Published in final edited form as:

Stat Biosci. 2018 August ; 10(2): 439–459. doi:10.1007/s12561-017-9200-5.

Bayesian Design of Non-Inferiority Clinical Trials via the Bayes Factor

Wenqing Li*, Ming-Hui Chen[†], Xiaojing Wangy[†], and Dipak K. Dey[†]

*Ventana Medical Systems, Inc., 1910 East Innovation Park Drive, Tucson, AZ 85755, USA

[†]Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, Connecticut 06269, USA

Abstract

We developed a Bayes factor based approach for the design of non-inferiority clinical trials with a focus on controlling type I error and power. Historical data are incorporated in the Bayesian design via the power prior discussed in Ibrahim and Chen (2000). The properties of the proposed method are examined in detail. An efficient simulation-based computational algorithm is developed to calculate the Bayes factor, type I error and power. The proposed methodology is applied to the design of a non-inferiority medical device clinical trial.

Keywords

Fitting prior; Power prior; Sampling prior; Sample size determination; Type I error

1 Introduction

Clinical trials are clinical investigations that examine and evaluate safety and efficacy of a drug used in human subjects. In general, it is more difficult to demonstrate superiority of a study drug to the standard of care (SOC) in an active-controlled clinical trial. A non-inferiority trial is often used to demonstrate efficacy of a new treatment (medication or medical device) relative to the known effective SOC. By now, there is substantial literature on the frequentist design of non-inferiority trials, including D'Agostino Sr. et al. (2003), Chow and Liu (2004), Kieser and Friede (2007), Kang and Tsong (2010), and Liu and Chang (2011).

Recently, the Bayesian methodology has become increasingly popular in conducting clinical trials, particularly in the early phase clinical trials and medical device studies. There has been a rich literature on Bayesian sample size determination (SSD), including, for example, Adcock (1997), Joseph and Belisle (1997), Wang and Gelfand (2002), De Santis (2007), M'Lan et al. (2006, 2008), Campbell (2011), and Chen et al. (2011). Unlike the use of the type I and type II errors in the frequentist SSD approach, most of the Bayesian SSD approaches are based on various performance criteria, including the average posterior variance criterion (Wang and Gelfand, 2002), the average coverage criterion, the average length criterion, or the worst outcome criterion (Joseph et al., 1995, 1997; M'Lan et al., 2006, 2008), and the Bayes factor (Weiss, 1997).

In 2010, the FDA released guidance for the use of Bayesian ideas in medical device clinical trials. The document titled “Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials” provides guidance on the evaluation of operating characteristics of a Bayesian clinical trial design, specifically on the evaluation of type I error (the probability of erroneously approving an ineffective or unsafe device) and type II error (the probability of erroneously disapproving a safe and effective device). Often there is no appropriate type I error or type II error defined in Bayesian SSD approaches except as noted in the following literature. Campbell (2011) provided a thorough update of the recent development of Bayesian statistics in medical devices. Chen et al. (2011) developed a new Bayesian design methodology with a focus on controlling type I error and power for non-inferiority trials. Ibrahim et al. (2012) developed a Bayesian meta-experimental design for non-inferiority trials using the partial borrowing power prior. Chen et al. (2014a) developed a Bayesian sequential meta-analysis design for a new antidiabetic drug development program. Chen et al. (2014b) extended the methodology of Chen et al. (2011) to the design of clinical trials for recurrent events data. More recently, Ibrahim et al. (2015b) developed the SAS Macro BSMED for Bayesian survival meta-experimental design using historical data. Gamalo-Siebers et al. (2015) provided a comprehensive overview of both frequentist and Bayesian approaches in non-inferiority trials.

Weiss (1997) proposed a Bayes factor-based approach, in which the cut-off value for the Bayes factor is either a fixed constant or is determined by using the fitting prior. However, this approach may not always guarantee nominal type I error, especially when future data does not support the fitting prior. Other approaches using the calibrated Bayes factor include Garcia-Donato and Chen (2005), Weiss and Wang (2010), and Yu et al. (2008). Osman and Ghosh (2011) developed a semiparametric testing approach based on the Bayes factor for non-inferiority trials with binary endpoints. Reyes and Ghosh (2013) proposed a framework for hypothesis testing and sample size determination using Bayesian average errors. In this paper, we develop a Bayes factor-based approach for the design of non-inferiority clinical trials where the sampling prior is used to calibrate the Bayes factor and the historical data are incorporated via the power prior of Ibrahim and Chen (2000) and Ibrahim et al. (2015a). The proposed approach has a better control of type I error and leads to a potential increase in statistical power and sample size reduction. In addition, we propose an efficient simulation-based computational algorithm via the dropping-needle method discussed in Hoijtink et al. (2008) to calculate the Bayes factor, type I error and power.

The rest of the paper is organized as follows. In Section 2, we briefly discuss the traditional Bayesian hypothesis testing approach via the Bayes factor with a fixed or flexible critical Bayes factor value and demonstrate the detailed development of the proposed Bayesian approach, and provide an example. In Section 3, computational algorithms are developed for computing the Bayes factor as well as type I error and power. In Section 4, the proposed Bayesian approach is applied to a two-arm non-inferiority clinical trial with a binary primary endpoint and a continuous secondary endpoint. Finally, the paper is concluded with a brief discussion in Section 5.

2 The Bayesian Method

2.1 Bayesian Hypothesis Testing via the Bayes Factor

One of the Bayesian hypothesis testing approaches is based on the statistics of the Bayes factor and summarizes the evidence provided by the data in favor of one model or theory represented by a statistical hypothesis, such as the null hypothesis versus the alternative hypothesis. The decision making is based on the value of the Bayes factor (B) from data in comparison to a pre-specified constant (B_0). Jeffreys (1961) gave the following scale for the interpretation of the Bayes factor: there is not worth more than a bare mention, substantial, strong, or decisive evidence against H_0 if $\log_{10}(B)$ is in $(0, 1/2]$, $(1/2, 1]$, $(1, 2]$, or $(2, \infty)$, respectively. A modified version of the above interpretation is given in Kass and Raftery (1995) and is based on the natural logarithm scale: there is not worth more than a bare mention, positive, strong, or very strong evidence against H_0 if $2 \log_e(B)$ is in $(0, 2]$, $(2, 6]$, $(6, 10]$, or $(10, \infty)$, respectively. As shown in the subsequent sections, type I error is not guaranteed when we make a decision by simply comparing the value of the Bayes factor B to B_0 .

Weiss (1997) proposed a Bayesian approach via the Bayes factor with a flexible cut-off value to control type I error. However, the same fitting prior is used in computing both the Bayes factor and the prior predictive probabilities for decision making. Due to the double use of the same fitting prior, the type I error is not fully controlled, especially when future data is not from this fitting prior.

2.2 The Proposed Bayes Factor Approach

Let θ denote the parameters of interest and also let ξ be the nuisance parameters, where θ and ξ are assumed to be distinct. The general hypotheses can be formulated as follows: $H_0: (\theta, \xi) \in \Theta_0$ versus $H_1: (\theta, \xi) \in \Theta_1$, where Θ_0 and Θ_1 are disjoint parameter subspaces, although often the hypothesis testing is on the parameters of interest, θ , only. A simple example of the hypotheses is $H_0: \theta = \theta_0$ versus $H_1: \theta = \theta_1$, where $\Theta_i = \{\theta: \theta = \theta_i\}$, $i = 0, 1$.

Let $\pi_i = P(\Theta_i)$, $i = 0, 1$, denote the prior probabilities of the parameter spaces under the null hypothesis and the alternative hypothesis, respectively. Denote the likelihood function of the data by $L(\theta, \xi|D) \equiv f(\mathbf{y}^{(n)}|\theta, \xi)$. Here we use D to denote the observed data $\mathbf{y}^{(n)}$, which is a function of sample size n . Let $\alpha_i = P(\Theta_i|D)$, $i = 0, 1$, be the corresponding posterior probabilities of the parameter spaces for a given prior and the data D . The Bayes factor $B(D)$ for H_0 against H_1 given the data D is defined as the ratio of α_0/α_1 and π_0/π_1 , that is,

$$B(D) = \frac{\alpha_0/\alpha_1}{\pi_0/\pi_1}.$$

A proper prior needs to be specified in the calculation of the posterior probabilities of α_i . The prior used for the analysis is called the fitting prior, which is the same concept used in Wang and Gelfand (2002). Following Berger (1980), we specify the fitting prior as

$$\pi^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \begin{cases} \pi_0 \pi_0^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) & \text{if } (\boldsymbol{\theta}, \boldsymbol{\xi}) \in \Theta_0, \\ \pi_1 \pi_1^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) & \text{if } (\boldsymbol{\theta}, \boldsymbol{\xi}) \in \Theta_1, \end{cases} \quad (2.1)$$

where $\pi_i^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi})$, $i = 0, 1$, are the joint fitting priors for $(\boldsymbol{\theta}, \boldsymbol{\xi})$ under the null and alternative hypotheses, respectively. Define the marginal distributions of the data under H_0 and H_1 , respectively, as

$$m_i^{(f)}(D) \equiv \int_{\Theta_i} \pi_i^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) L(\boldsymbol{\theta}, \boldsymbol{\xi} | D) d\boldsymbol{\theta} d\boldsymbol{\xi}, \quad i = 0, 1.$$

The Bayes factor can be re-written as

$$B(D) = \frac{m_0^{(f)}(D)}{m_1^{(f)}(D)} = \frac{\int_{\Theta_0} \pi_0^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) L(\boldsymbol{\theta}, \boldsymbol{\xi} | D) d\boldsymbol{\theta} d\boldsymbol{\xi}}{\int_{\Theta_1} \pi_1^{(f)}(\boldsymbol{\theta}, \boldsymbol{\xi}) L(\boldsymbol{\theta}, \boldsymbol{\xi} | D) d\boldsymbol{\theta} d\boldsymbol{\xi}}. \quad (2.2)$$

We define a decision rule based on the Bayes factor $B(D)$ to reject H_0 as

$$\psi(D) = \begin{cases} 1 & \text{if } B(D) < B_0, \\ 0 & \text{if } B(D) \geq B_0, \end{cases} \quad (2.3)$$

where B_0 is the critical value for the Bayes factor, which only depends on the study design, not on the data D .

In addition to the fitting prior, we adopt another prior to carry out the Bayes factor calibration. This prior is called the sampling prior. The sampling prior is used to generate the data from the prior predictive distribution under the null hypothesis or under the alternative hypothesis. Let $\pi_i^{(s)}(\boldsymbol{\theta}, \boldsymbol{\xi})$, $i = 0, 1$, be the joint sampling priors for $(\boldsymbol{\theta}, \boldsymbol{\xi})$ under H_0 and H_1 , respectively. Similar to Chen et al. (2011), we define the following key quantity:

$$\beta_s^{(n)} = E_s[\psi(D)], \quad (2.4)$$

where the expectation is taken with respect to the marginal distribution of D under the sampling prior $\pi^{(s)}(\boldsymbol{\theta}, \boldsymbol{\xi})$. Therefore, the Bayesian type I error and power, denoted by $\beta_{s,i}^{(n)}$, $i = 0, 1$, respectively, are the corresponding $\beta_s^{(n)}$ values calculated based on the marginal distributions of the data under $\pi_0^{(s)}(\boldsymbol{\theta}, \boldsymbol{\xi})$ and $\pi_1^{(s)}(\boldsymbol{\theta}, \boldsymbol{\xi})$, respectively.

Denote the marginal distribution of the data corresponding to $\pi_i^{(s)}(\boldsymbol{\theta}, \boldsymbol{\xi})$ by

$$m_i^{(s)}(D) \equiv \int_{\Theta_i} \pi_i^{(s)}(\theta, \xi) L(\theta, \xi | D) d\theta d\xi, \quad i = 0, 1.$$

For a given nominal type I error rate $\alpha > 0$, we specify the critical value B_0 such that

$$\alpha = \int \psi(D) m_0^{(s)}(D) dD. \quad (2.5)$$

Then, the Bayesian type I error, i.e., the probability to reject H_0 when H_0 is true, can be controlled to be α by observing that $\beta_{s,0}^{(n)}$ is essentially equal to α as given in (2.5).

To determine the Bayesian sample size, for given $\alpha > 0$ and $\beta_0 > 0$, we compute

$$n_\alpha = \min\{n: \beta_{s,0}^{(n)} \leq \alpha\}; \quad n_{\beta_0} = \min\{n: \beta_{s,1}^{(n)} \geq \beta_0\}. \quad (2.6)$$

Then, the Bayesian sample size is given by $n = \max\{n_\alpha, n_{\beta_0}\}$. Common choices of α and β_0 include $\alpha = 0.05$ and $\beta_0 = 0.8$ so that the type I error rate is less than or equal to 0.05 and the power is at least 0.8.

Remark 1—Since the sampling priors, $\pi_i^{(s)}(\theta, \xi)$'s, are often determined by the study design, they may take the form of $\pi_i^{(s)}(\theta)\pi_i^{(s)}(\xi)$, where $\pi_i^{(s)}(\theta)$ and $\pi_i^{(s)}(\xi)$ are independent point mass distributions for θ and ξ , respectively. When H_0 is composite or there are nuisance parameters ξ , $\pi_0^{(s)}(\theta, \xi)$ used at the design stage may not be the sampling prior from which the data are generated once the trial is completed. In this case, the type I error is not controlled even if B_0 is calculated via (2.5). We will discuss how to determine B_0 under a special sampling prior so that the type I error is controlled when the true sampling prior is different than the one used in computing B_0 in Sections 2.3.2, 4.1.2, and 4.2.2.

Remark 2—To carry out Bayesian inference of model parameters, the fitting priors $\pi_i^{(f)}(\theta, \xi)$ may be non-informative or even more non-informative than the sampling priors. However, the fitting priors must be proper for the Bayes factor-based approach since the Bayes factor is not defined when the priors are improper.

In the next section, we use a simple case to show how our proposed Bayesian method has desirable properties not possessed by other approaches.

2.3 Applications to One-Arm Trial with the Normal Model

We consider testing the mean of a univariate normal distribution with a known variance. Let $\{y_j, j = 1, 2, \dots, n\}$ be a random sample from a $N(\theta, \sigma^2)$ distribution, where θ is the unknown parameter of interest, which needs to be estimated/tested, and the variance σ^2 is a

known constant. Then, $\bar{y} = (1/n) \sum_{j=1}^n y_j$ is a sufficient and complete statistic for the parameter θ .

2.3.1 Two-Sided Hypothesis Testing—Consider a simple null hypothesis and a two-sided alternative hypothesis on the mean parameter θ . That is, the null and alternative hypotheses are defined as $H_0: \theta = \theta_0$ and $H_1: \theta \neq \theta_0$. At the design stage, let the sampling priors for θ be

$$\pi_0^{(s)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\pi_1^{(s)}(\theta) \sim N(\theta_1, \tau_{\theta_1}^2) \text{ if } \theta \in \Theta_1: \theta \neq \theta_0,$$

where θ_1 and $\tau_{\theta_1}^2$ are the known prior mean and variance. The distance between θ_0 and θ_1 is often called the effective treatment size and denoted by $\delta = \theta_0 - \theta_1 > 0$ (assuming $\theta_1 < \theta_0$). Furthermore, let the fitting priors be

$$\pi_0^{(f)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_0, \\ 0 & \text{otherwise,} \end{cases}$$

$$\pi_1^{(f)}(\theta) \sim N(\mu, \tau^2) \text{ if } \theta \in \Theta_1: \theta \neq \theta_0,$$

where μ and τ^2 are the known prior mean and variance. Usually μ needs to be appropriately chosen to reflect θ_0 , but not necessarily.

After some algebra, the Bayes factor of H_0 versus H_1 can be written as

$$B(D) = \left(\frac{n\tau^2 + \sigma^2}{\sigma^2} \right)^{1/2} \exp \left\{ -\frac{1}{2} \left(c - \frac{b^2}{a} \right) \right\} \exp \left\{ -\frac{1}{2} a \left(\bar{y} - \frac{b}{a} \right)^2 \right\}, \quad (2.7)$$

where, $a = \frac{n}{\sigma^2} - \frac{n}{n\tau^2 + \sigma^2} > 0$, $b = \frac{n}{\sigma^2} \theta_0 - \frac{n}{n\tau^2 + \sigma^2} \mu$ and $c = \frac{n}{\sigma^2} \theta_0^2 - \frac{n}{n\tau^2 + \sigma^2} \mu^2$.

Remark 3: For decision making, it is equivalent to apply a monotonic transformation to the Bayes factor and its critical value. For example, after the log transformation, (2.7) can be rewritten as

$$\log B(D) = \frac{1}{2} \log \frac{n\tau^2 + \sigma^2}{\sigma^2} - \frac{1}{2} \left(c - \frac{b^2}{a} \right) - \frac{1}{2} a \left(\bar{y} - \frac{b}{a} \right)^2. \quad (2.8)$$

We also have

$$\log B_0 = \frac{1}{2} \log \frac{n\tau^2 + \sigma^2}{\sigma^2} - \frac{1}{2} \left(c - \frac{b^2}{a} \right) - \frac{\sigma^2}{2n} a \chi_{1-\alpha, 1, \lambda_1}^2, \quad (2.9)$$

where $\chi_{1-\alpha, 1, \lambda_1}^2$ is the $(1 - \alpha)$ th upper tail quantile of a non-central chi-square distribution with 1 degree of freedom and a non-centrality parameter of $\lambda_1 = (\theta_0 - b/a)^2/(\sigma^2/n)$. For a given B_0 , the null hypothesis is rejected if $\log B(D) < \log B_0$ for the given data D . Power can be computed as follows

$$\beta^{(n)} = P(\log B(D) < \log B_0 | H_1, \pi_1^{(s)}) = P\left(\chi_{1, \lambda_2}^2 > \frac{\sigma^2}{\sigma^2 + \tau_{\theta_1}^2} \chi_{1-\alpha, 1, \lambda_1}^2\right),$$

where χ_{1, λ_2}^2 is a non-central chi-square distribution with 1 degree of freedom and a non-centrality parameter of $\lambda_2 = (\theta_1 - b/a)^2/(\{\sigma^2 + \tau_{\theta_1}^2\}/n)$. Using (2.6), for a given $\beta_0 > 0$, the Bayesian sample size is calculated by $\min\{n : \beta^{(n)} \geq \beta_0\}$.

Remark 4: The decision rule in (2.3) is equivalent to the following rule:

$$\psi(D) = \begin{cases} 1 & \text{if } |\bar{y} - b/a| > k_0 > 0, \\ 0 & \text{if } |\bar{y} - b/a| \leq k_0, \end{cases} \quad (2.10)$$

where k_0 is the critical value. That is, the determination of k_0 in the above decision rule is equivalent to the determination of B_0 in the decision rule of (2.7).

After some algebra, we can show that k_0 can be computed by solving the following equation:

$$\alpha = 1 - \Phi\left(\frac{\frac{\sigma^2/n}{\tau^2}(\theta_0 - \mu) + k_0}{\sigma n^{-1/2}}\right) + \Phi\left(\frac{\frac{\sigma^2/n}{\tau^2}(\theta_0 - \mu) - k_0}{\sigma n^{-1/2}}\right), \quad (2.11)$$

where $\Phi(\cdot)$ is the cumulative probability function of a standard normal distribution. Consequently, the critical value B_0 can be calculated by

$$B_0 = \frac{n\tau^2 + \sigma^2}{\sigma^2} \exp\left\{-\frac{1}{2}\left(c - \frac{b^2}{a}\right)\right\} \exp\left\{-\frac{ak_0^2}{2}\right\}.$$

Result 1: There exists a unique value of k_0 such that the equation (2.11) satisfies, which guarantees the existence of the critical value k_0 , hence B_0 .

Proof: Write

$$g(k_0) = 1 - \Phi\left(\frac{\frac{\sigma^2/n}{\tau^2}(\theta_0 - \mu) + k_0}{\sigma n^{-1/2}}\right) + \Phi\left(\frac{\frac{\sigma^2/n}{\tau^2}(\theta_0 - \mu) - k_0}{\sigma n^{-1/2}}\right).$$

It is easy to show that $g(k_0)$ is continuous and strictly decreasing in $k_0 \in (0, \infty)$, $g_{k_0 \rightarrow 0}(k_0) = 1$, and $g_{k_0 \rightarrow \infty}(k_0) = 0$. Therefore, the result follows.

Remark 5: We consider two special cases: (i) a non-informative prior for $\pi_1^{(f)}(\theta)$ (i.e., $\tau^2 \rightarrow \infty$) and (ii) $\mu = \theta_0$. If $\tau^2 \rightarrow \infty$ or if $\mu = \theta_0$, equation (2.11) reduces to $\alpha = 1 - \Phi(k_0 n^{1/2}/\sigma) + \Phi(-k_0 n^{1/2}/\sigma)$. Hence, $k_0 = z_{1-\alpha/2} \sigma/n^{1/2}$, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ th quantile of the standard normal $N(0, 1)$ distribution.

For a point mass sampling prior under H_1 (i.e., $\pi_1^{(s)}(\theta) = \Delta_{\{\theta = \theta_1\}}$ or $\tau_{\theta_1}^2 = 0$), we have $\beta = 1 - \Phi([n^{1/2}/\sigma](\delta + k_0)) + \Phi([n^{1/2}/\sigma](\delta - k_0))$ if $\tau^2 \rightarrow \infty$. Then, β can be approximated by $\Phi([n^{1/2}/\sigma](\delta - k_0))$ for a small type I error α and a large effective treatment difference δ . Therefore, the optimal sample size is given by $n = (\sigma/\delta)^2(z_{1-\alpha/2} + z_\beta)^2$, which is same as the sample size formula in the frequentist design for a given type I error α and power β .

We note that the critical value $B_0 \rightarrow \infty$ when $\tau^2 \rightarrow \infty$. In this case, (2.3) is not well defined if we compute $B(D)$ and B_0 separately. However, $\lim_{\tau^2 \rightarrow \infty} B(D)/B_0$ may still exist. For this simple case, by comparing $\log B(D)$ in (2.8) and $\log B_0$ in (2.9), $B(D) < B_0$ in (2.3) is equivalent to $(\bar{y} - b/a)^2 > (\sigma^2/n)\chi_{1-\alpha, 1, \lambda_1}^2$, whose limit is $(\bar{y} - \theta_0)^2 > (\sigma^2/n)\chi_{1-\alpha, 1}^2$ when $\tau^2 \rightarrow \infty$. Thus, this limit decision rule is precisely the same as the decision rule under the frequentist approach, which confirms that the frequentist approach is essentially the same as the proposed Bayesian approach when a non-informative prior is specified for $\pi_1^{(f)}$.

Moreover, in this limiting scenario, type I error is always controlled for a given sample size and a known population variance, and it does not depend on the hyper-parameters in the fitting prior.

Remark 6: Assuming $\theta_0 = \mu = 0$, we carried out simulations for the scenarios with $\theta_1 = -1$, $\sigma = 1.5$, $\tau = 3$ and $n = 20$. Type I error under the Bayesian hypothesis testing approach by comparing the Bayes factor with a fixed cut-off $B_0 = 2.718$ is 0.12 while type I error under the proposed Bayesian approach is 0.05. These results indicate that type I error is not controlled under the Bayesian hypothesis testing approach by comparing the value of the Bayes factor with a fixed cut-off B_0 , but it is controlled under the proposed approach.

2.3.2 Non-inferiority Hypothesis Testing—Now consider the non-inferiority hypotheses of the form: $H_0: \theta \geq \delta$ versus $H_1: \theta < \delta$, where $\delta(> 0)$ is the non-inferiority

margin. We assume encompassing fitting priors (Klugkist et al., 2005) for θ . Let $\pi^{(f)}(\theta)$ be the density of $N(\mu, \tau^2)$, where μ and τ^2 are the known mean and variance constants, respectively (assuming $\tau > 0$). We specify the fitting priors to be

$$\pi_i^{(f)}(\theta) = \frac{\pi^{(f)}(\theta)}{\int_{\Theta_i} \pi^{(f)}(\theta) d\theta} \mathbf{1}(\theta \in \Theta_i), \quad i = 0, 1.$$

After some algebra, we obtain

$$B(D) = \frac{1 - \Phi((\delta - b/a)/a^{-1/2})}{\Phi((\delta - b/a)/a^{-1/2})} \cdot \frac{\Phi((\delta - \mu)/\tau)}{1 - \Phi((\delta - \mu)/\tau)}, \quad (2.12)$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal $N(0, 1)$ distribution, $a = 1/\tau^2 + n/\sigma^2$, and $b = \mu/\tau^2 + n\bar{y}/\sigma^2$.

We assume the point mass sampling priors for θ , which are given by

$$\pi_i^{(s)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_i \\ 0 & \text{otherwise} \end{cases}, \quad i = 0, 1.$$

Let $\theta_1 = \theta_0$, and $\delta = \theta_0 - \theta_1$ be the non-inferiority margin. We carried out simulations to investigate the effect of the sampling prior parameter θ_1 on power and the characteristics of the proposed Bayesian approach. For $\theta_0 = 1$, $\sigma = 1.5$, $\mu = 1$, $\tau = 1.5$, $\delta = 1$, and $n = 15$, power as a function of θ_1 is plotted in Figure 1, which shows that power is nondecreasing as θ_1 decreases. Furthermore, Figure 1 shows that power for hypothesis testing is maximized at $\theta_1 = \theta_0 = \delta$ for $\theta_1 \geq \delta$, and is minimized at $\theta_1 = 0 = \theta_0 - \delta$ for $\theta_1 < 0$. This implies that for non-inferiority hypothesis testing, it is sufficient to compute the type I error at $\theta_1 = \theta_0 = \delta$.

Note that the second part of $B(D)$ (i.e., $\Phi((\delta - \mu)/\tau)/(1 - \Phi((\delta - \mu)/\tau))$) does not depend on the data. Therefore, computing the first part of $B(D)$ (i.e., $(1 - \Phi((\delta - b/a)/a^{-1/2}))/\Phi((\delta - b/a)/a^{-1/2})$) is sufficient for the computation and decision making because the decision rule in (2.3) is based on the relative magnitude of $B(D)$ and B_0 .

Remark 7: Under the encompassing fitting priors, we can show that both b and $B(D)$ in (2.12) are monotonic functions of $\bar{y}$. Therefore, the decision based on $B(D)$ is equivalent to the decision based on the sufficient statistic $\bar{y}$. In addition, power is independent of the hyperparameters of the fitting priors.

Remark 8: Consider a point mass fitting prior for $\pi_0^{(f)}(\theta)$ and a normal fitting prior for $\pi_1^{(f)}(\theta)$. Specifically, we take

$$\pi_0^{(f)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_f \geq \delta, \\ 0 & \text{otherwise,} \end{cases}$$

$$\pi_1^{(f)}(\theta) \sim \frac{1}{\Phi((\delta - \mu)/\tau)} N(\mu, \tau^2) \text{ if } \theta \in \Theta_1: \theta < \delta,$$

where μ and τ^2 are the known mean and variance constants, respectively. After some algebra, it can be shown that

$$\log(B(D)) = \log(\Phi((\delta - \mu)/\tau)\tau a^{-1/2}) + \left[\frac{c - b^2/a}{2} - \frac{(\bar{y} - \theta_f)^2}{2(\sigma^2/n)} - \log(1 - \Phi((\delta - \mu)/\tau)) \right],$$

where $a = 1/\tau^2 + n/\sigma^2$, $b = \mu/\tau^2 + n\bar{y}/\sigma^2$, and $c = \mu^2/\tau^2 + n\bar{y}^2/\sigma^2$. In this case, $B(D)$ is not a monotonic function of $\bar{y}$ in general. Thus, the decision based on $B(D)$ is not equivalent to the one based on $\bar{y}$ and power may depend on the hyperparameters of the fitting priors.

Remark 9: Suppose that we assume point mass fitting priors for both $\pi_0^{(f)}(\theta)$ and $\pi_1^{(f)}(\theta)$ as

$$\pi_0^{(f)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_{0f} \\ 0 & \text{otherwise} \end{cases}; \quad \pi_1^{(f)}(\theta) = \begin{cases} 1 & \text{if } \theta = \theta_{1f} \\ 0 & \text{otherwise} \end{cases}.$$

Then, we have $\log B(D) = n[\theta_{1f}^2 - \theta_{0f}^2]/(2\sigma^2) - n\bar{y}(\theta_{1f} - \theta_{0f})/\sigma^2$, which is a monotonic function of $\bar{y}$. Thus, the decision based on $B(D)$ is equivalent to the decision based on $\bar{y}$.

3 Computational Development

In order to determine Bayesian sample size, we need to compute type I error and power. Due to the complexity of the joint posterior distribution, an analytical evaluation such as the derivation of a closed form expression for the marginal posterior distribution does not appear possible. Hence, we adopt a Markov chain Monte Carlo (MCMC) sampling based method and develop a computationally efficient simulation algorithm to carry out the posterior computation.

3.1 Computational Algorithm for Type I Error and Power

We propose a simulation algorithm for the general hypothesis testing: $H_0 : (\theta, \xi) \in \Theta_0$ versus $H_1 : (\theta, \xi) \in \Theta_1$ as follows. The critical value B_0 of the Bayes factor is computed using the following algorithm:

Step 1. Generate $(\theta_i, \xi_i) \sim \pi_0^{(s)}(\theta, \xi)$.

Step 2. Generate $\mathbf{y}_i^{(n)} \sim f(\mathbf{y}^{(n)} | \theta_i, \xi_i)$.

Step 3. Calculate the Bayes factor $B(\mathbf{y}_i^{(n)})$.

Step 4. Repeat Steps 1–3 N times, and set B_0 to be the α th quantile of the $B(\mathbf{y}_i^{(n)})$.

Given the critical value B_0 , power is computed as follows:

Step 1. Generate $(\theta_i, \xi_i) \sim \pi_1^{(s)}(\theta, \xi)$.

Step 2. Generate $\mathbf{y}_i^{(n)} \sim f(\mathbf{y}^{(n)} | \theta_i, \xi_i)$.

Step 3. Calculate the Bayes factor $B(\mathbf{y}_i^{(n)})$ and check whether $B(\mathbf{y}_i^{(n)}) \leq B_0$ or not.

Step 4. Repeat Steps 1–3 N times. Then an estimate of power (i.e., $\beta_{s,1}^{(n)}$) is given by $(1/N) \sum_{i=1}^N \mathbf{1}\{B(\mathbf{y}_i^{(n)}) \leq B_0\}$.

This algorithm can be used for computing type I error if $(\theta_i, \xi_i) \sim \pi_0^{(s)}(\theta, \xi)$ is used in Step 1.

3.2 Computation of the Bayes Factor

The key step in type I error and power computation is to calculate the Bayes factor $B(D)$ for given data D . There are several Monte Carlo (MC) methods recently developed to calculate B . DiCicco et al. (1997) reviewed several MC methods of estimating the Bayes factor, including a simulated version of Laplace's method, a simulated version of the Bartlett correction, importance sampling, etc. Following Hoijtink et al. (2008), we propose a simple but efficient MC method to calculate the Bayes factor for non-inferiority clinical trials using the MC samples generated from the joint posterior distribution via the MCMC sampling algorithm or the Gibbs sampling algorithm.

Proposition 3.1—For the non-inferiority hypothesis testing scenario with the encompassing fitting prior, the Bayes factor can be written as

$$B(D) = \frac{E[\mathbf{1}\{(\theta, \xi) \in \Theta_0\} | D]}{E[\mathbf{1}\{(\theta, \xi) \in \Theta_1\} | D]} \cdot \frac{\int_{\Theta_1} \pi^{(f)}(\theta, \xi) d\theta d\xi}{\int_{\Theta_0} \pi^{(f)}(\theta, \xi) d\theta d\xi},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function, and the expectation is taken with respect to the joint posterior distribution of (θ, ξ) , i.e., $\pi^{(f)}(\theta, \xi | D)$.

Proof—Following Klugkist et al. (2005) assuming the encompassing fitting prior, $\pi^{(f)}(\theta, \xi)$, and the fitting priors under H_0 and H_1 :

$$\pi_i^{(f)}(\theta, \xi) = \frac{\pi^{(f)}(\theta, \xi)}{\int_{\Theta_i} \pi^{(f)}(\theta, \xi) d\theta d\xi}, \quad (\theta, \xi) \in \Theta_i, i = 0, 1.$$

Inserting the above fitting priors into (2.2), and letting $\pi^{(f)}(\theta, \xi | D)$ be the joint posterior distribution of (θ, ξ) , we can obtain

$$\begin{aligned}
 B(D) &= \frac{\int_{\Theta_0} \pi^{(f)}(\theta, \xi) L(\theta, \xi | D) d\theta d\xi}{\int_{\Theta_1} \pi^{(f)}(\theta, \xi) L(\theta, \xi | D) d\theta d\xi} \cdot \frac{\int_{\Theta_1} \pi^{(f)}(\theta, \xi) d\theta d\xi}{\int_{\Theta_0} \pi^{(f)}(\theta, \xi) d\theta d\xi} = \frac{\int_{\Theta_0} \pi^{(f)}(\theta, \xi | D) d\theta d\xi}{\int_{\Theta_1} \pi^{(f)}(\theta, \xi | D) d\theta d\xi} \cdot \frac{\int_{\Theta_1} \pi^{(f)}(\theta, \xi) d\theta d\xi}{\int_{\Theta_0} \pi^{(f)}(\theta, \xi) d\theta d\xi} \\
 &= \frac{E[1\{(\theta, \xi) \in \Theta_0 | D\}]}{E[1\{(\theta, \xi) \in \Theta_1 | D\}]} \cdot \frac{\int_{\Theta_1} \pi^{(f)}(\theta, \xi) d\theta d\xi}{\int_{\Theta_0} \pi^{(f)}(\theta, \xi) d\theta d\xi}.
 \end{aligned}$$

In general, we need to compute $B(D)$ and B_0 , and compare them for decision making as shown in (2.3). But note that the second part of $B(D)$, i.e., the ratio $(\int_{\Theta_1} \pi^{(f)}(\theta, \xi) d\theta d\xi) / (\int_{\Theta_0} \pi^{(f)}(\theta, \xi) d\theta d\xi)$, namely, the ratio of the probabilities of the parameter spaces under H_1 and H_0 , is a data independent constant. Therefore it is sufficient to calculate and compare the first part of $B(D)$, i.e., the ratio $E[1\{(\theta, \xi) \in \Theta_0 | D\}] / E[1\{(\theta, \xi) \in \Theta_1 | D\}]$, for decision making.

Further we define

$$g = \frac{E[1\{(\theta, \xi) \in \Theta_0 | D\}]}{E[1\{(\theta, \xi) \in \Theta_1 | D\}]}$$

and use the dropping-needle method discussed in Hoijtink et al. (2008) to calculate g . Specifically, let $\{(\theta_i, \xi_i), i = 1, 2, \dots, N\}$ represent an MC sample from the joint posterior distribution of (θ, ξ) . An MC estimate of g is given by

$$\hat{g} = \frac{\sum_{i=1}^N 1\{(\theta_i, \xi_i) \in \Theta_0\}}{\sum_{i=1}^N 1\{(\theta_i, \xi_i) \in \Theta_1\}}. \quad (3.1)$$

Under certain regularity conditions, such as ergodicity, we can show that $\lim_{N \rightarrow \infty} \hat{g} = g$, which indicates that $\hat{g}$ is a consistent estimator of g .

From (3.1), it is quite natural that one might think that a more efficient way to estimate g might involve generating two independent MC samples from the joint posterior distribution. One sample is used to compute $E[1\{(\theta, \xi) \in \Theta_0 | D\}]$, and the other is used to compute $E[1\{(\theta, \xi) \in \Theta_1 | D\}]$, that is,

$$\hat{g}^* = \frac{N_1^{-1} \sum_{i=1}^{N_1} 1\{(\theta_{1,i}, \xi_{1,i}) \in \Theta_0 | D\}}{N_2^{-1} \sum_{i=1}^{N_2} 1\{(\theta_{2,i}, \xi_{2,i}) \in \Theta_1 | D\}}, \quad (3.2)$$

where $\{(\theta_{1,i}, \xi_{1,i}), i = 1, 2, \dots, N_1\}$ and $\{(\theta_{2,i}, \xi_{2,i}), i = 1, 2, \dots, N_2\}$ are two independent MC random samples from the joint posterior distribution of (θ, ξ) .

Following Chen et al. (2008), it can be shown that the use of two MC samples to estimate g as in (3.2) may not necessarily lead to a more efficient estimate of g than the use of one MC sample to estimate g as in (3.1). In addition, generating two MC samples requires more computation time than generating one MC sample. Therefore, it is unnecessary to use two MC samples to estimate g .

We propose a simple simulation algorithm to obtain the estimate of g , i.e., $\hat{g}$, as follows: For the given data D , we first simulate N (usually big, e.g., 5000) samples of (θ_i, ξ_i) from their joint posterior distribution via the MCMC sampling algorithm or the Gibbs sampling algorithm. Then we count the number of times (θ_i, ξ_i) in the null parameter space Θ_0 and in the alternative parameter space Θ_1 , and denote them by N_0 and N_1 , respectively. We obtain $\hat{g} = N_0/N_1$.

In order to determine the critical value of $\hat{g}$, $\hat{g}_0$, for decision making we first draw $\{(\theta_i, \xi_i), i = 1, 2, \dots, M\}$ (M is often big, e.g., 10,000) from the sampling prior of $\pi_0^{(s)}(\theta, \xi)$, and generate M data set using (θ_i, ξ_i) . Then we obtain $\hat{g}$ for all the simulated data sets using the proposed simulation algorithm described above, and let the critical value of $\hat{g}$ be the α th quantile.

4 Application to a Non-inferiority Clinical Trial

In this section, the proposed Bayesian approach is applied to a published non-inferiority medical device clinical trial on two endpoints with unknown nuisance parameters: a binomially distributed endpoint and a normally distributed endpoint (Chen et al., 2011). The goal of the trial is to show the non-inferiority of the new generation of Drug-Eluting Stent (DES) to the first generation of DES. The primary endpoint is a binary response, the 12-month Target Lesion Failure (TLF), which is an ischemia-driven revascularization of the target lesion, myocardial infarction related to the target vessel, or (cardiac) death related to the target vessel. The secondary endpoint is a continuous response, the 9-month in-segment percent Diameter Stenosis (%DS), which is the percentage of narrowing in the coronary artery caused by plaque.

4.1 Binomial Responses

4.1.1 The Method—We assume that the response for the test group, Y_t , follows a $Bin(n_t, p_t)$ distribution, and the response for the control group, Y_c , follows a $Bin(n_c, p_c)$ distribution, where n_t and n_c are the total numbers of subjects in the test and control group, respectively, and p_t and p_c are the corresponding unknown 12-month TLF rates for the two groups, respectively. Often p_c is treated as a nuisance parameter in a non-inferiority hypothesis testing scenario.

We assume the non-inferiority hypotheses: $H_0: p_t - p_c \geq \delta$ versus $H_1: p_t - p_c < \delta$ ($0 < \delta < 1$), where δ is the non-inferiority margin. Moreover we assume two independent conjugate encompassing priors $Beta(a_t, b_t)$ and $Beta(a_c, b_c)$ for p_t and p_c , respectively. Therefore, the corresponding fitting priors for p_t and p_c are

$$\pi_0^{(f)}(p_t, p_c) = \begin{cases} \frac{1}{C_0} \frac{p_t^{a_t-1} (1-p_t)^{b_t-1} p_c^{a_c-1} (1-p_c)^{b_c-1}}{B(a_t, b_t) B(a_c, b_c)} & \text{if } p_t - p_c \geq \delta, \\ 0 & \text{otherwise,} \end{cases}$$

where a_t, b_t, a_c, b_c are known hyper-parameters, and C_0 is the normalization constant and

$$\pi_1^{(f)}(p_t, p_c) = \begin{cases} \frac{1}{1-C_0} \frac{p_t^{a_t-1} (1-p_t)^{b_t-1} p_c^{a_c-1} (1-p_c)^{b_c-1}}{B(a_t, b_t) B(a_c, b_c)} & \text{if } p_t - p_c < \delta, \\ 0 & \text{otherwise,} \end{cases}$$

After some algebra, the Bayes factor can be written as

$$B(D) = \frac{F_2(1-\delta) - \int_0^{1-\delta} f_2(p_c) F_1(p_c + \delta) dp_c}{1 - F_2(1-\delta) + \int_0^{1-\delta} f_2(p_c) F_1(p_c + \delta) dp_c} \times \frac{1 - F_4(1-\delta) + \int_0^{1-\delta} f_4(p_c) F_3(p_c + \delta) dp_c}{F_4(1-\delta) - \int_0^{1-\delta} f_4(p_c) F_3(p_c + \delta) dp_c},$$

where $F_1(\cdot)$ is the CDF of the $Beta(y_t + a_t, n_t - y_t + b_t)$ distribution; $f_2(\cdot)$ and $F_2(\cdot)$ are the probability distribution function (PDF) and CDF, respectively, of the $Beta(y_c + a_c, n_c - y_c + b_c)$ distribution; $F_3(\cdot)$ is the CDF of the $Beta(a_t, b_t)$ distribution; and $f_4(\cdot)$ and $F_4(\cdot)$ are the PDF and CDF, respectively, of the $Beta(a_c, b_c)$ distribution.

Moreover, we assume the point mass sampling priors for p_t and p_c as

$$\pi_0^{(s)}(p_t) = \begin{cases} 1 & \text{if } p_t = p_{0t}, \\ 0 & \text{otherwise} \end{cases}; \quad \pi_0^{(s)}(p_c) = \begin{cases} 1 & \text{if } p_c = p_{0c}, \\ 0 & \text{otherwise} \end{cases},$$

where p_{0t} and p_{0c} are the known 12-month TLF rates. Then the type I error and power are estimated via simulations by letting $p_{0t} = p_{0c} + \delta$ and $p_{0t} = p_{0c}$, respectively, where p_{0c} is a pre-specified failure rate for the control group as commonly elicited and used in the frequentist SSD approach.

Remark 11: Under this scenario, as the total number of the combination of random data points is finite for known integers n_t and n_c , the probability function of the Bayes factor can be fully determined given p_{0c} ; hence, the critical value B_0 can be determined from the distribution of the Bayes factor B . However, in practice, numerical simulations may still be needed if $n_t \times n_c$ is sufficiently large that the enumeration of the full distribution of the Bayes factor B is practically impossible within a reasonable time.

4.1.2 Application to the Medical Device Clinical Trial—Historical data are available for the first generation of DES from two previous trials. Summary of the historical data is given in Table 1 of Chen et al. (2011). The percentage of the 12-month TLF is estimated to be about 8.2% (44 failures of total 535) in the first historical trial, and is estimated to be about 10.9% (33 failures of total 304) in the second historical trial.

We use the power priors of Ibrahim and Chen (2000) and Ibrahim et al. (2015) as the encompassing fitting priors for p_t and p_c but with fixed power parameters a_{0k} , $k = 1, 2$, to incorporate the data from the two historical trials. In this case, it is equivalent to assuming *Beta* fitting priors as described in the previous section with hyper-parameters of $a = \sum_{k=1}^K a_{0k} y_{0k}$ and $b = \sum_{k=1}^K a_{0k} (n_{0k} - y_{0k})$ for the test and control groups. Here n_{0k} and y_{0k} represent the total number of subjects and the numbers of failures, respectively, in the k th historical data set, and a_{0k} is the corresponding power parameter in the power priors. For simplicity, we assume the same power parameters for the two historical data sets for the control group, i.e., $a_0 = a_{01c} = a_{02c}$. As there are no historical data for the new generation of DES, we choose an encompassing fitting prior for the test group with $a_t = b_t = 1$. In addition, the non-inferiority margin δ is set to be 0.041 throughout the simulations.

For a given range of p_{0c} , say, (p_{0ch}, p_{0cu}) , the goal of our study design is to find a minimum total sample size n with fixed ratio of n_t/n_c (say, 3) and a value of the power parameter a_0 so that type I error is controlled at the nominal level, say, 5%, and power is at least at the acceptance level, say, 80%. Through sensitivity analysis it has been recognized that in order to control type I error for all possible future values of p_{0c} , the design value of p_{0c} used in the computation of the critical value B_0 , i.e., the p_{0c} in the sampling prior $\pi_0^{(s)}(p_c)$, denoted by $p_{0c}^{(B_0)}$, has to be set at the lower bound of p_{0c} , i.e., p_{0ch} .

We first investigate the impact of the power parameter a_0 on type I error and power under the assumption that $(p_{0ch}, p_{0cu}) = (0.08, 0.092)$, $p_{0c}^{(B_0)} = 0.08$, $n_t = 900$, and $n_c = 300$. When $a_0 = 0$, the proposed procedure yields a type I error of 0.0493 and a power of 0.7302 when $p_{0c} = 0.092$. Also from Table 2 of Chen et al. (2011), we see that the z-test (unpooled) yields a power of 0.710 while the score test gives a power of 0.736. Therefore, without incorporation of historical data, the proposed Bayesian procedure is comparable to a frequentist procedure. The results summarized in Table 1 and Figure 2 demonstrate that the more historical data we borrow, the higher power we have, and that type I error is always no more than 5%. The full historical data should be borrowed and the power parameter a_0 should be 1.0 in order to maximize power, which implies that a random power parameter a_0 need not be considered here. We then examine type I errors and powers under different total sample sizes n by incorporating the two full historical data sets. The corresponding results are given in Table 2 and Figure 3, which indicate that $n = 880$ is the minimal total sample size for the study in terms of the primary endpoint, under which type I error is at most 5% and power is at least 80%.

It is worth pointing out that the design value of p_{0c} can be conservative by taking the minimum possible future value, so that the increase in study power might be very limited, especially for a wider range of the future values of p_{0c} .

If no historical data are borrowed, type I error is always controlled regardless of the choices of $p_{0c}^{(B_0)}$ and p_{0c} . Moreover, the power will not be impacted by the choice of $p_{0c}^{(B_0)}$, but it decreases as p_{0c} increases.

4.2 Normally Distributed Responses

4.2.1 The Method—Assume that the response in the test group, Y_t follows a $N(\mu_t, \sigma^2)$ distribution and the response in the control group, Y_c , follows a $N(\mu_c, \sigma^2)$ distribution, where μ_t , μ_c , and σ^2 are unknown parameters. Assume that a set of i.i.d. samples are collected: $\{y_{ti}, i = 1, 2, \dots, n_t\}$ and $\{y_{cj}, j = 1, 2, \dots, n_c\}$, and let D denote the collected data $\{y_{ti}, y_{cj}\}$.

Observe that $(\bar{y}_t, \bar{y}_c, s^2)$ are sufficient statistics, where $\bar{y}_t = \sum_i y_{ti}/n_t$, $\bar{y}_c = \sum_j y_{cj}/n_c$, and $s^2 = (\sum_i (y_{ti} - \bar{y}_t)^2 + \sum_j (y_{cj} - \bar{y}_c)^2)/(n_t + n_c - 2)$. The likelihood function can be written as

$$L(\mu_t, \mu_c, \sigma^2 | D) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n_t + n_c}{2}} \exp \left(- \frac{(n_t + n_c - 2)s^2 + n_t(\bar{y}_t - \mu_t)^2 + n_c(\bar{y}_c - \mu_c)^2}{2\sigma^2} \right).$$

Assume the following non-inferiority hypotheses: $H_0: \mu_t - \mu_c \geq \delta$ versus $H_1: \mu_t - \mu_c < \delta$ ($0 < \delta < 1$), where δ is the non-inferiority margin.

Furthermore, we assume that there are K historical data sets available for the control group. We use the power prior in Ibrahim and Chen (2000) and Ibrahim et al. (2015) as the corresponding encompassing fitting priors for the control group mean, μ_c , and the common variance, σ^2 . Let $\mathbf{y}_{0k} = (y_{0k1}, y_{0k2}, \dots, y_{0kn_{0k}})'$, $k = 1, 2, \dots, K$, be the K historical data sets. Write $\bar{y}_{0k} = \sum_i y_{0ki}/n_{0k}$ and $s_{0k}^2 = \sum_i (y_{0ki} - \bar{y}_{0k})^2/(n_{0k} - 1)$ and let a_{0k} be the fixed power parameter corresponding to the k th historical data set. As there are no historical data available for the test group, a non-informative encompassing fitting prior is assumed for the test group mean, μ_t . Then, the joint fitting prior can be written as

$$\begin{aligned} \pi^{(f)}(\mu_t, \mu_c, \sigma^2) &= \pi^{(f)}(\mu_t, \mu_c, \sigma^2 | \mathbf{y}_{0k}) \propto \pi(\mu_c, \sigma^2 | \mathbf{y}_{0k}, a_{0k}) \pi_0(\mu_t, \mu_c, \sigma^2) = \pi(\mu_c, \sigma^2 | \mathbf{y}_{0k}, a_{0k}) \cdot \frac{1}{\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{\sum_k a_{0k} n_{0k}}{2}} \exp \left(- \sum_k a_{0k} \frac{(n_{0k} - 1)s_{0k}^2 + n_{0k}(\bar{y}_{0k} - \mu_c)^2}{2\sigma^2} \right) \cdot \frac{1}{\sigma^2} \\ &\propto \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{2 + \sum_k a_{0k} n_{0k}}{2}} \exp \left(- \sum_k a_{0k} \frac{(n_{0k} - 1)s_{0k}^2 + n_{0k}(\bar{y}_{0k} - \mu_c)^2}{2\sigma^2} \right). \end{aligned}$$

After some algebra, the joint posterior distribution for μ_t, μ_c, σ^2 given the data D can be written as

$$\pi(\mu_t, \mu_c, \sigma^2 | D) \propto L(\mu_t, \mu_c, \sigma^2 | D) \pi^{(f)}(\mu_t, \mu_c, \sigma^2) \propto \left(\frac{1}{\sigma^2} \right)^{A/2} \exp \left(- \frac{B}{2\sigma^2} \right),$$

where $A = n_t + n_c + 2 + \sum_k a_{0k} n_{0k}$,

$$B = (n_t + n_c - 2)s^2 + n_t(\bar{y}_t - \mu_t)^2 + n_c(\bar{y}_c - \mu_c)^2 + \quad , \text{ and}$$

$$\begin{aligned} \sum_k a_{0k} [(n_{0k} - 1)s_{0k}^2 + n_{0k}(\bar{y}_{0k} - \mu_c)^2] &= C + n_t(\bar{y}_t - \mu_t)^2 + (n_c + \sum_k a_{0k} n_{0k}) \\ &\times \left(\mu_c - \frac{n_c \bar{y}_c + \sum_k a_{0k} n_{0k} \bar{y}_{0k}}{n_c + \sum_k a_{0k} n_{0k}} \right)^2 \\ C &= (n_t + n_c - 2)s^2 + \sum_k a_{0k} (n_{0k} - 1)s_{0k}^2 + n_c \bar{y}_c^2 + \sum_k a_{0k} n_{0k} \bar{y}_{0k}^2 - \frac{(n_c \bar{y}_c + \sum_k a_{0k} n_{0k} \bar{y}_{0k})^2}{n_c + \sum_k a_{0k} n_{0k}}. \end{aligned}$$

Therefore, the marginal posterior distribution of σ^2 is

$$\pi(\sigma^2 | D) = \int \int \pi(\mu_t, \mu_c, \sigma^2 | D) d\mu_t d\mu_c \propto \left(\frac{1}{\sigma^2} \right)^{\frac{A-2}{2}} \exp \left(-\frac{C}{2\sigma^2} \right),$$

which is a scaled inverse chi-square distribution, $\text{inv-}\chi^2(A-4, C/(A-4))$, or equivalently the inv-gamma $((A-4)/2, C/2)$ distribution. The full conditional distribution of μ_t given σ^2 and D is $\pi(\mu_t | \sigma^2, D) \sim N(\bar{y}_t, \sigma^2/n_t)$, and the full conditional distribution of μ_c given (σ^2, D) is

$$\pi(\mu_c | \sigma^2, D) \sim N \left(\frac{n_c \bar{y}_c + \sum_k a_{0k} n_{0k} \bar{y}_{0k}}{n_c + \sum_k a_{0k} n_{0k}}, \frac{\sigma^2}{n_c + \sum_k a_{0k} n_{0k}} \right).$$

The marginal posterior distribution of (μ_c, σ^2) is

$$\pi(\mu_c, \sigma^2 | D) = \int \pi(\mu_t, \mu_c, \sigma^2 | D) d\mu_t \propto \left(\frac{1}{\sigma^2} \right)^{\frac{A-1}{2}} \exp \left(-\frac{C + (n_c + \sum_k a_{0k} n_{0k})(\mu_c - E)^2}{2\sigma^2} \right),$$

where $E = (n_c \bar{y}_c + \sum_k a_{0k} n_{0k} \bar{y}_{0k}) / (n_c + \sum_k a_{0k} n_{0k})$. Hence, the marginal posterior distribution of μ_c is

$$\pi(\mu_c | D) = \int \pi(\mu_c, \sigma^2 | D) d\sigma^2 \propto \frac{\Gamma(\frac{A-3}{2})}{\left(\frac{C + (n_c + \sum_k a_{0k} n_{0k})(\mu_c - E)^2}{2} \right)^{\frac{A-3}{2}}},$$

which is a three parameter Student's t distribution with location parameter E , inverse scale parameter (i.e., precision) $(A-4)(n_c + \sum_k a_{0k} n_{0k})/C$, and $A-4$ degrees of freedom. After the transformation, $((A-4)(n_c + \sum_k a_{0k} n_{0k})/C)^{1/2}(\mu_c - E)D$ follows a Student-t distribution with $A-4$ degrees of freedom.

Finally, we assume the point mass sampling priors for μ_b , μ_c , and σ^2 as

$$\pi_0^{(s)}(\mu_t) = \begin{cases} 1 & \text{if } \mu_t = \mu_{0t}; \\ 0 & \text{otherwise} \end{cases}; \quad \pi_0^{(s)}(\mu_c) = \begin{cases} 1 & \text{if } \mu_c = \mu_{0c}; \\ 0 & \text{otherwise} \end{cases}; \quad \pi_0^{(s)}(\sigma^2) = \begin{cases} 1 & \text{if } \sigma^2 = \sigma_0^2; \\ 0 & \text{otherwise} \end{cases},$$

where μ_{0t} , μ_{0c} , σ_0^2 are known constants. Then type I error and power are estimated via simulations by letting $\mu_{0t} = \mu_{0c} + \delta$ and $\mu_{0t} = \mu_{0c}$, respectively, where μ_{0c} is a pre-specified value for the control group mean as commonly elicited and used in the frequentist SSD approach.

4.2.2 Application to the Medical Device Clinical Trial—For the medical device example, the historical data for the 9-month in-segment %DS, which is the secondary endpoint, are available from the two previous trials. A summary of the historical data is given in Table 1 of Chen et al. (2011). The mean $\pm$ SD for the 9-month in-segment %DS is 3.0891 ± 0.6315 from the first historical trial, and is 3.1849 ± 0.5811 from the second historical trial. The sample sizes are 242 and 263 for these two trials, respectively. We assume that the non-inferiority margin δ is 0.2.

For a given range of μ_{0c} , say, (μ_{0cl}, μ_{0cu}) , and a range of σ_0^2 , say, $(\sigma_{0l}^2, \sigma_{0u}^2)$, the goal of the study design is to find a minimum total sample size n with a fixed ratio of n_t/n_c (say, 3) and a value of the power parameter a_0 so that type I error is controlled at the nominal level of 5%, and power is at least at the acceptance level of 80%.

Through a sensitivity analysis, it is found that the design values for μ_{0c} and σ_0^2 used in the computation of the critical value B_0 , i.e., the μ_{0c} in the sampling prior $\pi_0^{(s)}(\mu_c)$ and the σ_0^2 in $\pi_0^{(s)}(\sigma^2)$, denoted by $\mu_{0c}^{(B_0)}$ and $\sigma_0^{2(B_0)}$, respectively, have to be set at μ_{0cl} and σ_{0u}^2 , respectively, in order to control type I error for all possible future values of μ_{0c} and σ_0^2 . The analysis of the impact of the power parameter shows that the more historical data we incorporate, the higher power we have. That is to say, we can incorporate the full historical data, i.e., $a_0 = a_{01} = a_{02} = 1.0$, to maximize study power.

Table 3 reports type I errors and powers under different total sample sizes n by incorporating the two full historical data sets, and it shows that $n = 180$ is the minimal total sample size for the study in terms of the secondary endpoint, under which type I error is at most 5% and power is at least 80%. Combined with the results for the primary endpoint, the minimal total sample size for the study is $n = 880$.

We note that μ_{0cl} and σ_{0u}^2 are the two key quantities, which will affect the type I error.

Choosing a small value μ_{0cl} coupled with a large value σ_{0u}^2 leads to a small type I error.

However, such a choice will also lead to a low power, implying a large sample size. In practice, the lower limit of a 95% confidence interval (CI) for μ_c and the upper limit of a 95% CI for σ^2 based on the data from the historical trials are a reasonable choice.

5 Conclusion and Discussion

In this paper, we develop a general Bayesian approach using formal Bayesian hypothesis testing via the Bayes factor, apply the method to a published medical device clinical trial, and demonstrate that the proposed Bayesian approach can control type I error appropriately. We empirically show that compared with the frequentist approach, the Bayesian approach has higher power and leads to sample size reduction by incorporating historical data. Although we mainly consider non-inferiority clinical trials, the proposed approach can be extended to other types of clinical trials by appropriately modifying the design settings and the computational algorithm.

The advantages of the Bayesian approach include the accountability of parameter uncertainties; the ability to control type I error; and the ability to incorporate historical data, leading to a potential power increase and sample size reduction. There are challenges associated with the Bayesian approach. The Bayesian approach often involves extensive computation. Hence, it may require a longer planning time. In addition, the Bayes factor-based approach requires specifying proper fitting priors, which may not be an easy task especially when historical data are not available. Compared to Weiss (1997), the proposed approach does not have a “rule of thumb” type of interpretation but has a better control of type I error than Weiss’s approach. Compared to Chen et al. (2011), the proposed approach has a better control of type I error and allows full borrowing of the historical data. However, computing the Bayes factors is quite time consuming. Therefore, there is a need to develop user-friendly software for implementing the proposed Bayes factor based approach, which is currently under investigation.

Acknowledgments

The authors would like to thank the Editor, the Guest Editors, and two anonymous referees for their helpful comments. The authors would also like to thank Jim Ranger-Moore, Sr. Director, Biostatistics & Data Management, Ventana Medical Systems, Inc. for reading the manuscript carefully and providing useful comments. These comments have lead to a much improved version of the paper. Dr. M.-H. Chen’s research was partially supported by NIH grants #GM70335 and #P01CA142538.

References

- Adcock CJ. Sample size determination: a review. *The statistician*. 1997; 44:143–154.
- Berger J. *Statistical Decision Theory: Foundations, Concepts, and Methods*. Second. New York: Springer; 1980.
- Campbell G. Bayesian statistics in medical devices: innovation sparked by the FDA. *Journal of Biopharmaceutical Statistics*. 2011; 21:871–887. [PubMed: 21830920]
- Chen M-H, Huang L, Ibrahim JG, Kim S. Bayesian variable selection and computation for generalized linear models with conjugate priors. *Bayesian Analysis*. 2008; 3:585–614. [PubMed: 19436774]
- Chen M-H, Ibrahim JG, Lam P, Yu A, Zhang Y. Bayesian design of non-inferiority trials for medical devices using historical data. *Biometrics*. 2011; 67:1163–70. [PubMed: 21361889]
- Chen M-H, Ibrahim JG, Xia HA, Liu T, Hennessey V. Bayesian sequential meta-analysis design in evaluating cardiovascular risk in a new antidiabetic drug development program. *Statistics in Medicine*. 2014a; 33:1600–1618. [PubMed: 24343859]
- Chen M-H, Ibrahim JG, Zeng D, Hu K, Jia C. Bayesian Design of superiority clinical trials for recurrent events data with applications to bleeding and transfusion events in myelodysplastic syndrome. *Biometrics*. 2014b; 70:1003–1013. [PubMed: 25041037]

- Chow SC, Chang M, Pong J. Statistical consideration of adaptive methods in clinical development. *Journal of Biopharmaceutical Statistics*. 2005; 15:575–591. [PubMed: 16022164]
- Chow SC, Liu JP. *Design and Analysis of Clinical Trials, Concepts and Methodologies*. Second. New York: Wiley; 2004.
- D'Agostino RB Sr, Massaro JM, Sullivan LM. Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics. *Statistics in Medicine*. 2003; 22:169–186. [PubMed: 12520555]
- De Santis F. Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*. 2004; 124:121–144.
- De Santis F. Using historical data for Bayesian sample size determination. *Journal of the Royal Statistical Society, Series A*. 2007; 170:95–113.
- DiCiccio TJ, Kass RE, Raftery A, Wasserman L. Computing Bayes factor by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*. 1997; 92:903–915.
- Gamalo-Siebers M, Gao A, Lakshminarayanan M, Liu G, Natanegara F, Railkar R, Schmidli H, Song G. Bayesian methods for the design and analysis of noninferiority trials. *Journal of Biopharmaceutical Statistics*. 2015; 26:823–841. [PubMed: 26247350]
- Garcia-Donato G, Chen M-H. Calibrating Bayes factor under prior predictive distributions. *Statistica Sinica*. 2005; 15:359–380.
- Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Second. Boca Raton, FL: Chapman & Hall; 2006. 80–83.
- Hoijtink H, Klugkist I, Boelen P. *Bayesian Evaluation of Informative Hypotheses*. New York: Springer; 2008.
- Ibrahim JG, Chen M-H. Power prior distributions for regression models. *Statistical Science*. 2000; 15:46–60.
- Ibrahim JG, Chen M-H, Gwon Y, Chen F. The power prior: theory and applications. *Statistics in Medicine*. 2015a; 34:3724–3749. [PubMed: 26346180]
- Ibrahim JG, Chen M-H, Xia HA, Liu T. Bayesian meta-experimental design: evaluating cardiovascular risk in new antidiabetic therapies to treat type 2 diabetes. *Biometrics*. 2012; 68:578–586. [PubMed: 21955084]
- Ibrahim JG, Gwon Y, Chen M-H. SAS Macro BSMED: Bayesian survival meta-experimental design using historical data. In: Menon SM, Zink RC, editors *Modern Approaches to Clinical Trials Using SAS: Classical, Adaptive, and Bayesian Methods*. Gary, NC: SAS Institute Inc.; 2015b. 107–131.
- Jeffreys H. *Theory of Probability*. 3. London: Oxford University Press; 1961.
- Joseph L, Wolfson DB, du Berger R. Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician*. 1995; 44:143–154.
- Joseph L, Belisle P. Bayesian sample size determination for normal means and differences between normal means. *The Statistician*. 1997; 46:209–226.
- Kang SH, Tsong Y. Strength of evidence of non-inferiority trials - The adjustment of the type I error rate in non-inferiority trials with the synthesis method. *Statistics in Medicine*. 2010; 29:1477–1487. [PubMed: 20535762]
- Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association*. 1995; 90:773–795.
- Kieser M, Friede T. Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine*. 2007; 26:253–273. [PubMed: 16685703]
- Klugkist I, Kato B, Hoijtink H. Bayesian model selection using encompassing priors. *Statistica Neerlandica*. 2005; 59:57–69.
- Liu KJ, Chang KC. Test non-inferiority and sample size determination based on the odds ratio under a cluster randomized trial with noncompliance. *Journal of Biopharmaceutical statistics*. 2011; 21:94–110. [PubMed: 21191857]
- M'lan CE, Joseph L, Wolfson DB. Bayesian sample size determination for case-control studies. *Journal of the American Statistical Association*. 2006; 101:760–772.

- M'lan CE, Joseph L, Wolfson DB. Bayesian sample size determination for binomial proportions. *Bayesian Analysis*. 2008; 3:269–296.
- Osman M, Ghosh SK. Semiparametric Bayesian Testing Procedure for Noninferiority Trials with Binary Endpoints. *Journal of Biopharmaceutical Statistics*. 2011; 21:920–937. [PubMed: 21830923]
- Reyes EM, Ghosh SK. Bayesian average error-based approach to sample size calculations for hypothesis testing. *Journal of Biopharmaceutical Statistics*. 2013; 23:569–588. [PubMed: 23611196]
- Wang F, Gelfand AE. A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*. 2002; 17:193–208.
- Weiss RE. Bayesian sample size calculations for hypothesis testing. *The Statistician*. 1997; 46:185–191.
- Weiss RE, Wang Y. Experimental design and sample size computations for longitudinal models. In: Chen M-H, Dey DK, Müller P, Sun D, Ye K, editors *Frontiers of Statistical Decision Making and Bayesian Analysis*. New York: Springer; 2010. 270–276.
- Yu F, Chen M-H, Kuo L. Detecting differentially expressed genes using calibrated Bayes factors. *Statistica Sinica*. 2008; 18:783–802.

Figure 1.
The plot of the power as a function of θ_1 under $\theta_0 = 1$, $\sigma = 1.5$, $\mu = 1$, $\tau = 1.5$, $\delta = 1$, and $n = 15$

Figure 2.

Type I error and power versus a_0 under $p_{0c}=(0.08, 0.092)$, $p_{0c}^{(B_0)} = 0.08$, $n_t = 900$, and $n_c = 300$

Figure 3.

Type I error and power versus total sample size n under $p_{0c} = (0.08, 0.092)$, $p_{0c}^{(B_0)} = 0.08$, $a_0 = 1$, and $n_t/n_c = 3$

Table 1

Type I error and power as a function of the power prior parameter a_0 under $p_{0c}^{(B_0)} = 0.08$, $n_t = 900$, and $n_c = 300$

a_0	$p_{0c} = 0.08$		$p_{0c} = 0.092$	
	type I error	power	type I error	power
0	0.0493	0.7846	0.0493	0.7302
0.1	0.0496	0.8774	0.0355	0.7902
0.3	0.0506	0.9503	0.0220	0.8519
0.5	0.0500	0.9733	0.0152	0.8767
0.7	0.0507	0.9822	0.0118	0.8897
0.9	0.0504	0.9858	0.0100	0.8941
1	0.0505	0.9875	0.0091	0.8965

Table 2

Power and type I error under $p_{0c}^{(B_0)} = 0.08, a_0 = 1$

Sample size			$p_{0c} = 0.08$		$p_{0c} = 0.092$	
n	n_t	n_c	type I error	power	type I error	power
840	630	210	0.0494	0.9519	0.0100	0.7819
880	660	220	0.0494	0.9587	0.0102	0.7995
920	690	230	0.0491	0.9643	0.0096	0.8136
1000	750	250	0.0497	0.9742	0.0095	0.8429
1200	900	300	0.0505	0.9875	0.0091	0.8965

Table 3

The type I error and power for $\mu_{0c}^{(B_0)} = 3.10, \sigma_0^{2(B_0)} = 0.7, a_{01} = a_{02} = 1$

n	n_t	n_c	μ_{0c}	σ_0^2	Type I error	Power
260	195	65	3.10	0.61	0.039	0.997
			3.10	0.70	0.048	0.988
			3.15	0.61	0.003	0.957
			3.15	0.70	0.006	0.920
200	150	50	3.10	0.61	0.040	0.986
			3.10	0.70	0.050	0.965
			3.15	0.61	0.004	0.901
			3.15	0.70	0.008	0.853
180	135	45	3.10	0.61	0.037	0.975
			3.10	0.70	0.051	0.950
			3.15	0.61	0.004	0.865
			3.15	0.70	0.009	0.811