached by Psioda and Ibrahim [18] whose derivations were based on a power prior. From the above, for a non-favorable or slightly favorable inferential prior, we want to increase the α in (8) to avoid being conservative when the posterior probability is applied for testing H_0. Also, for a too favorable inferential prior, we need to reduce α to avoid being liberal. These will be confirmed by an example and simulations in Section 7. In all, an informative (no matter favorable or non-favorable) inferential prior should not be used in Bayesian hypothesis testing with a frequentist characteristic, in order to have the type I error probability exactly controlled at the nominal level.

Several authors discuss different approaches for type I error probability calculations for Bayesian hypothesis testing. Psioda and Ibrahim [18] first derive the conditional null prior $\pi(\theta|H_0)$ based on the inferential prior $\pi(\theta)$, and then calculate the average type I error probability under this conditional null prior as (11). This average type I error probability depending on the specific $\pi(\theta)$ is smaller than the type I error probability calculated at the boundary $\theta = \theta_0$ and therefore can be controlled at the nominal level even with certain borrowing of the favorable history information. Nonetheless, this average type I error probability is not the type I error rate in the traditional frequentist sense.

Hobbs et al. [25] use a Commensurate Power Prior (CPP discussed later) to borrow the historical data in a dynamic way through the informative inferential prior. To control the type I error, they use simulation to obtain the calibrated adjusted critical value or threshold like the one of (15). In the simulation, they assume no treatment effect or $\theta = \theta_0$ to generate new data for the current study. However, there may still be a treatment effect embedded in the informative inferential prior or historical data, which would make it difficult to interpret the type I error probability. In that case, the simulated data for the current study will not be consistent with any favorable inferential prior, and the historical data will be substantially discounted similar to no borrowing due to the use of the calibrated critical value. Nonetheless, their simulation demonstrates that their method still has slightly higher power compared to the one without borrowing any information. One interpretation could be as follows. For the fixed borrowing, the levels of borrowing are the same (e.g., power = 0 in the power prior) for the derivation of the calibrated critical value and power calculation, which results in no borrowing for the purpose of type I error control. For the dynamic borrowing, the levels of borrowing for the derivation of the calibrated critical value and power calculation are different. There may be a substantial discount in the historical data for the calibrated critical value calculation when the inferential prior is not so consistent with the simulated data for the current study where no treatment effect is assumed, if for example when the power is much smaller than 1 or close to 0 in the power prior. For power calculations when the inferential prior is somewhat consistent with the simulated data assuming a certain treatment effect for the current study, one may borrow more historical data if for example the power for the power prior may be closer to 1. If we use simulation to verify the type I error rate control, we should assume no treatment effects for both the historical and current studies rather than just the current study to generate data.

Based on above discussion, to use a substantially favorable inferential prior in Bayesian hypothesis testing and at the same time control the type I error, we need to treat the historical data presented in the inferential prior as a part of the overall data (see also [20] and (9)). Then the inferential prior for the analysis of the current study can be viewed as the posterior distribution of the historical study with a non-informative prior. Thus, the type I error will be controlled for the combined data of the historical and current studies even we use the nominal threshold level for (8). As a result, we do not really need to use simulation to obtain the calibrated threshold as in [25].

Before making the decision of borrowing the historical data, we first need to assess the combinability. This should be done preferably before any data including historical data unblinding. It is reasonable to combine data from multiple studies in a stratified analysis manner if they are conducted using similar protocols, patient populations (inclusion/exclusion criteria), same endpoint, same treatment and similar study centers [21]. Adjustment by covariates in the analysis models will also help the exchangeability of the parameters of the treatment effects across the sources. Moreover, just in case the observed data from different data sources after unblinding reveal heterogeneity, a mechanism can be put in place to determine the amount of the historical data borrowing based on the degree of the heterogeneity.

There are many methods for this purpose.One is the commensurate power prior framework of Hobbs et al. [25] mentioned previously. Assuming different parameters θ_h and θ for the treatment effects of the historical and current studies, the initial prior for the historical study should be non-informative $\theta_h \propto 1$. The posterior distribution for θ_h for the historical study is then $N(\theta_h|\, a,\, \tau^2) = q_0(\theta_h|\, a,\, \tau^2)$. If we further assume $\theta \sim N(\theta_h, \eta^2)$ where η^2 is the commensurate parameter with a prior $\pi_0(\eta^2)$ and a power prior $\pi_0(c|\,\eta)$ $(0 \le c \le 1)$, the whole prior for the current study will be

$$\pi^{CPP}(\theta, c, \eta) = N\left(\theta \mid a, \eta^2 + \frac{\tau^2}{c}\right)\pi_0(c \mid \eta)\pi_0(\eta^2) \tag{17}$$

where τ^2 is the variance of the estimate a of the treatment effect based on the historical data. When $\pi_0(\eta^2 = \infty) = 1$ or $\pi_0(\tau^2 = \infty) = 1$ (the latter will not hold as long as the historical study has any data), the prior for θ of the current study will be a non-informative prior and no historical data will be borrowed in the analysis. Otherwise, the CPP method will borrow historical information even if θ_h and θ are different. Note that CPP allows two channels for discounting the historical data; one is through the power prior and another through treating θ_h and θ differently. The degree of the heterogeneity of the two data sets will determine the amount of borrowing. A special case of (17) when $\eta^2 = 0$ with probability 1 is the Modified Power Prior of Duan et al. [26]

$$\pi^{MPP}(\theta, c) = N\left(\theta \mid a, \frac{\tau^2}{c}\right)\pi_0(c).$$

For a supervised power prior, Psioda and Ibrahim [18] propose to set $c = \hat{c} = \left(\frac{L(a \mid D)}{L(\bar{Y} \mid D)}\right)^{s_0} < 1$, where L is the likelihood, D represents the data from the current study and s_0 is a calibration parameter. Evidently, such a power $\hat{c}$ is determined by the distance between the prior mean a and the sample mean $\overline{Y}$ of the current study, or $|a - \overline{Y}|$ for normal distributed data. When the two means are very different, $\hat{c}$ will be close to zero and less historical information will be borrowed.

Even though the historical and current studies may be very similar in terms of the trial characteristics, and a stratified combined analysis may have been pre-specified, the sponsor will only conduct the new study and borrow information from the historical study if the historical study is positive,. As a result, there will always be potential selection bias when the historical data are incorporated in the Bayesian analysis. This selection bias may be a concern but cannot be easily addressed with any methods.

4. Power calculation

In order to design a study allowing historical data borrowing under the assumption of combinability of data from the historical and current studies through Bayesian hypothesis testing, we need to calculate the sample size for the current study to ensure desired assurance probability or power for (8). The calculation can be based on an assumed prior distribution for the data of the current study, just as we assume $\theta = \theta_1$ in the alternative hypothesis domain for the frequentist setting. This prior distribution can be called the sampling prior [24] or more specifically the design prior. Psioda and Ibrahim [18] use the conditional prior $\pi(\theta|\, H_1)$ as the design prior where $\pi(\theta)$ is the inferential prior. Nonetheless, if $\pi(\theta)$ is a non-informative inferential prior, $\pi(\theta|\, H_1)$ may also be a non-informative prior and such an approach will not work. As θ may truly have value below θ_0 with a small probability, we may not confine the value of θ to be within H_1 and instead use a more general design prior $\theta \sim N(b, \gamma^2)$ which could be determined through historical information and other assumptions. Here the γ^2 is a measure of the between study variability and is different from the τ^2 in the inferential prior as $1/\tau^2$ is a measure of the amount of historical data borrowing. Under the design prior, the marginal distribution of $\bar{y}$ is

$$\bar{y} \sim N\left(b, \gamma^2 + \frac{\sigma^2}{n}\right).$$

Assuming certain discounting has already been taken into account in the inferential prior $N(a, \tau^2)$ which will be used directly in the inference, if no further discounting will be considered in data analysis for the purpose of sample size calculation, the assurance probability for (8) is then

$$\begin{aligned} Pr\Bigg[(\bar{y} - b)/\sqrt{\gamma^2 + \frac{\sigma^2}{n}} \ge \Bigg(\frac{\sigma^2}{n}\left(\theta_0\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) + z_{1-\alpha}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a}{\tau^2}\right) \\ - b\Bigg)/\sqrt{\gamma^2 + \frac{\sigma^2}{n}}\Bigg] \end{aligned}$$

$$= 1 - \Phi\left[\frac{\left(\frac{\sigma^2}{n}\left(\theta_0\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) + z_{1-\alpha}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a}{\tau^2}\right) - b\right)}{\sqrt{\gamma^2 + \frac{\sigma^2}{n}}}\right]. \tag{18}$$

For (18) to be at least $1 - \beta$, the sample size n and the other parameters should satisfy

$$\frac{\left(b - \frac{\sigma^2}{n}\left(\theta_0\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) + z_{1-\alpha}\sqrt{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} - \frac{a}{\tau^2}\right)\right)}{\sqrt{\gamma^2 + \frac{\sigma^2}{n}}} \ge z_{1-\beta}. \tag{19}$$

If $\tau^2 = \infty$ or with a non-informative inferential prior, (19) becomes

$$z_{1-\alpha}\sqrt{\frac{\sigma^2}{n}} + z_{1-\beta}\sqrt{\gamma^2 + \frac{\sigma^2}{n}} \le b - \theta_0. \tag{20}$$

Since the left side of (20) is a decreasing function of n, for a reasonably small γ, there is an n such that (20) holds. Actually, (18) with $\tau^2 = \infty$ is often called the probability of success, and the n derived through (20) is the sample size for a $1 - \beta$ probability of success (or average power under the distribution $\theta \sim N(b, \gamma^2)$) in a frequentist setting. The sample size derived from (20) can be much larger than the one from (5) as variability $\gamma^2 > 0$ is incorporated. If we further assume $\gamma^2 = 0$ (a degenerated design prior) in (20), the required sample size will be

$$n = (z_{1-\alpha} + z_{1-\beta})^2\left(\frac{\sigma}{b - \theta_0}\right)^2$$

which is the same as (5) with θ_1 replaced by b. Note that even when $\tau^2 = \infty$, if γ^2 is not small and b is not large enough, there may not be an

n such that (20) holds. Rather than a design prior distribution $\theta \sim N(b, \gamma^2)$, similar to Psioda and Ibrahim [18], we can also consider a truncated distribution such that θ is larger than θ_0 with a margin so that the assurance probability or power is always guaranteed with a reasonable sample size. If the required sample size for the desired probability of success is very large, we can calculate the minimum observed treatment effect that is statistically significant for the final analysis based on the sample size. If such an observed treatment effect is too small to be clinically meaningful, we may just decrease the sample size accordingly even though the probability of success may not be at the desired level.

There may be very rich data from historical studies such as in the case when we extrapolate clinical trial results from an adult population to a pediatric population. Even if the pediatric study may appear to be very consistent with the adult study in all aspects other than age, we may still want to limit the amount of borrowing from the adult population for the assessment of the treatment effect for the pediatric population. For that, we need first evaluate the effective sample size from the historical data set. From the estimation perspective, Pennello and Thompson [20] use

$$ESS = n\frac{Var(\theta \mid D)}{Var(\theta \mid D, \pi(\theta))} \tag{21}$$

as the effective sample size for the new analysis. If $\tau^2 = \frac{\sigma^2}{n_0}$, (21) will become $n\frac{1/n}{1/(n_0+n)} = n_0 + n$ if the distributions of the endpoint and the inferential prior are both normal distributions. For the purpose of testing H_0, the required sample size for the current study n^* obtained through solving (20) assuming $\tau^2 = \infty$ (no borrowing) may not be the same as the n' obtained through (19) under $\tau^2 \neq \infty$ (with borrowing) plus n_0 if $\gamma \neq 0$. In any case, we may just use $n' + n_0$ as the effective sample size for the new analysis. If n_0 is too large relative to n', we may increase τ^2 or equivalently reduce n_0 or use a small power in the power prior to curb the borrowing of historical information. With the updated inference prior, the sample size for the new study then has to be recalculated.

In the above sample size and power calculation, we assume that the full pre-specified inferential prior will be used for the final analysis and no further discounting will be considered. However, as alluded to earlier, in the actual analysis, the amount of borrowing of the historical data through the commensurate or power prior approach will depend on the degree of consistency between the observed data of the historical and current studies. As there will then be no corresponding closed form formula for the sample size calculation, simulation should be conducted assuming various design priors after the specification of the inferential prior to select an appropriate sample size for the current study.

The above is for the case of a one sample continuous endpoint. Methods and formulae for the other cases including between treatment comparisons and time-to-event endpoints are similar, we just need to change $\bar{y}$ to the estimator of the treatment effect and replace $\frac{\sigma^2}{n}$ by the asymptotic variance of the estimator of the treatment effect. Of note, if the hazard ratio is used for measuring the magnitude of the treatment effect for an adverse time-to-event endpoint such as a cardiovascular event, the smaller the hazard ratio of the treatment versus the control, the better the treatment outcome. Therefore, all the formulae should be modified correspondingly. This also applies when a decrease in the value of an endpoint represents a better treatment outcome.

5. Binary endpoint

We now consider the analysis of a binary endpoint starting from the one sample case as in Simon's [27] optimal two stage design. Given the event rate p and sample size n of the current study, the number of events X follows a binomial distribution $(X|p) \sim B(p, n)$. We commonly use a conjugate prior distribution for p which is a Beta distribution $p \sim Beta(a, b)$, $a \geq 0$ and $b \geq 0$. If X takes the value of x, the posterior distribution of p is still a Beta distribution $Beta(a + x, b + n\text{-}x)$ with parameters $a + x$ and $b + n\text{-}x$. The posterior distribution of a previous study can be the informative inferential prior distribution of a new study particularly if data from the two studies can be justified for the combined statistical inference via the criteria of Pocock [21]. To evaluate $H_0 : p \leq p_0$ at level α, we need the critical value x such that

$$\Pr(p \leq p_0 \mid X = x, n) \leq \alpha. \tag{22}$$

Here, no calibration on α is needed when combinability of data from the two sources can be justified. Based on the posterior distribution of p, (22) is the same as

$$\frac{1}{B(a + x, b + n - x)} \int_0^{p_0} p^{a+x-1}(1-p)^{b+n-x-1} dp \leq \alpha. \tag{23}$$

A critical value $x_0(\alpha)$ for X can be derived numerically for (23) to hold. On the other hand, after deriving $x_0(\alpha)$, for a null prior distribution $\pi_0(p)$ on $H_0 : p \leq p_0$, the average error rate similar to (11) for testing $H_0 : p \leq p_0$ can also be evaluated as

$$\int_0^{p_0} \sum_{x \geq x_0(\alpha)} \binom{n}{x} p^x (1-p)^{n-x} \pi_0(p) dp. \tag{24}$$

If $\pi_0(p)$ is a distribution with all the mass on p_0, (24) will become

$$\sum_{x \geq x_0(\alpha)} \binom{n}{x} p_0^x (1-p_0)^{n-x}. \tag{25}$$

Note that (23) (with $x = x_0(\alpha)$) and (25) use different concepts: (23) treats p as a random variable with a beta distribution and data x as fixed observed information for Bayesian analysis while (25) treats $p = p_0$ as a fixed parameter and x as a random variable with a binomial distribution in a kind of frequentist analysis. The frequentist approach without borrowing any prior information uses critical value $x'_0(\alpha)$ derived from

$$\sum_{x \geq x'_0(\alpha)} \binom{n}{x} p_0^x (1-p_0)^{n-x} \leq \alpha,$$

which can be compared to $x_0(\alpha)$ when $a = 1$ and $b = 1$ or the inferential prior is non-informative.

If $a > 1$ and $b > 1$, we can treat the informative prior as data of a study with a events among $a + b$ patients. Then the corresponding frequentist analysis can be viewed as the combined analysis of the observed data from the previous and current studies. The critical value for the combined data can be derived as

$$\sum_{x \geq x_0^*(\alpha)} \binom{a+b+n}{x} p_0^x (1-p_0)^{a+b+n-x} \leq \alpha. \tag{26}$$

As a result, the critical value for the current study is

$$x'_0(\alpha) = x_0^*(\alpha) - a.$$

Using the relationship between the beta distribution and binomial distribution, the left side of formula (23) can be rewritten as

$$\sum_{x \geq x_0^*(\alpha)} \binom{a+b+n-1}{x} p_0^x (1-p_0)^{a+b+n-x-1}, \tag{27}$$

which is different compared to formula (26) only on the number of total patients. Thus, using the prior information of Beta(a,b) in the Bayesian analysis is equivalent to adding a events among a + b-1 patients in the current study in a combined frequentist analysis for a binary endpoint. Comparing formula (26) and (27), the critical value found by either a Bayesian or frequentist analysis will be the same or very close.

For the power calculation for a Bayesian analysis of a binary endpoint, we use a design prior distribution $\pi_1(p)$, so that the average power is

$$\int \sum_{x \geq x_0(\alpha)} \binom{n}{x} p^x (1-p)^{n-x} \pi_1(p) dp. \tag{28}$$

The required sample size for (28) to be larger than the desired

power $1-\beta$ can be derived but has no closed form formula. Note that again for certain $\pi_1(p)$ which has mass close to H_0, we will need a huge sample size to achieve the desired power. Therefore, a truncated beta distribution for the power prior can be considered.

For between-treatment comparisons on a binary endpoint, we can first obtain the posterior distribution for each treatment then derive the posterior distribution for the difference of rates, the risk ratio or odds ratio [28]. For easy determination of the level of the historical data borrowing based on the degree of heterogeneity between the data sources, hierarchical models can also be utilized.

6. Interim analysis

Interim analyses may be planned for long term and large scale confirmative trials just in case the trial should be stopped early for futility or an efficacy claim, or should be modified based on the interim results for the desired trial characteristics. In certain disease areas, particularly if the enrollment is anticipated to be slow, the sponsor may want to perform interim looks of the trial data to timely obtain treatment information for facilitating the whole new drug development process. With repeated analyses, it is well understood that the appropriate adjustment is necessary to control the overall type I error rate under the frequentist analysis framework. Since the frequentist analysis is a special case of the Bayesian analysis, the corresponding adjustment should also be required for the Bayesian analysis [29].

When the informative commensurate power prior or other power prior approach is used to dynamically borrow historical data in a Bayesian analysis, the amount of borrowing for the interim analysis will depend on the level of consistency of the historical data and the observed interim data from the current study and cannot be pre-specified. For the final analysis, if the amount of borrowing will be determined further by consistency of the historical data and the final data of the whole current study, the amount of borrowing for the final analysis may be different from the one for the interim analysis. The calculations of the critical values for the interim and the final analyses will not be straightforward.

A simple but conservative approach for handling the issue is to use a Bonferroni type multiplicity adjustment. That is, we can use α_1 for the interim analysis then use $\alpha-\alpha_1$ for the final analysis. This approach will always control the type I error probability as long as the test at each stage is valid.

7. Example and simulation

To illustrate the methods and serve as the setting for simulation, we use a slightly modified version of the example of Sahu and Smith [30]. The endpoint follows a normal distribution with mean θ and standard deviation $\sigma=2$. The hypotheses are of the form: $H_0:\theta\le 0$ and $H_1:\theta>0$ with a value of $\theta>0$ favoring the treatment. With a classical or frequentist setup, to detect a treatment effect of $\theta_a=0.29$, 0.39 and 0.56 at a one-sided significance level of 2.5% with 80% power, the required sample size would be 376, 209 and 102, respectively. For a Bayesian analysis, an inferential prior for θ summarized by $\theta\sim N(0.39, 0.2^2)$ reflects an enthusiastic treatment effect in the mean and an effective sample size from the historical source of 100 ($0.2^2=\sigma^2/100$) in the variance.

For power calculations and comparisons, 3 different means for the design priors are considered: $\theta\sim N(0.29,\gamma^2)$, $N(0.39,\gamma^2)$ and $N(0.56,\gamma^2)$ where $\gamma=0$, 0.025, 0.05 and 0.1. Table 1 shows the results for the fixed 0% or 100% borrowing of the historical information. With no borrowing, the power for $\gamma^2=0$ is just the regular frequentist power which is 80% by design for the corresponding sample size, although due to rounding in the sample size, the power could be slightly different from 80%. The power for $\gamma^2>0$ is the probability of success which is smaller than the one of $\gamma^2=0$ particularly if γ^2 is substantially larger than zero. With the fixed 100% borrowing through the very favorable inferential prior, we basically combine data from the historical and current studies in the analysis. The power for the combined analysis via (18) can be as high as around 97% and decreases with γ^2. The unadjusted type I error probability via (12) can be larger than 20% in some cases, with a non-ignorable posterior mean (a weighted combination of a zero treatment effect for the current study and a positive treatment effect in the inference prior) which decreases with the sample size of the current study. The required calibrated adjusted alpha (see (8) and (15)) for the posterior probability to control the type I error at the nominal level is from 0.0028 to 0.0042 which is much smaller than 2.5%. The use of this calibrated adjusted alpha results in the power of 0% borrowing.

For the dynamic borrowing, we consider the commensurate prior approach of Hobbs et al. [25,31]. Since the amount of borrowing is not fixed and relies on the consistency of the observed historical and current data, simulation has to be used for generating the results presented in Table 2. As a favorable inferential prior of $\theta\sim N(0.39\ 0.2^2)$ is used, the mean of the posterior distribution combining dynamically the fixed zero treatment effect for the current study and the inconsistent positive inferential prior mean is much larger than zero, with a range from 0.0337 to 0.0962 depending on the sample size of the current study. As a result, dynamic borrowing without the calibration of the critical value still inflates the type I error probability which is in the range of 0.0496–0.0606. Correspondingly, the power depending on the design prior and sample size is higher than the one of the fixed 0% borrowing but lower than the one of the fixed 100% borrowing presented in Table 1. The posterior mean is a dynamic weighted combination of the means of the inferential prior and the design prior of the first column. With the use of the calibrated critical value associated with a calibrated adjusted alpha, which is much lower than the nominal alpha, to control the type I error, the power in general is slightly lower than the probability of success (power of 0% borrowing in the table) presented in Table 1 when $\gamma^2>0$. That is, as anticipated, calibration has the effect of eliminating the very favorable informative inferential prior in order to control the type I error probability.

Sahu and Smith [30] actually use $\theta\sim N(0.12, 0.19^2)$ as the inferential prior in their example. Clearly, the mean of this inferential prior is much lower than those $\theta_a=0.29$, 0.39 and 0.56 in the design priors for the current study. It would be of interest to see the impact of such a low mean in the inferential prior on the characteristics of different settings compared to the results of Tables 1 and 2. Results for the fixed 0% and 100% borrowing are presented in Table 3. The power of the 0% borrowing remains the same while the power for the 100% borrowing is much lower compared to the corresponding one in Table 1. A full 100% borrowing of historical data is associated with a smaller posterior mean or treatment effect, and the type I error rate of 0.0152 for the combined analysis when the sample size for the current study is small is even lower than the nominal level of 0.025. In this case, the power of 0.7466 for the combined analysis is lower than 0.80 and the calibrated adjusted alpha (see (8) and (15)) of 0.0348 for the type I error control can even be higher than 0.025.

For the dynamic borrowing with inferential prior $\theta\sim N(0.12, 0.2^2)$ (Table 4), even though the dynamically combined posterior mean, assumed a 0 treatment effect for the current study, is still slightly larger than 0, the type I error rate (0.0214) could be smaller than 0.025 due to the inclusion of the variability in the inferential prior in the analysis. To make the type I error rate to be close to the nominal level, the calibrated adjusted alpha (0.027) could be larger than 0.025 as in the case of the fixed borrowing.

Sample sizes for achieving 80% power based on (19) for a one-sided test with a significance level of 2.5%, different inferential priors and design priors are provided in Table 5. If the mean of the inferential prior is not large enough relative to the mean of the design prior, borrowing information from the prior may actually increase the required sample size compared to the one of no borrowing.

Table 1
Power and Type I error for a fixed (0% or 100%) borrowing for inferential prior $\theta \sim N(0.39, 0.2^2)$ and different design priors.

Design prior $\theta_d \sim N(b, \gamma^2)$, sample size	Power 0% borrowing	Power Posterior mean (SD) 100% borrowing	Type I error rate Posterior mean (SD) 100% borrowing	Calibrated adjusted alpha
$\theta_d \sim N(0.29,0)$, $n = 376$	0.8028	0.9465 0.3110 (0.0917)	0.1151 0.0819 (0.0917)	0.0042
$\theta_d \sim N(0.39,0)$, $n = 209$	0.8048	0.9628 0.3900 (0.1138)	0.1505 0.1262 (0.1138)	0.0033
$\theta_d \sim N(0.56,0)$, $n = 102$	0.8072	0.9773 0.4758 (0.1407)	0.2040 0.1931 (0.1407)	0.0028
$\theta_d \sim N(0.29,0.025^2)$, $n = 376$	0.7961	0.9414 0.2983 (0.0937)	0.1151 0.0819 (0.0937)	0.0042
$\theta_d \sim N(0.39,0.025^2)$, $n = 209$	0.8011	0.9605 0.3817 (0.1150)	0.1505 0.1262 (0.1150)	0.0033
$\theta_d \sim N(0.56,0.025^2)$, $n = 102$	0.8054	0.9764 0.4714 (0.1413)	0.2040 0.1931 (0.1413)	0.0028
$\theta_d \sim N(0.29,0.05^2)$, $n = 376$	0.7782	0.9266 0.2674 (0.0994)	0.1151 (0.0819 (0.0994)	0.0042
$\theta_d \sim N(0.39,0.05^2)$, $n = 209$	0.7904	0.9534 0.3595 (0.1185)	0.1505 0.1262 (0.1185)	0.0033
$\theta_d \sim N(0.56,0.05^2)$, $n = 102$	0.8000	0.9738 0.4589 (0.1429)	0.2040 0.1931 (0.1429)	0.0028
$\theta_d \sim N(0.29,0.1^2)$, $n = 376$	0.7296	0.8764 0.2000 (0.1167)	0.1151 0.0819 (0.1167)	0.0042
$\theta_d \sim N(0.39,0.1^2)$, $n = 209$	0.7569	0.9259 0.2995 (0.1298)	0.1505 0.1262 (0.1298)	0.0033
$\theta_d \sim N(0.56,0.1^2)$, $n = 102$	0.7807	0.9629 0.4184 (0.1485)	0.2040 0.1931 (0.1485)	0.0028

Table 2
Type I error and Power for a dynamic borrowing for inferential prior $\theta \sim N(0.39, 0.2^2)$ and different design priors.

Design prior $\theta_d \sim N(b, \gamma^2)$, sample size	Calibrated adjusted Power	Power Posterior mean (SD)	Type I error Posterior mean (SD)	Calibrated adjusted alpha
$\theta_d \sim N(0.29, 0)$, $n = 376$	0.7900	0.8646 0.2983 (0.0984)	0.0496 0.0337 (0.0996)	0.012
$\theta_d \sim N(0.39, 0)$, $n = 209$	0.7966	0.8796 0.3908 (0.1276)	0.0500 0.0535 (0.1301)	0.011
$\theta_d \sim N(0.56, 0)$, $n = 102$	0.7896	0.8926 0.5165 (0.1718)	0.0606 0.0962 (0.1760)	0.010
$\theta_d \sim N(0.29, 0.025^2)$, $n = 376$	0.7940	0.871 0.3008 (0.0984)	0.0496 0.0337 (0.0996)	0.012
$\theta_d \sim N(0.39, 0.025^2)$, $n = 209$	0.7944	0.8832 0.3903 (0.1276)	0.0500 0.0535 (0.1301)	0.011
$\theta_d \sim N(0.56, 0.025^2)$, $n = 102$	0.7914	0.8820 0.5183 (0.1719)	0.0606 0.0962 (0.1760)	0.010
$\theta_d \sim N(0.29, 0.05^2)$, $n = 376$	0.7818	0.8490 0.3004 (0.0984)	0.0496 0.0337 (0.0996)	0.012
$\theta_d \sim N(0.39, 0.05^2)$, $n = 209$	0.7942	0.8768 0.3948 (0.1276)	0.0500 0.0535 (0.1301)	0.011
$\theta_d \sim N(0.56, 0.05^2)$, $n = 102$	0.7870	0.8862 0.5177 (0.1719)	0.0606 0.0962 (0.1760)	0.010
$\theta_d \sim N(0.29, 0.1^2)$,$n = 376$	0.7270	0.7952 0.3007 (0.0985)	0.0496 0.0337 (0.0996)	0.012
$\theta_d \sim N(0.39, 0.1^2)$,$n = 209$	0.7510	0.8282 0.391 (0.1278)	0.0500 0.0535 (0.1301)	0.011
$\theta_d \sim N(0.56, 0.1^2)$, $n = 102$	0.7766	0.8754 0.5198 (0.1722)	0.0606 0.0962 (0.1760)	0.010

8. Discussion

As discussed, there are many approaches for performing Bayesian hypothesis testing. Some of them may lack the traditional optimal frequentist operating characteristics. In this paper, we focus on the reconciliation between Bayesian and frequentist hypothesis testing. Rather than controlling the average type I error probability or average weighted combination of type I and type II errors probability, our objective is to control the maximum type I error probability. To harmonize the quantification of treatment effect and hypothesis testing, Bayesian hypothesis testing can be based on the credible probability derived from the posterior distribution which can incorporate historical information through the informative inferential prior.

No matter whether from a hypothesis testing or estimation perspective, a Bayesian analysis is basically a combined analysis. The informative inference prior for the current study could be considered as the posterior distribution of a historical study with a non-informative inferential prior to form the combined analysis. One necessary condition for the justification of the historical data borrowing is consistency of treatment effects between the historical and current studies. Without any general consistency assurance as the a priori before data unblinding, a combined analysis should not be planned and the Bayesian hypothesis testing with the historical data borrowing could be invalid. A discount of the historical data could be reflected in the inferential

Table 3
Power and Type I error for a fixed (0% or 100%) borrowing for inferential prior $\theta \sim N(0.12, 0.2^2)$ and different design priors.

Design prior $\theta_d \sim N(b, \gamma^2)$, sample size	Power 0% borrowing	Power Posterior mean (SD) 100% borrowing	Type I error rate Posterior mean (SD) 100% borrowing	Calibrated adjusted alpha
$\theta_d \sim N(0.29,0)$, $n = 376$	0.8028	0.8201 0.2543 (0.0917)	0.0290 0.0252 (0.0917)	0.0219
$\theta_d \sim N(0.39,0)$, $n = 209$	0.8048	0.8026 0.3026 (0.1138)	0.0245 0.0388 (0.1138)	0.0254
$\theta_d \sim N(0.56,0)$, $n = 102$	0.8072	0.7466 0.3422 (0.1407)	0.0152 0.0594 (0.1407)	0.0348
$\theta_d \sim N(0.29,0.025^2)$, $n = 376$	0.7961	0.8133 0.2416 (0.0937)	0.0290 0.0252 (0.0937)	0.0219
$\theta_d \sim N(0.39,0.025^2)$, $n = 209$	0.8011	0.7988 0.2943 (0.1150)	0.0245 0.0388 (0.1150)	0.0254
$\theta_d \sim N(0.56,0.025^2)$, $n = 102$	0.8054	0.7449 0.3377 (0.1413)	0.0152 0.0594 (0.1413)	0.0348
$\theta_d \sim N(0.29,0.05^2)$, $n = 376$	0.7782	0.7951 0.2107 (0.0994)	0.0290 0.0252 (0.0994)	0.0219
$\theta_d \sim N(0.39,0.05^2)$, $n = 209$	0.7904	0.7882 0.2721 (0.1185)	0.0245 0.0388 (0.1185)	0.0254
$\theta_d \sim N(0.56,0.05^2)$, $n = 102$	0.8000	0.7401 0.3252 (0.1429)	0.0152 0.0594 (0.1429)	0.0348
$\theta_d \sim N(0.29,0.1^2)$, $n = 376$	0.7296	0.7446 0.1433 (0.1167)	0.0290 0.0252 (0.1167)	0.0219
$\theta_d \sim N(0.39,0.1^2)$, $n = 209$	0.7569	0.7548 0.2121 (0.1298)	0.0245 0.0388 (0.1298)	0.0254
$\theta_d \sim N(0.56,0.1^2)$, $n = 102$	0.7807	0.7232 0.2847 (0.1485)	0.0152 0.0594 (0.1485)	0.0348

Table 4
Type I error and power for a dynamic borrowing for inferential prior $\theta \sim N(0.12, 0.2^2)$ and different design priors.

Design prior $\theta_d \sim N(b, \gamma^2)$, sample size	Calibrated adjusted Power	Power Posterior mean (SD)	Type I error Posterior mean (SD)	Calibrated adjusted alpha
$\theta_d \sim N(0.29,0)$, $n = 376$	0.7936	0.8092 0.2758 (0.0986)	0.0280 0.0117 (0.0985)	0.022
$\theta_d \sim N(0.39,0)$, $n = 209$	0.8124	0.8072 0.352 (0.1287)	0.0246 0.0176 (0.1278)	0.026
$\theta_d \sim N(0.56,0)$, $n = 102$	0.7966	0.785 0.4589 (0.1777)	0.0214 0.0313 (0.1714)	0.027
$\theta_d \sim N(0.29,0.025^2)$, $n = 376$	0.7816	0.7984 0.2746 (0.0986)	0.0280 0.0117 (0.0985)	0.022
$\theta_d \sim N(0.39,0.025^2)$, $n = 209$	0.8018	0.7984 0.3504 (0.1288)	0.0246 0.0176 (0.1278)	0.026
$\theta_d \sim N(0.56,0.025^2)$, $n = 102$	0.7886	0.7774 0.4565 (0.1776)	0.0214 0.0313 (0.1714)	0.027
$\theta_d \sim N(0.29,0.05^2)$, $n = 376$	0.7686	0.7832 0.2746 (0.0986)	0.0280 0.0117 (0.0985)	0.022
$\theta_d \sim N(0.39,0.05^2)$, $n = 209$	0.7976	0.794 0.3501 (0.1288)	0.0246 0.0176 (0.1278)	0.026
$\theta_d \sim N(0.56,0.05^2)$, $n = 102$	0.7746	0.7656 0.4565 (0.1777)	0.0214 0.0313 (0.1714)	0.027
$\theta_d \sim N(0.29,0.1^2)$,$n = 376$	0.7242	0.7352 0.2754 (0.0987)	0.0280 0.0117 (0.0985)	0.022
$\theta_d \sim N(0.39,0.1^2)$,$n = 209$	0.7590	0.7536 0.3519 (0.129)	0.0246 0.0176 (0.1278)	0.026
$\theta_d \sim N(0.56,0.1^2)$, $n = 102$	0.7616	0.749 0.4567 (0.1779)	0.0214 0.0313 (0.1714)	0.027

prior with a suitable τ^2 for the effective sample size from the historical source. As perfect consistency is unlikely, a dynamic borrowing may be applied to data analysis so that a further discounting from the inferential prior can be performed based on the level of the heterogeneity between the observed treatment effects of the current study and the inferential prior. Nonetheless, there is no need for the use of the calibrated critical value derived assuming a zero treatment effect for the current study combined with an inconsistent positive treatment effect in the inferential prior. Confirmed by simulation, the application of the calibrated critical value for the dynamic borrowing will eventually eliminate the historical information in the analysis which is equivalent to a fixed 0% borrowing.

A Bayesian analysis is also often applied to borrow historical control data rather than active treatment data for a new analysis. With the rapid evolution of standard of care in some therapeutic areas, the borrowing of only control data may bring more bias as the between treatment difference may have been adjusted for all the confounding effects including the effect of standard of care. That is, the borrowing of the historical control data will at least inherit the same issues discussed above. We need to ensure combinability of the historical and current control data before the Bayesian method is applied. Since there are also frequentist methods through the shrinkage estimators that borrow information from multiple sources and are adjusted for between source variability, the advantage of the Bayesian analysis with the capability of borrowing the historical data may be over stated.

Table 5
Sample sizes for 80% power for a one-sided test at a significance level of 2.5% and different inferential priors.

Design prior $\theta_d \sim N(b, \gamma^2)$	No borrowing	Inferential prior $\theta \sim N(0.39, 0.2^2)$	Inferential prior $\theta \sim N(0.12, 0.2^2)$
$\theta_d \sim N(0.29,0)$	376	152	359
$\theta_d \sim N(0.39,0)$	209	63	209
$\theta_d \sim N(0.56,0)$	102	23	114
$\theta_d \sim N(0.29,0.025^2)$	381	155	365
$\theta_d \sim N(0.39,0.025^2)$	210	64	211
$\theta_d \sim N(0.56,0.025^2)$	102	23	114
$\theta_d \sim N(0.29,0.05^2)$	402	163	383
$\theta_d \sim N(0.39,0.05^2)$	216	65	217
$\theta_d \sim N(0.56,0.05^2)$	104	23	116
$\theta_d \sim N(0.29,0.1^2)$	494	202	464
$\theta_d \sim N(0.39,0.1^2)$	242	72	240
$\theta_d \sim N(0.56,0.1^2)$	109	24	121

Acknowledgements

We would like to thank Dr. Louise Traylor, the three referees and the associate editor for their helpful comments and suggestions that have improved the presentation of this paper.

References

[1] R. Weiss, Bayesian sample size calculations for hypothesis testing, J. R. Stat. Soc. Ser. D Stat. 46 (1997) 185–191.
[2] L.Y.T. Inoue, D.A. Berry, G. Parmigiani, Relationship between Bayesian and frequentist sample size determiniation, Am. Stat. 59 (2005) 79–87.
[3] G. Casella, R.L. Berger, Reconciling Bayesian and frequentist evidence in the one-sided testing problem, J. Am. Stat. Assoc. 82 (1987) 106–111.
[4] A. Gelman, J. Hill, M. Yajima, Why we (usually) don't have to worry about multiple comparisons, J. Res. Educ. Eff. 5 (2012) 189–211.
[5] T. Sellke, M.J. Bayarri, J.O. Berger, Calibration of p values for testing precise null hypotheses, Am. Stat. 55 (2001) 62–71.
[6] M. Bogdan, J.K. Ghosh, S.T. Tokdar, A comparison of the Benjamin-Hochberg procedure with some Bayesian rules for multiple testing, IMS Collections: Beyond Parameters in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen, 1 2008, pp. 211–230.
[7] H. Pezeshk, Bayesian techniques for sample size determination in clinical trials: a short review, Stat. Methods Med. Res. 12 (2003) 489–504.
[8] F. Sadia, S.S. Hossain, Contrast of Bayesian and classical sample size determination, J. Mod. Appl. Stat. Methods 13 (2014) 420–431.
[9] Y. Ji, Y. Lu, G.B. Mills, Bayesian models based on test statistics for multiple hypothesis testing problems, Bioinformatics. 24 (2008) 943–949.
[10] US FDA, Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials, Available at https://www.fda.gov/MedicalDevices/ucm071072.htm, (February 2010) accessed date: April 5, 2019.
[11] J. Cao, J.J. Lee, S. Alber, Comparison of Bayesian sample size criteria: ACC, ALC and WOC, J. Stat. Plan Infer. 139 (2009) 4111–4122.
[12] C.J. Adcock, A Bayesian approach to calculating sample sizes, The Statistician 37 (1988) 433–439.
[13] L. Joseph, P. Belisle, Bayesian sample size determination for normal mean and differences between normal means, The Statistician 46 (1997) 209–266.
[14] S.N. Goodman, Toward evidence-based medical statistics. 1: the p value fallacy, Ann. Intern. Med. 130 (1999) 995–1013.
[15] S.N. Goodman, Toward evidence-based medical statistics. 2: the Bayes factor, Ann. Intern. Med. 130 (1999) 1019–1021.
[16] J.O. Berger, T. Sellke, Testing a point null hypothesis: the irreconcilability of P values and evidence, J. Am. Stat. Assoc. 82 (1987) 112–122.
[17] E.M. Reyes, S.K. Ghosh, Bayesian average error based approach to sample size calculations for hypothesis testing, J. Biopharm. Stat. 23 (2013) 569–588.
[18] M.A. Psioda, J.G. Ibrahim, Bayesian clinical trial design using historical data that inform the treatment effect, Biostatistics (2018), https://doi.org/10.1093/biostatistics/kxy009.
[19] J. Berger, Lecture 2: Bayesian hypothesis testing, CBMS Conference on Model Uncertainty and Multiplicity, July 23–28, 2012 Available at https://cbms-mum.soe.ucsc.edu/lecture2.pdf (Accessed data: September 25, 2018).
[20] G. Pennello, L. Thompson, Experience with reviewing Bayesian medical device trials, J. Biopharm. Stat. 18 (2008) 81–115.
[21] S. Pocock, The combination of randomized and historical controls in clinical trials, J. Chronic Dis. 29 (1976) 175–188.
[22] M. Guo, D. Heitjan, Multiplicity-calibrated Bayesian hypothesis tests, Biostatistics 11 (2010) 473–483.
[23] J.O. Berger, B. Boukai, Y. Wang, Unified frequentist and Bayesian testing of a precise hypothesis, Stat. Sci. 12 (1997) 133–160.
[24] F. Wang, A.E. Gelfand, A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models, Stat. Sci. 17 (2002) 193–208.
[25] B.P. Hobbs, B.P. Carlin, S.J. Mandrekar, D.J. Sargent, Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials, Biometrics 67 (2011) 1047–1056.
[26] Y. Duan, K. Ye, E.P. Smith, Evaluating water quality using power priors to incorporate historical information, Environmetrics 17 (2006) 95–106.
[27] R. Simon, Optimal two-stage designs for phase II clinical trials, Control. Clin. Trials 10 (1989) 1–10.
[28] O. Sverdlov, Y. Ryeznik, S. Wu, Exact Bayesian inference comparing binomial proportions, with application to proof-of concept clinical trials, Ther. Innov. Regul. Sci. 49 (2015) 163–174.
[29] C. Jennison, B.W. Turnbull, Statistical approaches to interim monitoring of medical trials: a review and commentary, Stat. Sci. 5 (1990) 299–317.
[30] S.K. Sahu, T.M.F. Smith, A Bayesian method of sample size determination with practical applications, J. R. Statist. Soc. A. 169 (2006) 235–253.
[31] B.P. Hobbs, D.J. Sargent, B.P. Carlin, Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models, Bayesian Anal. 7 (3) (2012) 639–674.