

Performance of Bayesian methods in non-inferiority tests based on relative risk and odds ratio for dichotomous data

Muhtarjan Osman and Sujit K. Ghosh
*Department of Statistics, NC State University,
Raleigh, NC 27695-8203, USA*

Abstract

In a non-inferiority trial, the experimental treatment is compared against an active control instead of placebo. The goal of these studies is often to show the experimental treatment is non-inferior to the control by some pre-specified margin. The standard approach for these problems, which relies on asymptotic normality, usually requires large sample size to achieve some desired power level. In this paper, we propose alternative approaches based on Bayes factor and posterior probability for testing non-inferiority in the context of two-sample dichotomous data. Results based on simulated data indicate that both of the proposed Bayesian approaches provide significant improvement in terms of statistical power as well as the total error rate over the popularly used frequentist procedures. This in turn indicates the required sample size to achieve certain power level could be substantially lowered by using the proposed Bayesian approaches. The two Bayesian methods however are comparable and are found to be not significantly different in terms of statistical power and total error rates.

Keywords: Bayes factor; non-inferiority; odds ratio; posterior probability; relative risk

1 Introduction

Selecting an appropriate control group is a very important step in many medical applications, such as showing therapeutic effect of a certain medical intervention. Normally, a placebo group is the most ideal candidate for the control. In some situations, however, a placebo control is considered not feasible due to ethical concerns about randomizing patients with life-threatening diseases to the placebo group when a widely accepted treatment is already available. In some other cases, a placebo control is just impossible due to the nature of some treatment methods like device implant or surgery. Therefore, an active control is used to compare against the study treatment. Generally, the best available treatment is selected as the active control in these trials to address the ethical issue mentioned above and to avoid “biocreep” [1]. Establishing superiority of a new treatment over the active control that already has the best proven efficacy is desirable but usually to be a difficult task. Instead, sometimes it is acceptable to show the experimental treatment is not inferior to the standard treatment by some small margin if the experimental treatment is believed to have other advantages compared to the standard treatment. These ancillary benefits may include less toxicity, easier administration, or lower cost.

For two-sample dichotomous data, suppose the success rates of the active control group and the experimental group are denoted by θ_1 and θ_2 , respectively. The hypotheses for testing non-inferiority can be formulated as: $H_0 : \theta_2 - \theta_1 \leq -\delta$ vs. $H_1 : \theta_2 - \theta_1 > -\delta$, where $\delta > 0$ is the pre-specified non-inferiority margin. When the sample size is large and δ is relatively small, the Wald type test statistic $Z = (\hat{\theta}_2 - \hat{\theta}_1 + \delta) / \sqrt{\hat{\theta}_1(1 - \hat{\theta}_1)/n_1 + \hat{\theta}_2(1 - \hat{\theta}_2)/n_2}$ will approximately follow normal distribution under the null hypothesis, where $\hat{\theta}_1$ and $\hat{\theta}_2$ are sample proportions and n_1 and n_2 denote the corresponding sample sizes [2]. Given a level α , the null hypothesis will be rejected if $Z > Z_{1-\alpha}$, where $Z_{1-\alpha}$ denotes the $(1 - \alpha)$ th quantile of a standard normal distribution. Farrington and Manning [3] further suggested replacing the sample proportions in the standard error with the restricted maximum likelihood estimates of θ_1 and θ_2 under the null hypothesis. Due to the regulatory requirement on the assay sensitivity [4], the non-inferiority margin is often set to a small number (e.g., $\delta = 0.1$). Therefore, this test based on asymptotic normality usually results in very large sample size. As an example reported in [5], the study GUSTO III with over 15,000 patients is still under-powered for its setup to assess non-inferiority. Since both θ_1 and θ_2 are positive numbers between 0

and 1, other dissimilarity measures such as relative risk (θ_2/θ_1) and odds ratio ($\frac{\theta_2(1-\theta_1)}{(1-\theta_2)\theta_1}$) also can be used to test non-inferiority. One advantage of using relative risk and odds ratio is that they take the sizes of θ_1 and θ_2 into account. As pointed out in [6], the non-inferiority test based on the absolute difference is not meaningful when the reference θ_1 is close to 0. Similar to the previous case, test statistics using relative risk and odds ratio can be constructed using the standard “delta” method approximation (see [7, 8]).

Based on a Bayesian framework, Wellek [6] suggested a test using posterior probability of the alternative hypothesis with Jeffreys’ prior distribution for θ_1 and θ_2 . All of the three dissimilarity measures (absolute difference, relative risk, and odds ratio) were considered. Williamson [9] proposed a test based on Bayes factor in the context of equivalence tests. The Bayesian tests are usually based on comparing the posterior probability of the competing hypotheses when prior probability of the hypotheses are chosen to be unequal. Bayes factor, which is defined as the ratio of posterior odds of the hypotheses to the corresponding prior odds, also can be used to measure the evidence against (or in favor of) the null hypothesis. A Bayes factor can be sensitive to the selection of prior distribution, Williamson argues that more than one prior should be attempted. In a simulation study carried out by the author, 4 different sets of prior distributions for θ_1 and θ_2 were investigated. The rejection region of the testing method in [9] was determined in an ad-hoc manner, a formal justification for the choice of the the cut-off value of the Bayes factor was not provided.

In this paper, we propose a thorough testing procedure based on Bayes factor for non-inferiority trials with binary data. Priors are chosen to minimize the difference in probability of the two competing hypotheses. More importantly, the key aspect that distinguishes this approach from the currently available Bayesian testing procedures in the similar context is the determination of the threshold or critical value of Bayes factor used in the decision rule. Instead of being pre-specified to some fixed value, the critical value of the Bayes factor is determined using a criterion that approximately maintains the frequentist type I error rate to a desired level. Consequently, the cut-off value of Bayes factor in the decision rule will depend on the design parameters like sample sizes and the non-inferiority margin. Additionally, the posterior probability approach is also considered but with different prior specification from the one given in [6]. A detailed description of these Bayesian approaches will be given in Section 2. In Section 3, we carried out several simulation studies to explore the sampling properties of the Bayesian methods and compared our methods to the popularly used frequentist procedures. An

application to real data using the proposed Bayesian methods is provided in Section 4. Finally, we conclude with some discussions in Section 5.

2 Bayesian Methods for Non-inferiority Test

Suppose in a two-arm study, n_1 subjects are randomized to receive the standard treatment (active control) and the experimental treatment group has n_2 subjects. Let X_1 and X_2 denote the number of successes in the corresponding treatment groups, we further assume X_1 and X_2 are independent and follow binomial distribution, i.e.,

$$\begin{aligned} X_1|\theta_1 &\sim \text{Bin}(n_1, \theta_1) \text{ and} \\ X_2|\theta_2 &\sim \text{Bin}(n_2, \theta_2), \end{aligned} \quad (2.1)$$

where θ_1 and θ_2 are the success rates of the active control and the experimental treatment, respectively. Following the standard Bayesian inferential procedure, the parameters are assigned prior distribution as $(\theta_1, \theta_2) \sim \pi(\cdot, \cdot)$. Next we develop suitable prior distribution for a given hypothesis test.

2.1 Bayes Factor Approach

Under the general framework described above, the null and alternative hypotheses for non-inferiority tests can be expressed as:

$$H_0 : \theta_2 \leq g(\theta_1, \rho) \text{ vs. } H_1 : \theta_2 > g(\theta_1, \rho), \quad (2.2)$$

where ρ is a pre-determined real-valued quantity and $g(\cdot, \cdot)$ is a continuous function of θ_1 and ρ . For example, when $g(\theta_1, \rho) = \theta_1 - \rho$, it leads to the null hypothesis of comparing differences $H_0 : \theta_2 - \theta_1 \leq -\rho$, where ρ is usually called the non-inferiority margin. Also $g(\theta_1, \rho) = \rho\theta_1$ leads to the null hypothesis to compare the relative risk $H_0 : \theta_2 \leq \rho\theta_1$, and $g(\theta_1, \rho) = \frac{\rho\theta_1}{1-\theta_1+\rho\theta_1}$ leads to the null hypothesis of comparing the odds ratio $H_0 : \frac{\theta_2}{1-\theta_2} \leq \rho \frac{\theta_1}{1-\theta_1}$. The hypotheses (2.2) can equivalently be expressed as $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, where $\theta = (\theta_1, \theta_2)$, $\Theta_0 = \{(\theta_1, \theta_2) \in [0, 1]^2 : \theta_2 \leq g(\theta_1, \rho)\}$, and $\Theta_1 = \{(\theta_1, \theta_2) \in [0, 1]^2 : \theta_2 > g(\theta_1, \rho)\}$.

Next we develop a Bayes factor based test for the general form of the hypotheses in (2.2) using independent conjugate priors $\theta_1 \sim Beta(a(\rho), a(\rho))$ and $\theta_2 \sim Beta(a(\rho), a(\rho))$, where the prior parameter $a(\rho)$ is allowed to depend on ρ specified in (2.2). In order to achieve balance between H_0 and H_1 in prior selection we determine

$$\tilde{a}(\rho) = \arg \min_{a \in [0,1]} |P[\theta_2 \leq g(\theta_1, \rho) | a(\rho) = a] - 0.5|. \quad (2.3)$$

In general, there may not exist an $a(\rho) = a$ such that $P[\theta_2 \leq g(\theta_1, \rho) | a(\rho) = a] = 0.5$. But $\tilde{a}(\rho)$ as defined in (2.3) always exists as $P[\theta_2 \leq g(\theta_1, \rho) | a(\rho) = a]$ is a continuous function of a defined over a closed and bounded interval $[0, 1]$. In order to determine $\tilde{a}(\rho)$ in (2.3) we compute

$$\begin{aligned} & Pr[(\theta_1, \theta_2) \in \Theta_0] \\ &= Pr[\theta_2 \leq g(\theta_1, \rho) | a(\rho) = a] \\ &= \frac{1}{B(a, a)^2} \int_0^1 \left\{ \int_0^{g(\theta_1, \rho)} \theta_2^{a-1} (1 - \theta_2)^{a-1} d\theta_2 \right\} \theta_1^{a-1} (1 - \theta_1)^{a-1} d\theta_1 \\ &= \frac{1}{B(a, a)} \int_0^1 F^*(g(\theta_1, \rho); a, a) \theta_1^{a-1} (1 - \theta_1)^{a-1} d\theta_1 \\ &= \int_0^1 F^*(g(\theta_1, \rho); a, a) f^*(\theta_1; a, a) d\theta_1 \\ &= E_{\theta_1}[F^*(g(\theta_1, \rho); a, a)], \end{aligned} \quad (2.4)$$

where $f^*(t; a, b)$ and $F^*(t; a, b)$ denote probability density function (pdf) and cumulative density function (cdf) of $Beta(a, b)$ distribution, respectively, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, and $\Gamma(\cdot)$ denotes the Gamma function.

Once the prior $Beta(\tilde{a}, \tilde{a})$ is determined for a given value of ρ , the posterior becomes

$$\begin{aligned} \theta_1 | x_1 &\sim Beta(\tilde{a} + x_1, \tilde{a} + n_1 - x_1) \text{ and} \\ \theta_2 | x_2 &\sim Beta(\tilde{a} + x_2, \tilde{a} + n_2 - x_2). \end{aligned} \quad (2.5)$$

Hence, by conjugacy it follows that the posterior probability of the null hypothesis in

(2.2) can be written as:

$$\begin{aligned}
& Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2] \\
&= Pr[\theta_2 \leq g(\theta_1, \rho) | X_1 = x_1, X_2 = x_2] \\
&= \int_0^1 F^*(g(\theta_1, \rho); \tilde{a} + x_2, \tilde{a} + n_2 - x_2) f^*(\theta_1; \tilde{a} + x_1, \tilde{a} + n_1 - x_1) d\theta_1 \\
&= E_{\theta_1 | x_1} [F^*(g(\theta_1, \rho); \tilde{a} + x_2, \tilde{a} + n_2 - x_2)]. \tag{2.6}
\end{aligned}$$

This again only involves one dimensional integration, which can be computed fairly accurately via numerical integration algorithms available in various statistical softwares like R and SAS. Next the Bayes factor in favor of the alternative hypothesis is defined as the ratio of the posterior odds to the prior odds:

$$\begin{aligned}
& BF(x_1, x_2) \\
&= \frac{Pr[(\theta_1, \theta_2) \in \Theta_1 | X_1 = x_1, X_2 = x_2] / Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2]}{Pr[(\theta_1, \theta_2) \in \Theta_1] / Pr[(\theta_1, \theta_2) \in \Theta_0]} \\
&= \frac{1 - Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2]}{Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2]} \cdot \frac{Pr[(\theta_1, \theta_2) \in \Theta_0]}{1 - Pr[(\theta_1, \theta_2) \in \Theta_0]}, \tag{2.7}
\end{aligned}$$

where $Pr[(\theta_1, \theta_2) \in \Theta_0]$ and $Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2]$ are given in (2.4) and (2.6), respectively. Clearly, the Bayes factor is a non-negative function of x_1 and x_2 , and it is a discrete random variable with its distribution determined by $P(x_1, x_2)$, the unconditional joint distribution of (X_1, X_2) :

$$\begin{aligned}
& P(x_1, x_2) \\
&= Pr[X_1 = x_1, X_2 = x_2] \\
&= \binom{n_1}{x_1} \binom{n_2}{x_2} \frac{B(\tilde{a} + x_1, \tilde{a} + n_1 - x_1) B(\tilde{a} + x_2, \tilde{a} + n_2 - x_2)}{B(\tilde{a}, \tilde{a})^2}. \tag{2.8}
\end{aligned}$$

As a result, every possible value of $BF(x_1, x_2)$ is assigned a corresponding probability $P(x_1, x_2)$ given in (2.8).

Since the Bayes factor here is defined in favor of the alternative hypothesis, large values of $BF(x_1, x_2)$ will be regarded as evidence against the null hypothesis. Therefore, the rejection region of the test could be constructed as $\{(x_1, x_2) : BF(x_1, x_2) > B_0\}$ for some suitable $B_0 > 0$. There are several rules that can be used to determine the critical value B_0 . A popular choice is $B_0 = 1$ and many researchers have advocated the use of Jeffreys'

empirical scale of evidence (see [10]). However such choices are generally not used in practice and more specifically not by regulatory agencies (e.g., FDA) and industries. An alternative approach is to search among all possible outcomes of $BF(x_1, x_2)$ for the largest B_0 that still controls the overall frequentist type I error rate under some desired level. A Bayesian version of the type I error rate of the proposed test can be defined as:

$$\begin{aligned}
& Pr[BF(X_1, X_2) > B_0 | (\theta_1, \theta_2) \in \Theta_0] \\
&= \frac{Pr[BF(X_1, X_2) > B_0, (\theta_1, \theta_2) \in \Theta_0]}{Pr[(\theta_1, \theta_2) \in \Theta_0]} \\
&= \frac{\sum_{x_2=0}^{n_2} \sum_{x_1=0}^{n_1} I_{(B_0, \infty)}(BF(X_1, X_2)) Pr[(\theta_1, \theta_2) \in \Theta_0 | X_1 = x_1, X_2 = x_2] P(x_1, x_2)}{Pr[(\theta_1, \theta_2) \in \Theta_0]},
\end{aligned} \tag{2.9}$$

where $I_A(x)$ denotes the indicator function that equals to 1 if $x \in A$ and 0 otherwise. To compute B_0 , after sorting all possible values of $BF(x_1, x_2)$ given n_1 and n_2 , search can start from the maximum value. The critical value B_0 will be the last value in the sorted sequence that still satisfies $Pr[BF(x_1, x_2) > B_0 | H_0] \leq \alpha$, where $\alpha \in (0, 1)$. Notice that the above definition of the type I error is different from the usual definition of the size of a frequentist test, which is given by $\sup_{\theta \in \Theta_0} Pr[BF(X_1, X_2) > B_0 | \theta]$. However, it easily follows that above defined Bayes type I error is always less than the usual size of the frequentist test for a given B_0 . Thus, given $\alpha \in (0, 1)$ we compute the cut-off value as $B_0(\alpha) = \inf\{B_0 > 0 : Pr[BF(X_1, X_2) > B_0 | \theta \in \Theta_0] \leq \alpha\}$. As a side note, it is more convenient to work on the log scale of the Bayes factor. Since $\{(x_1, x_2) : BF(x_1, x_2) > B_0\}$ is equivalent to $\{(x_1, x_2) : \log(BF(x_1, x_2)) > L_0\}$ with $L_0 = \log(B_0)$, the testing procedure described above remains the same in the log scale.

2.2 Posterior Probability Approach

Another method in Bayesian framework is to directly use the posterior probability of the null or alternative hypotheses. This approach is much straight forward conceptually, that is, the null hypothesis will be rejected if the posterior probability of the null hypothesis falls below some pre-specified threshold. For this approach, prior selection is rather important. Usually, non-informative prior is suggested to minimize subjectivity. Weelek [6] used independent Jeffreys' prior $Beta(0.5, 0.5)$ for each of the proportion parameters. In this section, a different set of non-informative prior is considered. If we let $\eta =$

$\eta(\theta_1, \theta_2, \rho) = \theta_2 - g(\theta_1, \rho)$, then the hypotheses in (2.2) are equivalent to

$$H_0 : \eta \leq 0 \text{ vs. } H_1 : \eta > 0. \quad (2.10)$$

Next we construct a prior for (θ_1, θ_2) in such a way that induces uniform distribution on η . Since $0 < \theta_2 < 1$ and η is monotone function of θ_2 , it follows that $\eta \in (-g(\theta_1, \rho), 1 - g(\theta_1, \rho))$. The prior that we propose for the posterior probability approach is specified as follows

$$\begin{aligned} \theta_1 &\sim Unif[0, 1] \text{ and} \\ \eta|\theta_1 &\sim Unif[-g(\theta_1, \rho), 1 - g(\theta_1, \rho)]. \end{aligned} \quad (2.11)$$

Note that (2.11) leads to a joint prior for (θ_1, η)

$$\pi(\theta_1, \eta) = I_{(0,1)}(\theta_1)I_{(-g(\theta_1, \rho), 1 - g(\theta_1, \rho))}(\eta).$$

Consequently, the likelihood function of θ_1 and η becomes product of two binomial distributions: $X_1|\theta_1 \sim Bin(n_1, \theta_1)$ and $X_2|\theta_1, \eta \sim Bin(n_2, \eta + g(\theta_1, \rho))$. The joint posterior distribution of (θ_1, η) for $\theta_1 \in (0, 1)$ and $\eta \in (-g(\theta_1, \rho), 1 - g(\theta_1, \rho))$ can be written as:

$$\begin{aligned} &f(\theta_1, \eta|x_1, x_2) \\ &= \frac{f^*(\theta_1; x_1 + 1, n_1 - x_1 + 1)f^*(\eta + g(\theta_1, \rho); x_2 + 1, n_2 - x_2 + 1)}{\int_0^1 \int_{-g(\theta_1, \rho)}^{1-g(\theta_1, \rho)} f^*(\theta_1; x_1 + 1, n_1 - x_1 + 1)f^*(\eta + g(\theta_1, \rho); x_2 + 1, n_2 - x_2 + 1)d\theta_1 d\eta}, \end{aligned} \quad (2.12)$$

where f^* is as defined right (2.4) in Section 2.1.

The decision rule is to reject the null hypothesis if the posterior probability of the null

hypothesis $\Pi(x_1, x_2)$ is less than some cut-off value Π_0 , where

$$\begin{aligned}
\Pi(x_1, x_2) &= Pr[\eta \leq 0 | X_1 = x_1, X_2 = x_2] \\
&= \int_0^1 \int_{-g(\theta_1, \rho)}^0 f(\theta_1, \eta | x_1, x_2) d\theta_1 d\eta \\
&= \frac{\int_0^1 \int_{-g(\theta_1, \rho)}^0 f^*(\theta_1; x_1 + 1, n_1 - x_1 + 1) f^*(\eta + g(\theta_1, \rho); x_2 + 1, n_2 - x_2 + 1) d\theta_1 d\eta}{\int_0^1 \int_{-g(\theta_1, \rho)}^{1-g(\theta_1, \rho)} f^*(\theta_1; x_1 + 1, n_1 - x_1 + 1) f^*(\eta + g(\theta_1, \rho); x_2 + 1, n_2 - x_2 + 1) d\theta_1 d\eta}.
\end{aligned} \tag{2.13}$$

The computation of $\Pi(x_1, x_2)$ involves 2-dimensional integration, but it still can be computed numerically using 2-dimensional Gaussian quadrature methods [11] instead of using Monte Carlo methods. Once again we can determine the cut-off value Π_0 which satisfies $Pr[\Pi(X_1, X_2) \leq \Pi_0 | \eta \leq 0] \leq \alpha$. In other words, given $\alpha \in (0, 1)$ we can compute $\Pi_0(\alpha) = \sup\{\Pi_0 \in (0, 1) : Pr[\Pi(X_1, X_2) \leq \Pi_0 | \eta \leq 0] \leq \alpha\}$ by following the similar algorithm described at the end of Section 2.1.

3 Simulation Study

In this section, the performance of the proposed Bayesian tests is investigated and it is compared with the standard Blackwelder type method based on several simulated datasets. Both relative risk and odds ratio based tests are considered. As the reasons will be discussed later, tests under two dissimilarity measures will be examined in slightly different scenarios with regard to the size of θ_1 , which is the success rate of the active control group. For relative risk approach, true values of θ_1 are selected relatively far from the boundaries, while it is set close to 1 for the odds ratio based tests. We have used the R function “integrate” to compute the one-dimensional integrals in (2.4) and (2.6) numerically, whereas the R package “adapt” is used to compute the two-dimensional integrals in (2.13). For more details, the readers may find the corresponding R manuals useful.

3.1 Tests Based on Relative Risk

Data are generated under the model $X_1|\theta_1 \sim \text{Bin}(n_1, \theta_1)$ and $X_2|\theta_2 \sim \text{Bin}(n_2, \theta_2)$. True values of θ_1 are chosen to be $\{0.3, 0.5, 0.8\}$. The true value of the test group proportion $\theta_2 = \eta + \rho\theta_1$, where η is an experimental factor that takes values $\{-0.2, -0.15, -0.1, -0.05, 0.05, 0.1, 0.15, 0.2\}$. Note that the negative values of η imply the null hypothesis is true, while positive values favor alternative hypothesis. The non-inferiority “margin” ρ is set be 0.8706. Next we discuss the motivation behind this choice for ρ .

In practice, selecting an appropriate ρ is very important and there are many alternative values that have been suggested in literature. The main concern is that of the “assay sensitivity” issue mentioned in ICH E10 [4], which states that, the experimental treatment should be superior to placebo even though there’s no placebo arm in non-inferiority trials. As there are extensive literature available targeting this problem, we will not get into this topic in this paper. However, we feel it is necessary to provide readers why $\rho = 0.8706$ is picked for this simulation study. Here we followed the “putative placebo” argument given by Ng [14]. Under the constancy assumption, Ng suggested $\rho = (\theta_0/\theta_1)^\epsilon$, where θ_0 is the putative placebo effect and $(1 - \epsilon)$ is the fraction of the effect size of the active control one desires to preserve for the experiment treatment. Clearly, ϵ should be between 0 and 1 and should be more close to 0 to be on the conservative side. As Ng pointed out, the rationale behind this should be apparent in the log scale. When the non-inferiority is claimed (i.e. the null hypothesis is rejected), we have $\theta_2/\theta_1 > \rho = (\theta_0/\theta_1)^\epsilon$. After taking logarithm, it becomes $\log \theta_2 - \log \theta_1 > (1 - \epsilon)(\log \theta_1 - \log \theta_0)$, which implies the experimental treatment preserves at least $(1 - \epsilon)$ of the standard treatment effect with respect to placebo. Of course, in real applications, the true value of θ_0/θ_1 is unknown, hence one has to use the estimate, which usually comes from the previous placebo control trial for the standard treatment. Therefore, the “putative placebo” approach relies heavily on the validity of the constancy assumption. This problem, however, does not exist for our simulation study for obvious reasons. For further information on the choice of ρ , the readers may find the recent articles [12, 13, 14] useful.

For each value of θ_1 , we set θ_0 to be $\frac{\theta_1}{2}$ and $\epsilon = 0.2$ to preserve at least 80% of the standard treatment effect when claiming non-inferiority, it naturally follows $\rho = (\theta_0/\theta_1)^\epsilon = (0.5)^{0.2} = 0.8706$. Since sample size is an important aspect in clinical trials, it is also included as an experimental factor in our simulation study. For simplicity, sizes of two groups are set equal, i.e., $n = n_1 = n_2 \in \{10, 20, 30, 50, 80, 120\}$. For each

combination of experimental factors, 10,000 replicates were generated, and then analyzed by two Bayesian methods described in Section 2 and the frequentist method using the Blackwelder type test statistic

$$Z_{RR}(x_1, x_2) = \frac{\hat{\theta}_2 - \rho\hat{\theta}_1}{\sqrt{\hat{\theta}_2(1 - \hat{\theta}_2)/n_2 + \rho^2\hat{\theta}_1(1 - \hat{\theta}_1)/n_1}}, \quad (3.1)$$

where $\hat{\theta}_1 = x_1/n_1$ and $\hat{\theta}_2 = x_2/n_2$ are the observed sample proportions. The null hypothesis is rejected if $Z_{RR}(x_1, x_2) > Z_{1-\alpha}$, where α represents the level of the test.

Table 1 goes here**

Results for Monte Carlo (MC) average type I error and power of the three methods are given in Table 1. Only the scenarios with $\theta_1 = 0.8$ are shown here. The level of the Blackwelder testing method is set at 0.05. Note that the α value used to determine the critical values of the Bayesian methods is set to the half of the level used in the Blackwelder approach. The reason for this adjustment is that $\alpha = 0.05$ made the tests based on Bayes factor and posterior probability too liberal when η approaches to 0 from the negative side. When $\alpha = 0.25$, the cutoff values L_0 for the Bayes factor are 1.743, 1.289, 1.005, 0.745, 0.501, and 0.254 for $n=10, 20, 30, 50, 80,$ and 120 , respectively, whereas the corresponding critical values Π_0 for the posterior probability approach are 0.117, 0.161, 0.185, 0.226, 0.230, and 0.339, respectively. Generally, both of the Bayesian methods outperform the frequentist method by a large margin in terms of power. The advantage of Bayesian methods is particularly evident when η is small or n is small. The most remarkable case occurs at $\eta = 0.05$ and $n = 80$, the power of the Bayes factor test is 3 times the power obtained by the frequentist method. The two Bayesian methods appears to have very similar power and type I error rate, although the Bayes factor test has slightly higher power and the posterior probability approach has slightly lower type I error rate. Some inflation of type I error rate was observed when η is -0.05 , but it is controlled at an acceptable level in most cases for both of the Bayesian methods. It also appears that the Blackwelder approach pretects Type I error rate too conservatively as it is much lower than 0.05 in most cases. This may have caused huge loss of power as a result.

In order to compare power and type I error together, the total error (type I + type II) is also examined. In Table 1, we let $\eta_0 \in \{-0.2, -0.15, -0.1, -0.05\}$, which are the negative

η values associated with the type I error, and similarly we let $\eta_1 \in \{0.2, 0.15, 0.1, 0.05\}$. For each combination of (η_0, η_1) , we extract a type I error rate and a power, hence the corresponding total error rate $TR = \text{type I error} + (1 - \text{power})$. Define a metric of (η_0, η_1) as $\Delta\eta = |\eta_0 - \eta_1|$, then the performance of the tests in terms of total error can be evaluated by examining how TR is different among the tests along $\Delta\eta$. This is shown in Figures 1, 2, and 3. Similar patterns to previous ones were observed with respect to the total error. The Bayesian testing procedures have almost uniformly lower total error across all sample sizes and $\Delta\eta$'s. There also appears to be no significant difference between the two Bayesian methods as far as the total error rate is compared.

The results for $\theta_1 = 0.5$ and $\theta_1 = 0.3$ (not shown) are very similar. All patterns mentioned above are consistent across all θ_1 's studied.

3.2 Tests Based on Odds Ratio

As it is pointed out by Ng [14], tests based on relative risk can be problematic when the active control success rate is close to boundary values. In a hypothetical example of this situation given by Ng, suppose $\theta_1 = 0.95, \theta_0 = 0.45, \epsilon = 0.2$, then $\rho = (\theta_0/\theta_1)^\epsilon = 0.8612$. When the null $\theta_2 \leq \rho\theta_1$ is rejected, we only can conclude $\theta_2 > 0.8181$. In the worst case scenario, the failure rate of θ_2 is at most 0.1819, which is more than 3 times the failure rate of θ_1 . Claiming that the treatment 2 is non-inferior to the treatment 1 could be questionable in this case. Therefore, the author suggested using odds ratio as the dissimilarity measure to test non-inferiority when θ_1 is close to 1 or 0. In this section, a simulation is performed for this setup to compare the proposed Bayesian methods and the Blackwelder approach when odds ratio is used to formulate hypotheses.

For $H_0 : \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} \leq \rho$, let $\theta_1 = 0.95$ and $\eta^* = (1 + \rho\theta_1 - \theta_1)\theta_2 - \rho\theta_1 \in \{-.12, -.08, -.04, -.02, .02, .04\}$. Note that the levels of η^* are different from the ones of η in Section 3.1, as a larger value of η^* will result θ_2 exceeding 1. The corresponding θ_2 values to η^* values are $\{0.694, 0.766, 0.838, 0.874, 0.910, 0.946, 0.982\}$. The non-inferiority ‘‘margin’’ is determined in the similar manner as before $\rho = [\text{odds}(\theta_0)/\text{odds}(\theta_1)]^{0.2} = 0.533$, where $\text{odds}(\theta_0) = \theta_0/(1 - \theta_0)$. The same set of sample sizes are used as in Section 3.1. And the Blackwelder-type Wald test statistics is given by

$$Z_{OR}(x_1, x_2) = \frac{\log(\hat{\phi}) - \log(\phi)}{\hat{\sigma}_{or}}, \quad (3.2)$$

where

$$\log(\hat{\phi}) = \log \frac{(x_2 + 0.5)(n_1 - x_1 + 0.5)}{(x_1 + 0.5)(n_2 - x_2 + 0.5)}$$

and

$$\hat{\sigma}_{or}^2 = \frac{1}{x_1 + 0.5} + \frac{1}{n_1 - x_1 + 0.5} + \frac{1}{x_2 + 0.5} + \frac{1}{n_2 - x_2 + 0.5}.$$

Note that 0.5 is added to avoid empty cells that make the test statistics not well defined [7]. The test rejects the null hypothesis if $Z_{OR}(x_1, x_2) > Z_{1-\alpha}$ for a given $\alpha \in (0, 1)$. This adjustment is particularly necessary for the situations in which θ_1 and θ_2 take extreme values such as 0.95. This ad-hoc adjustment can be formally justified by using Jeffreys' priors for θ_1 and θ_2 .

*****Table 2 goes here*****

Power and type I error of the tests based on odds ratio are given in Table 2. Again, the level of the Bayesian tests is set to $\alpha = 0.025$, which is the half of the level used in the frequentist approach. Accordingly, the cutoff values L_0 for the Bayes factor are -.002, -.150, -.141, -.313, -.400, and -.485 for $n=10, 20, 30, 50, 80,$ and $120,$ respectively, and the corresponding critical values Π_0 for the posterior probability approach are 0.112, 0.143, 0.158, 0.205, 0.246, and 0.304, respectively. Among all, the Bayes factor approach has the highest power but also leads to a few cases with inflation of type I error rate when $\eta^* = -0.02$ and $\eta^* = -0.04$. It is followed by the posterior probability test which maintains much acceptable type I error rates. Again, the frequentist method overprotects type I error rate at a cost of substantial loss of power. In terms of the total error rate shown in Figures 4, 5, and 6, both of the Bayesian approaches performs better than the frequentist method across all conditions studied. Between two Bayesian methods, it appears that the posterior probability approach has lower total error rates when the parameters are close to the boundary of two hypotheses while the Bayes factor test does better on the other end.

4 A Case Study: Streptococcal Pharyngitis Trial

As an illustration, we will apply the proposed Bayesian procedures to real data in an example from [15]. In a phase IV trial for streptococcal pharyngitis reported in [16], the study drug clarithromycin was compared to the standard treatment erythromycin.

Compared to erythromycin, the ancillary benefit of clarithromycin is its lower toxicity. Hence the goal is to show its efficacy is non-inferior to the standard treatment. The data are presented in Table 3.

*****Table 3 goes here*****

From the observed data, it appears that the success rate of the active control is close to 1. So odds ratio is used as the dissimilarity measure. The non-inferiority margin is specified as $\rho = 0.5$. Given $n_1 = 107$, $n_2 = 106$, and $\alpha = 0.025$, the critical value of Bayes factor in log scale is $L_0 = -0.456$. For the data $x_1 = 97$ and $x_2 = 98$, the observed log Bayes factor $\log(BF(97, 98)) = 3.39 > L_0$, hence the null hypothesis $H_0 : \frac{\theta_2(1-\theta_1)}{\theta_1(1-\theta_2)} \leq \rho$ can be rejected and we may conclude that clarithromycin is non-inferior to erythromycin. Note that the critical value L_0 is determined to control the Bayesian type I error rate below 0.025, the actual type I error rate for this case stops around 0.007 right before crossing 0.025 benchmark. So even if we set $\alpha = 0.0125$ for the Bayes factor test, the null hypothesis can be rejected. For the posterior probability approach, the cut-off value $\Pi_0 = 0.278$ when $\alpha = 0.025$ and $\Pi_0 = 0.162$ when $\alpha = 0.0125$ as defined at the end of Section 2.2. It turns out that the posterior probability of the null hypothesis $\Pi(x_1, x_2) = 0.040$ for the observed data, again the null hypothesis will be rejected if we use either $\alpha = 0.025$ or $\alpha = 0.0125$. The p-value based on the Blackwelder type test is 0.029 (test statistics $Z_{OR} = 1.894$ as define in (3.2)), which indicates the null hypothesis can not be rejected at the $\alpha = 0.025$ level. Siqueira et al. [17] also analyzed this particular dataset with the same non-inferiority margin $\rho = 0.5$ using the Wald, score, and likelihood ratio tests. The Wald test statistic they used is equivalent to the one defined in (3.2) but without continuity adjustment (i.e., not adding 0.5 to each cell). The p-values for the Wald, score, and likelihood ratio tests are 0.031, 0.027, and 0.031, respectively. Again, none of these frequentist test rejects the null hypothesis at the $\alpha = 0.025$ level. For this example, the Bayesian methods proposed lead to the different conclusion from the frequentist approaches.

5 Conclusions and Discussions

In this paper, two Bayesian alternatives to the classical techniques for testing non-inferiority are provided in the context of two-sample binary data. It is demonstrated that the approaches based on Bayes factor and posterior probability both provide com-

elling improvement in terms of statistical power as compared to the popularly used Blackwelder type testing procedures. Further, the Bayesian approaches have almost uniformly lower total error rate (type I plus type II) than the frequentist method with a very large margin. This in turn indicates that the problem of large sample size, which currently challenges the administration of non-inferiority trials the most, could be alleviated by the Bayesian approaches proposed. The standard frequentist method based on the asymptotic normality appears to over-protect the type I error rate in such a conservative way that it is close to zero in most cases that were explored in this article. Between the two Bayesian methods proposed, there is very little difference in terms of statistical power as well as total error rate.

One advantage of the approaches in this paper over previous Bayesian methods is that its prior selection not only maintains relative objectivity but also accounts for the design parameter that controls the non-inferiority boundary of the particular trial. Additionally, the critical value used in the decision rule is also function of the sample sizes and the non-inferiority margin. In this sense, the selection of prior distribution and cut-off values for the proposed Bayesian methods are design adaptive. Also, the Bayesian approaches described here are constructed for the general form of the hypotheses for testing non-inferiority of two binomial proportion parameters. Therefore, many other possible dissimilarity measures of two proportion parameters can be fitted into this general framework. Although many authors ([6, 14] among others) discussed the choice of the dissimilarity measure from the statistical point of view, we think other practical issues should also be put into the consideration. The choice of the dissimilarity measure links closely to the problem of the determination of the non-inferiority margin (in terms of difference, relative risk, and odds ratio). In most cases, the historical data are the major basis to set the non-inferiority margin, hence the validity of the constancy assumption is crucial to the statistical analysis following the trial. Accordingly, it is necessary at the design stage of the trial to evaluate which dissimilarity measure of the placebo and the active control effects will be most likely to retain constant so that the historical estimate of such measure can be carried along to set non-inferiority margin. This decision of course should be made based on the consensus of statisticians and more importantly by the subject matter clinical experts. As pointed out in [6] and [14], non-inferiority tests based on absolute difference and relative risk have theoretical flaws when the proportion parameter of the active control is close to 0 or 1, thus odds ratio is recommended for these situations. On the other hand, using absolute difference and relative risk do have an advantage that the concept is well understood and appreciated by practitioners. This

aspect facilitates the communication between statisticians and non-statisticians, which is important when determining the non-inferiority margin. Therefore, absolute difference or relative risk as a dissimilarity measure should not be completely ruled out when there is confidence that the active control proportion parameter is far from the boundaries.

Finally, in addition to the prior selection methods presented in this paper, we also tried the empirical Bayes way of determining the prior parameter. In this approach, the prior parameter is selected to maximize the marginal likelihood of (x_1, x_2) . Unfortunately, the sampling properties of the empirical Bayes approach are not found to be very satisfactory, hence its performance is not reported in this article.

References

- [1] D'Agostino, R.B., Massaro, J.M., and Sullivan, L. (2003) Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics, *Statistics in Medicine*, **22**, 169–186.
- [2] Blackwelder, W.C. (1982) “Proving the null hypothesis” in clinical trials, *Controlled Clinical Trials*, **3**, 345–353.
- [3] Farrington, C.P. and Manning, G. (1990) Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk, *Statistics in Medicine*, **9**, 1447–1454.
- [4] ICH E10 (2000). International Conference on Harmonization Guideline: Guidance on Choice of Control Group and Related Design and Conduct Issues in Clinical Trials. Food and Drug Administration, DHHS, July 2000.
- [5] Kaul, S. and Diamond, G.A. (2006) Good Enough: A Primer on the Analysis and Interpretation of Noninferiority Trials, *Annals of Internal Medicine*, **145**, 62–69.
- [6] Wellek, S. (2005) Statistical methods for the analysis of two-arm non-inferiority trials with binary outcomes, *Biometrical Journal*, **47**, 48–61.
- [7] Tu, D. (1998) On the use of the ratio or the odds ratio of cure rates in therapeutic equivalence clinical trials with binary endpoints, *Journal of Biopharmaceutical Statistics*, **8**, 263–282.

- [8] Laster, L.L. and Johnson, M.F. (2006) Non-inferiority trials: “the at least as good as” criterion for dichotomous data, *Statistics in Medicine*, **22**, 187–200.
- [9] Williamson, P.P. (2007) Bayesian equivalence tesing for binomail random variables, *Journal of Statistical Computation and Simulation*, **77**, 739–755.
- [10] Kass, R.E. and Raftery, A.E. (1995) Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795.
- [11] Berntsen, J., Espelid, T.O., and Genz, A. (1991) An Adaptive Algorithm for the Approximate Calculation of Multiple Integrals, *ACM Transactions on Mathematical Software*, **17**,437–451.
- [12] Hung, H-MJ., Wang, S-J., and O’Neill, R.T. (2005) A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials, *Biometrical Journal*, **47**, 28–36.
- [13] Chow, S-C. and Shao J. (2006) On non-inferiority margin and statistical tests in active control trial, *Statistics in Medicine*, **25**, 1101–1113.
- [14] Ng, T.H. (2008) Noninferiority hypotheses and choice of noninferiority margin, *Statistics in Medicine*, **27**, 5392–5406.
- [15] Wellek, S. (2003) *Testing Statistical Hypotheses of Equivalence*. Chapman & Hall/CRC: London.
- [16] Scaglione, F. (1990) Comparison of the clinical and bacteriological efficacy of clarithromycin and erythromycin in the treatment of streptococcal pharyngitis, *Current Medical Research Opinion*, **12**, 25–33.
- [17] Siqueira, A.L., Whitehead, A., and Todd, S. (2008) Active-control trials with binary data: A comparison of methods for testing superiority or non-inferiority using the odds ratio, *Statistics in Medicine*, **27**, 353–370.

Table 1: MC Type I error and Power of Posterior Probability(PP, $\alpha=0.025$), Bayes Factor(BF, $\alpha=0.025$), and Blackwelder(BW, $\alpha=0.05$), $\theta_1=0.8$, the relative risk cut-off $\rho=0.8706$, $\eta=\theta_2-\rho\theta_1$, $H_0 : \eta \leq 0$, $N=10,000$.

η	method	n (per group)					
		10	20	30	50	80	120
-0.2	PP	0.02	0.01	0.00	0.00	0.00	0.00
	BF	0.02	0.01	0.01	0.00	0.00	0.00
	BW	0.01	0.00	0.00	0.00	0.00	0.00
-0.15	PP	0.03	0.02	0.01	0.01	0.00	0.00
	BF	0.03	0.02	0.02	0.01	0.00	0.00
	BW	0.01	0.01	0.00	0.00	0.00	0.00
-0.1	PP	0.05	0.04	0.04	0.03	0.02	0.01
	BF	0.05	0.05	0.05	0.04	0.03	0.02
	BW	0.02	0.01	0.01	0.00	0.00	0.00
-0.05	PP	0.08	0.10	0.10	0.11	0.09	0.10
	BF	0.08	0.11	0.12	0.12	0.11	0.12
	BW	0.04	0.03	0.02	0.01	0.01	0.01
0.05	PP	0.21	0.30	0.35	0.50	0.61	0.72
	BF	0.21	0.31	0.40	0.53	0.64	0.76
	BW	0.12	0.15	0.14	0.17	0.21	0.26
0.1	PP	0.31	0.46	0.55	0.74	0.87	0.95
	BF	0.31	0.47	0.62	0.78	0.89	0.96
	BW	0.20	0.27	0.30	0.40	0.52	0.65
0.15	PP	0.44	0.65	0.77	0.91	0.98	1.00
	BF	0.44	0.65	0.82	0.93	0.98	1.00
	BW	0.31	0.44	0.53	0.68	0.84	0.94
0.2	PP	0.60	0.84	0.92	0.99	1.00	1.00
	BF	0.60	0.84	0.95	0.99	1.00	1.00
	BW	0.47	0.68	0.78	0.91	0.98	1.00

Table 2: MC Type I error and Power of Posterior Probability(PP, $\alpha=0.025$), Bayes Factor(BF, $\alpha=0.025$), and Blackwelder(BW, $\alpha=0.05$), $\theta_1=0.95$, the odds ratio cut-off $\rho=0.533$, $\eta^* = (1 + \rho\theta_1 - \theta_1)\theta_2 - \rho\theta_1$, $H_0 : \eta^* \leq 0$, $N=10,000$.

η^*	method	n (per group)					
		10	20	30	50	80	120
-0.12	PP	0.01	0.01	0.00	0.00	0.00	0.00
	BF	0.10	0.06	0.02	0.01	0.00	0.00
	BW	0.00	0.00	0.00	0.00	0.00	0.00
-0.08	PP	0.03	0.02	0.01	0.00	0.00	0.00
	BF	0.16	0.12	0.08	0.05	0.02	0.01
	BW	0.00	0.00	0.00	0.00	0.00	0.00
-0.04	PP	0.07	0.07	0.04	0.02	0.02	0.02
	BF	0.24	0.23	0.21	0.18	0.14	0.11
	BW	0.00	0.00	0.00	0.00	0.00	0.00
-0.02	PP	0.11	0.12	0.09	0.06	0.07	0.08
	BF	0.28	0.31	0.31	0.31	0.30	0.30
	BW	0.00	0.01	0.01	0.01	0.01	0.01
0.02	PP	0.24	0.33	0.37	0.48	0.76	0.89
	BF	0.37	0.53	0.61	0.73	0.83	0.87
	BW	0.01	0.03	0.06	0.11	0.17	0.25
0.04	PP	0.33	0.52	0.74	0.90	1.00	1.00
	BF	0.39	0.63	0.76	0.91	0.97	0.99
	BW	0.01	0.06	0.13	0.30	0.49	0.69

Table 3: Response counts (rates) for the Streptococcal Pharyngitis Trial [16]

Treatment	Successfully Treated	
	Yes	No
erythromycin	97 (90.7%)	10 (9.3%)
clarithromymin	98 (92.5%)	8 (7.5%)

Figure 1: Total error difference (Posterior Probability approach minus Blackwelder approach), tests based on relative risk, $\theta_1=0.8$, the dashed line denotes 0.

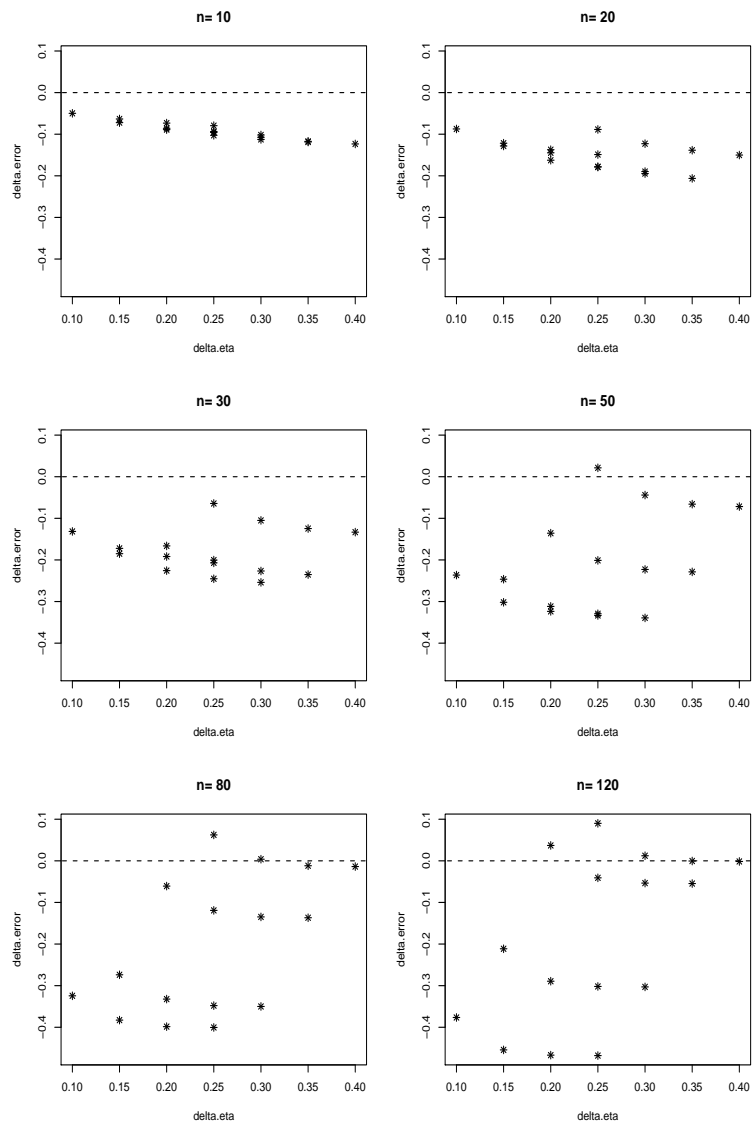


Figure 2: Total error difference (Bayes Factor approach minus Blackwelder approach), tests based on relative risk, $\theta_1=0.8$, the dashed line denotes 0.

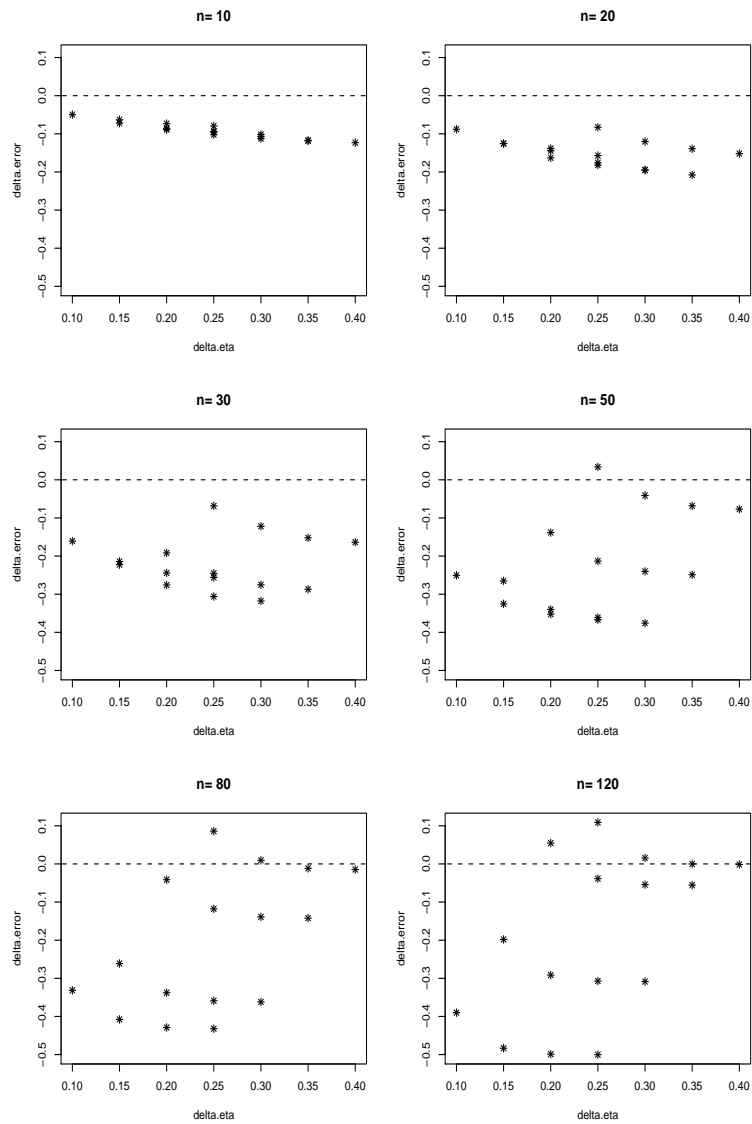


Figure 3: Total error difference (Posterior Probability approach minus Bayes Factor approach), tests based on relative risk, $\theta_1=0.8$, the dashed line denotes 0.

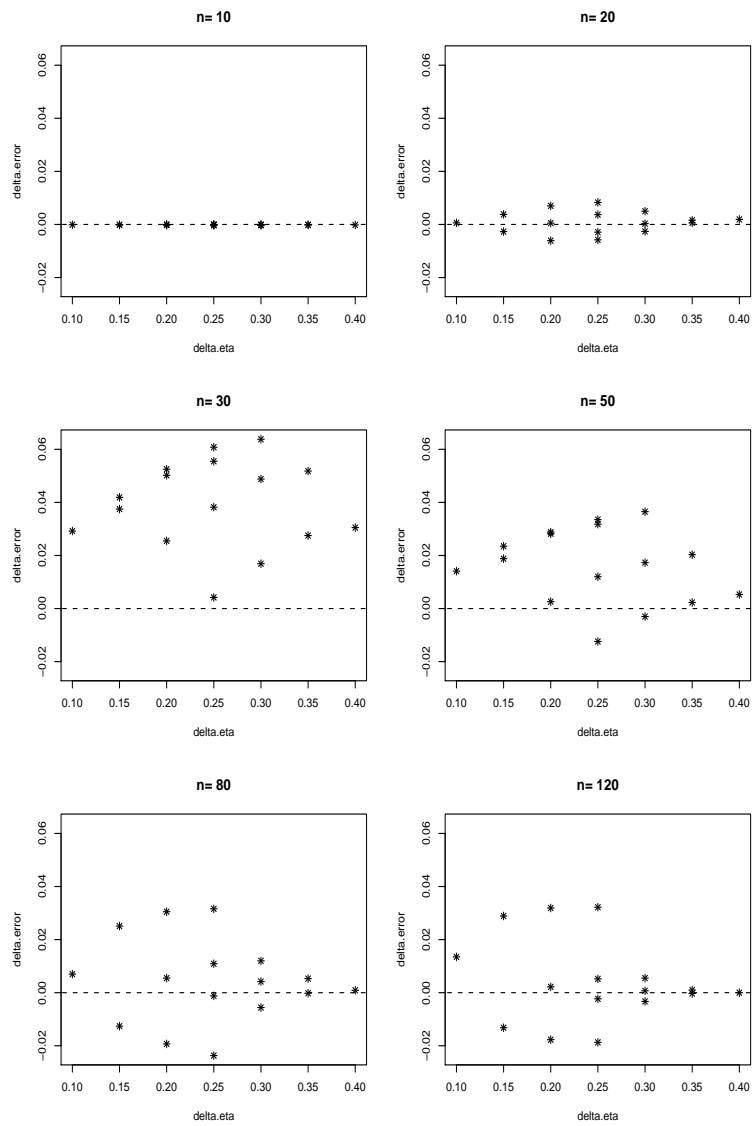


Figure 4: Total error difference (Posterior Probability approach minus Blackwelder approach), tests based on odds ratio, $\theta_1=0.95$, the dashed line denotes 0.

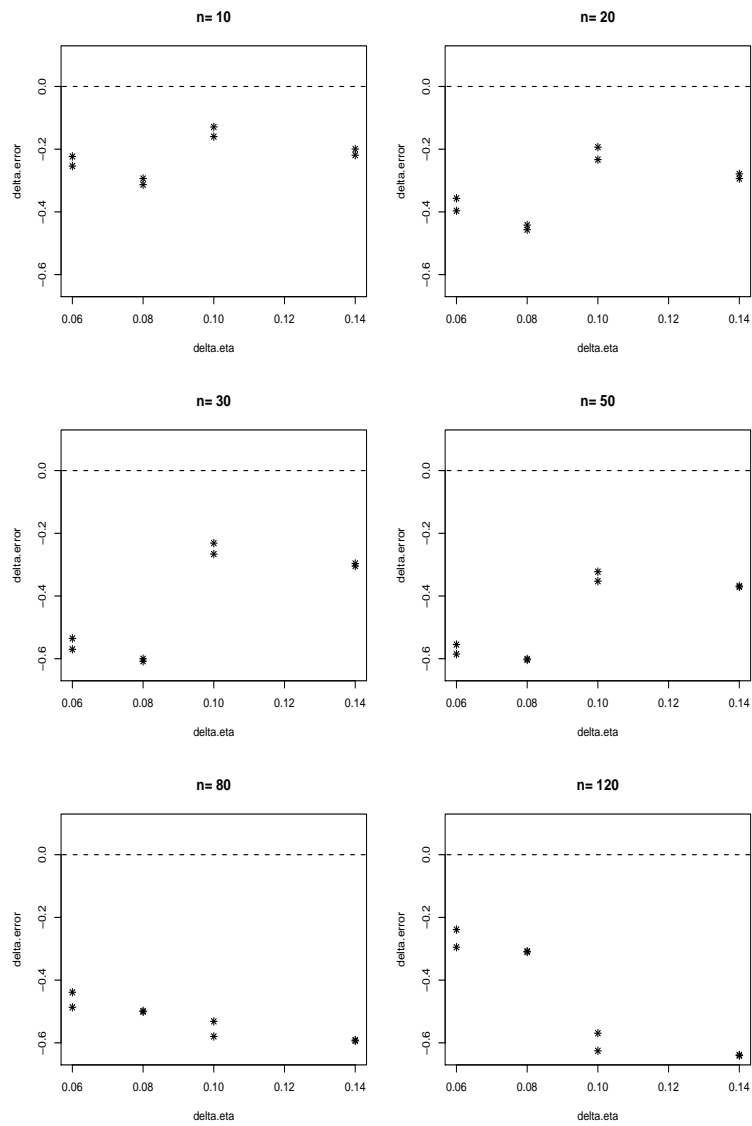


Figure 5: Total error difference (Bayes Factor approach minus Blackwelder approach), tests based on odds ratio, $\theta_1=0.95$, the dashed line denotes 0.

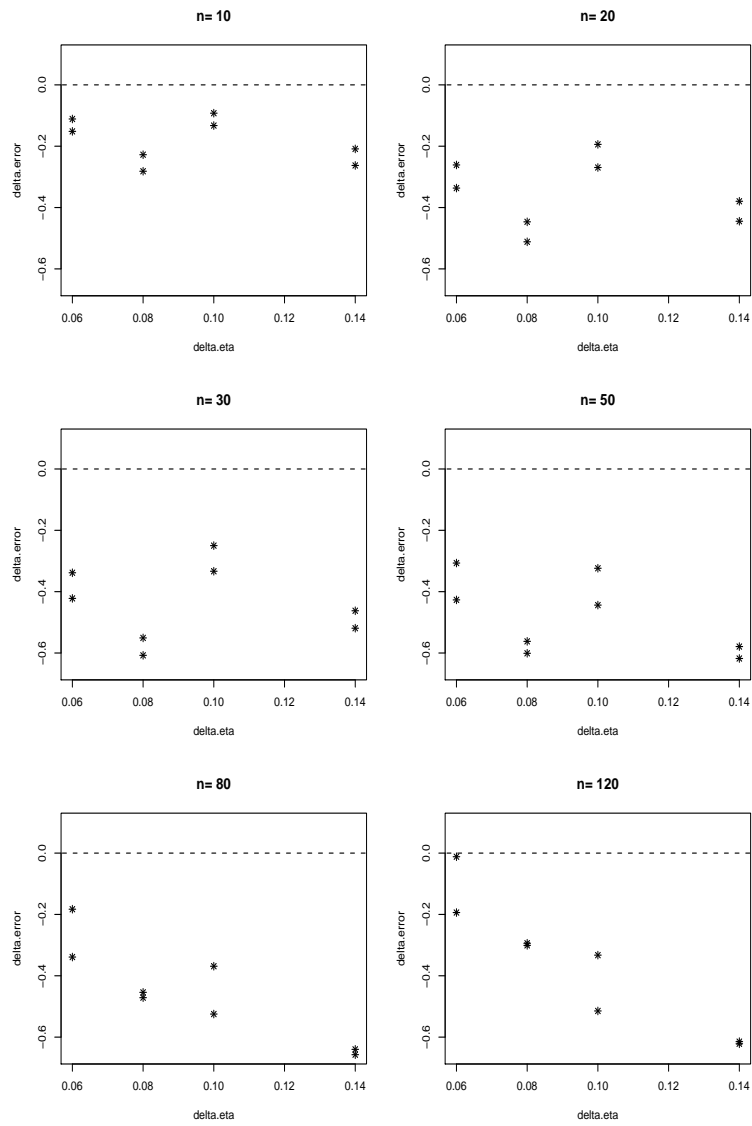


Figure 6: Total error difference (Posterior Probability approach minus Bayes Factor approach), tests based on odds ratio, $\theta_1=0.95$, the dashed line denotes 0.

