Published in final edited form as:

Stat Med. 2016 February 28; 35(5): 695–708. doi:10.1002/sim.6746.

Bayesian approach for assessing non-inferiority in a three-arm trial with pre-specified margin

Samiran Ghosh^{1,2}, Santu Ghosh², and Ram C. Tiwari³

¹Wayne State University, Department of Family Medicine & Public Health Sciences

²Center of Molecular Medicine and Genetics

³Office of Biostatistics, CDER, FDA, USA

Abstract

Non-inferiority trials are becoming increasingly popular for comparative effectiveness research. However, inclusion of placebo arm whenever possible gives rise to a three-arm trial which has lesser burdensome assumptions than a standard two-arm non-inferiority trial. Most of the past developments in a three-arm trial consider defining a pre-specified fraction of unknown effect size of reference drug, i.e. without directly specifying a fixed non-inferiority margin. However, in some recent developments more direct approach is being considered with pre-specified fixed margin albeit in the frequentist setup. Bayesian paradigm provides a natural path to integrate historical as well as current trial information via sequential learning. In this paper, we propose a Bayesian approach for simultaneous testing of non-inferiority and assay sensitivity in a three arm trial with normal response. For the experimental arm, in absence of historical information, non-informative priors are assumed under two situations, namely when a) variance is known and b) variance is unknown. A Bayesian decision criteria is derived and compared with the frequentist method using simulation studies. Finally, several published clinical trial examples are reanalyzed to demonstrate the benefit of the proposed procedure

Keywords

Assay Sensitivity; Bayesian Method; Jeffreys' prior; Markov chain Monte Carlo; Noninferiorty margin

1. Introduction

With the availability of established reference drugs in many disease areas non-inferiority (NI) trials are gaining increasing importance in the recent drug development, where primary aim is to show that a new test (experimental) treatment is not worse than an active control (reference) by more than a specified margin. For example, a trial might seek to determine if an inexpensive and/or easy to administer and/or less toxic test treatment maintains a substantial portion of the effect of the active control [1]. The new treatment though may be slightly inferior to the established active control; however, it can provide an attractive

[†]The article reflects the views of the author and should not be construed to represent FDA's views or policies.

options for patients and thus benefitting the health care system in general. Nevertheless the choice of allowable loss of effectiveness (i.e. the NI margin) is a critical issue. Though regulatory agencies provided broad guidelines [2, 3, 4, 5] on the choice of NI margin, no single consensus solution is reached so far. Hence, the NI trials need to be administered with extreme caution, a detailed discussion of which can be found in the references [6, 4, 7].

One interesting aspect of a two-arm NI trial is that no subjects are exposed to the placebo, which is also ethically attractive. However, apart from the dilemma of NI margin, absence of the placebo arm in the current trial brings the question of “*assay sensitivity*”, i.e. ability to distinguish the effective treatment from a less effective or an ineffective treatment ([3], Chapter 1.5). Hence, a NI trial should not only show comparative effectiveness but it also must demonstrate that active control (or reference drug) has similar effect size that was reported in historical placebo-controlled studies. To alleviate this problem, the inclusion of a placebo arm is recommended in a NI trial by EMA [5] when it is feasible and ethically justified. The resulting three-arm trial is also known as “*gold standard design*” and is preferable due to its greater confidence concerning assay sensitivity (AS) and lesser concern related to external validity.

Both in the two-arm and three-arm trials one critical issue is the choice of NI margin. Pigeot et al. [8] proposed the choice of NI margin as the fraction of the unknown effect size of reference drug, i.e. without directly specifying a fixed non-inferiority margin. This is also known as effect retention hypothesis testing approach. Recently, Hida and Tango [9] considered an approach where the NI margin or δ is defined as a pre-specified fixed difference between treatments in historical placebo controlled trial, and then tested simultaneously for both NI as well as AS hypotheses. While Hida and Tango [9] proposed the same value of δ for testing both the AS and NI, FDA [4] recommends, in many cases, that the desirable magnitude of the NI margin be only a small fraction of the active control effect to ensure a reasonable (neither too restrictive nor too liberal) portion of treatment can be preserved. Kwong et al. [10] proposed a modified version which takes into account the above consideration; we shall henceforth term it as MHT (Modified Hida and Tango) procedure. Nevertheless as mentioned in Hida and Tango [9] the choice of δ should come from past placebo-controlled trials of the reference drug and clinical judgement. Since usage of historical information is an integral part of the NI trial, Bayesian paradigm provides a natural path in which past trial information can be combined with the current trial to glean most possible information. Bayesian methods tend to be flexible and the posterior distribution can be computed efficiently and accurately using simulation based approach. More importantly since we can draw samples directly from the posterior distribution, joint test of NI and AS can be carried out directly without relying on asymptotic assumptions and p-values. Usage of Bayesian approaches in clinical trials and more specifically in the NI trials has a long history and can be found in, e.g. [11, 12, 13, 14, 15, 16], among others. Here, we propose Bayesian version of MHT procedure which uses historical placebo-controlled data (or summary result) to perform Bayesian hypothesis testing in the current trial. We develop methods for both known and unknown variance cases when the outcome of interest is treatment mean (μ). Posterior distributions are derived explicitly for all situations under both informative and non-informative priors depending upon the situation. The simultaneous testing of AS and NI hypotheses is performed on three published clinical trial

datasets with application in Mental Health, Cardiology and Asthma. The results are compared with those from the frequentist approach.

The rest of the article is organized as follows. In the next two Sections we have introduced a novel Bayesian methodology to design and to perform the analysis of a three-arm NI trial. We have considered both known and unknown variance cases separately. Section 4 presents the algorithm for simulation studies. We consider three published clinical trial datasets and apply our proposed methodology in Section 5. We conclude the article with some discussion in Section 6.

2. Hypothesis Testing in a Three-arm NI Trial: MHT Approach

For three-arm trial from the hypothesis testing point of view and the corresponding decision rule three methods are popular. They are namely, 1. Fractional margin, 2. Fixed Margin and 3. Group Sequential decision rules. The first two methods differ in the construction of the NI margin, the most important and pivotal quantity in NI testing. While the last one is an approach when multiple hypotheses need to be tested which has sequential precedence. In a sense the third approach is applicable to both first and second, once we agree on the number of hypotheses that need to be tested, their precedence and most importantly the computation of NI margin. Since our primary goal is to prescribe Bayesian solution for the fixed margin approach we will not deal with the group sequential decision rules in this article. Nevertheless Bayesian group sequential decision rule is also studied though not in the non-inferiority context elsewhere (see [17], [18]). Koch and Röhmel [19] discussed in detail the possible hypotheses that could be tested in a three-arm NI trial with Experiential treatment (E), Reference treatment (R) and Placebo (P). They proposed following hierarchical order of testing, 1. Test $H_0^{EP} : \mu_E \leq \mu_P$, 2. $H_0^{ER} : \mu_E \leq \mu_R - \Delta$ (NI hypothesis), and 3. $H_0^{RP} : \mu_R \leq \mu_P$ (AS hypothesis). They also commented that due to close testing procedure no multiple testing adjustments are needed. They did not comment on choosing NI margin as such. Fractional margin also known as effect retention approach is popularized by Pigeot et al. [8], a Bayesian version of this published by Ghosh et al. [14] recently. In this approach NI margin or Δ is constructed as a function of treatment difference between reference and placebo arms in the current three-arm trial. As mentioned in Pigeot et al. [8], before one constructs such a margin one needs to reject the null hypothesis $H_0^{EP} : \mu_E \leq \mu_P$ concerning reference and placebo as otherwise NI testing does not make much sense. One difficulty in choosing NI margin in this fashion is that NI margin is unknown before the conduct of the current trial and often Δ is defined as a pre-specified difference between treatment and placebo in many situations. It is also not clear how historical information can be used in the construction of NI margin in this approach directly. The method proposed in MHT (see Section 1) one tries to establish jointly, (i) the superiority of the reference and experimental treatments over placebo and (ii) the non-inferiority of the experimental treatment to the reference treatment. Apart from the construction of NI margin the other key difference between Pigeot et al. [8] and MHT [9, 10] approach is the specification of AS hypothesis. While the first one requires simple rejection AS hypothesis, the latter one calls for much stricter or substantial superiority of reference over placebo as specified by the AS margin. We believe in practice both have application, and the acceptability of each depends upon the strictness of assay

sensitivity definition. While simple superiority somewhat validates reference drug's usefulness over placebo in the current trial, the substantial superiority also tests the retention of historical effect-size (i.e. constancy) specified by AS margin via MHT approach.

Following the notations used in Kwong et al. [10], consider the one-way fixed effect model in a three-arm NI trial with Experiential treatment (E), Reference treatment (R) and Placebo (P) given by,

$$X_{ij} = \mu_i + \epsilon_{ij}; i = E, R, P; j = 1, \dots, n_i,$$

where X_{ij} is the j -th response from the i -th treatment, μ_i is the i -th treatment mean and ϵ_{ij} is the random i.i.d. normal error. Without loss of generality we assume that larger μ_i corresponds to better efficacy for the i -th treatment. To compare the three arms in the trial, Hida and Tango [9] proposed simultaneous testing of two null hypotheses H_0 and K_0 corresponding to NI and AS respectively:

$$\text{Null, } H_0: \mu_E \leq \mu_R - \Delta, \quad K_0: \mu_R \leq \mu_P + \Delta, \quad (1)$$

$$\text{Against the alternatives, } H_1: \mu_E > \mu_R - \Delta, \quad K_1: \mu_R > \mu_P + \Delta. \quad (2)$$

We want to reject both the null hypotheses H_0 and K_0 simultaneously. The above two alternative hypotheses H_1 and K_1 can be combined to obtain the inequality:

$$\mu_P + \Delta < \mu_R < \mu_E + \Delta, \quad (3)$$

which if satisfied warrants two objectives, a) AS: the superiority of the R to P by more than Δ and b) NI: the non-inferiority of E with a NI margin of Δ . However, since Δ is used in testing both the NI and AS, the original approach of Hida and Tango [9] seems too liberal. Draft guidance of U.S. FDA [4] mentions that the treatment effect of R should be considered to have a much larger magnitude than the NI margin (Δ). Hence, the inequality (3) is modified by Kwong et al. [10] as,

$$\mu_P + \Delta < \mu_R < \mu_E + r\Delta, \text{ where } 0 < r \leq 1. \quad (4)$$

This corresponds to the following sets of null and alternative hypotheses,

$$\begin{aligned} \text{NI: } H_0: \mu_E \leq \mu_R - r\Delta & \quad \text{vs.} \quad H_1: \mu_E > \mu_R - r\Delta, \\ \text{AS: } K_0: \mu_R \leq \mu_P + \Delta & \quad \text{vs.} \quad K_1: \mu_R > \mu_P + \Delta. \end{aligned} \quad (5)$$

When $r = 1$, (5) reduces to HT, but when $r < 1$ we term (5) as MHT or Modified Hida-Tango procedure which is more stringent (see [10]). Notably in both the HT and MHT is pre-specified and can be determined on the basis of the information from previous/historical trial(s). However, in their papers Hida and Tango [9] and Kwong et al. [10] did not employ historical trials to construct δ , but they rather used a convenient value of δ based on the clinical judgement only. Next, we propose a fixed margin Bayesian approach which takes into account the historical placebo-controlled trial.

3. Bayesian Approach to Three-arm NI Trial

In this section, we develop a Bayesian version of the MHT approach which closely follows frequentist fixed-margin approach. In particular, the NI margin will be determined through the lower bound of the credible interval of the treatment effect of the active control in historical trials. For the two arm NI trial context that precludes placebo arm, a somewhat similar Bayesian approach has been proposed in Gamalo et al. [15]. Though both approaches are Bayesian in spirit, however quite different in terms of trial design (two-arm vs. three-arm) and the hypothesis being tested (only NI testing vs. joint testing of NI and AS). We consider both the known and unknown variance cases for the responses separately. Notably for the unknown variance case the frequentist MHT approach relies solely on large sample property [9].

3.1. Known Variance Case

We first consider a historical placebo controlled trial for the reference treatment. This could very well be the superiority trial used for establishing the efficacy of the reference treatment. Let $X_{ij}^H (i = R, P \text{ and } j = 1, \dots, n_i^H)$ denote the observations corresponding to the j -th response from the i -th treatment in the historical trial. Following the one-way fixed effect model stated earlier,

$$\begin{aligned} X_{Pj}^H | \mu_{PH}, \sigma_{PH}^2 &\sim N(\mu_{PH}, \sigma_{PH}^2), j = 1, \dots, n_P^H, \\ X_{Rj}^H | \mu_{RH}, \sigma_{RH}^2 &\sim N(\mu_{RH}, \sigma_{RH}^2), j = 1, \dots, n_R^H, \end{aligned} \quad (6)$$

where μ_{PH} and μ_{RH} denote the mean responses in the historical trial for placebo and reference treatment and, σ_{PH}^2 and σ_{RH}^2 denote the associated variances, which are assumed to be known. Since the historical trial is not superseded by other trials, we assume non-informative priors for μ_{PH} and μ_{RH} as $\pi(\mu_{PH}) \propto 1$ and $\pi(\mu_{RH}) \propto 1$. The posterior distributions for μ_{PH} and μ_{RH} are given by,

$$\mu_{PH} | \bar{X}_P^H, \sigma_{PH}^2 \sim N\left(\bar{X}_P^H, \frac{\sigma_{PH}^2}{n_P^H}\right); \mu_{RH} | \bar{X}_R^H, \sigma_{RH}^2 \sim N\left(\bar{X}_R^H, \frac{\sigma_{RH}^2}{n_R^H}\right). \quad (7)$$

From the above the posterior distribution of $\mu_{RH} - \mu_{PH}$ can be derived as,

$$\mu_{RH} - \mu_{PH} | \bar{X}_P^H, \bar{X}_R^H, \sigma_{PH}^2, \sigma_{RH}^2 \sim N \left(\bar{X}_R^H - \bar{X}_P^H, \frac{\sigma_{RH}^2}{n_R^H} + \frac{\sigma_{PH}^2}{n_P^H} \right). \quad (8)$$

A threshold value or Δ^B is obtained by solving,

$$\begin{aligned} & P(\mu_{RH} - \mu_{PH} \geq \Delta^B | \bar{X}_P^H, \bar{X}_R^H, \sigma_{PH}^2, \sigma_{RH}^2) \\ &= P \left(Z \geq \frac{\Delta^B - (\bar{X}_R^H - \bar{X}_P^H)}{\sqrt{\frac{\sigma_{RH}^2}{n_R^H} + \frac{\sigma_{PH}^2}{n_P^H}}} | \bar{X}_P^H, \bar{X}_R^H, \sigma_{PH}^2, \sigma_{RH}^2 \right) = 1 - \alpha/2. \end{aligned} \quad (9)$$

This essentially corresponds to the lower limit of the $100(1 - \alpha)\%$ credible interval of the difference of the means $\mu_R - \mu_P$ which is the true effect-size. For the frequentist's setup the

NI margin is explicitly given by $\delta_{NI}^F = (1 - \lambda)\Delta^F = (1 - \lambda) \left(\bar{X}_R^H - \bar{X}_P^H - z_{1 - \alpha/2} \sqrt{\frac{\sigma_{RH}^2}{n_R^H} + \frac{\sigma_{PH}^2}{n_P^H}} \right)$,

where λ is a fraction denoting the preservation level or desired proportion of the control effect to be retained which needs to be guided by clinical judgement. Notably since non-informative priors are used, both the Bayesian and frequentist margins should be same, except for sampling errors. However, it is more challenging to define AS margin in the MHT context. Following the inequality (4), the AS margin can be defined as

$\delta_{AS}^B = \delta_{NI}^B / r = \frac{(1 - \lambda)\Delta^B}{r}$. For all practical purpose $r \in [1 - \lambda, 1]$ as when $r < (1 - \lambda)$ the AS margin asks for even stricter rejection criteria than the historical placebo controlled trial, and for $r = 1$, MHT reduces to Hida-Tango procedure [9]. For the examples discussed in the application section, we let $r = 1 - \lambda$, which boils down to choosing the full Δ^B as the AS margin, and $(1 - \lambda)$ fraction of the Δ^B as the NI margin. It is possible to choose other values of r in practice however those choices should be guided by practical as well as clinical justifications.

3.1.1. Posterior Distributions in Current Trial: Let us now consider the current NI trial consisting of three arms E, R and P. Since no information is available for the experiential treatment, we assume a non-informative prior for the experimental treatment effect i.e. $\pi(\mu_E) \propto 1$. On the other hand, since constancy is assumed for active control and placebo, we would like to use the posterior distributions of μ_R and μ_P obtained from the historical trial as their informative priors in the current trial. This informative prior setup justifies the meaningful interpretation of the NI trial. As noted in Gamalo et al. [16], if

several historical placebo control trials are available a meta-analysis could be carried out first to construct the prior distributions for μ_R and μ_P in the current trial. The posterior distributions for μ_E , μ_R and μ_P are,

$$\mu_E \propto N\left(\bar{X}_E, \frac{\sigma_E^2}{n_E}\right); \mu_l \propto N(\tilde{\mu}_l, \tilde{\sigma}_l^2) \text{ for } l \in \{R, P\}, \quad (10)$$

respectively, where

$$\tilde{\mu}_l = \tilde{\sigma}_l^2 \left(\frac{n_l \bar{X}_l}{\sigma_l^2} + \frac{n_l^H \bar{X}_l^H}{\sigma_{lH}^2} \right) \text{ and } \tilde{\sigma}_l^2 = \left(\frac{n_l}{\sigma_l^2} + \frac{n_l^H}{\sigma_{lH}^2} \right) \text{ for } l \in \{R, P\}. \quad (11)$$

Also the posterior distributions of $(\mu_E - \mu_R)$ (for NI) and $(\mu_R - \mu_P)$ (for AS) follow normal distribution given by,

$$(\mu_E - \mu_R) | \bar{X}_E, \bar{X}_R, \sigma_E^2, \sigma_R^2 \propto N\left(\bar{X}_E - \tilde{\mu}_R, \frac{\sigma_E^2}{n_E} + \tilde{\sigma}_R^2\right), \quad (12)$$

$$(\mu_R - \mu_P) | \bar{X}_R, \bar{X}_P, \sigma_R^2, \sigma_P^2 \propto N(\tilde{\mu}_R - \tilde{\mu}_P, \tilde{\sigma}_R^2 + \tilde{\sigma}_P^2). \quad (13)$$

Since both the distributions in (12) and (13) are in standard forms, sampling from them is rather straight forward. However, it should be noted that the posterior distributions of $\mu_E - \mu_R$ and $\mu_R - \mu_P$ in the above equation are not independent. The variance-covariance matrix of their joint bivariate normal distribution is given by,

$$\Sigma = \begin{bmatrix} \sigma_E^2 + \tilde{\sigma}_R^2 & -\tilde{\sigma}_R^2 \\ -\tilde{\sigma}_R^2 & \tilde{\sigma}_R^2 + \tilde{\sigma}_P^2 \end{bmatrix}.$$

For a specified NI margin (δ_{NI}^B) and preservation level (λ) the decision criterion is as follows. The experimental treatment is non-inferior to the active control with assay sensitivity at δ_{NI}^B/r if,

$$\begin{aligned}
 p(\delta_{NI}^B, \lambda, r, 1 - \alpha/2) &= P(\mu_E - \mu_R \geq -\delta_{NI}^B \cap \mu_R - \mu_P \geq \delta_{NI}^B/r | \bar{X}_E, \bar{X}_R, \bar{X}_P) \\
 &= P \left[Z_1 \geq \frac{-\delta_{NI}^B - (\bar{X}_E - \bar{\mu}_R)}{\sqrt{\frac{\sigma_E^2}{n_E} + \tilde{\sigma}_R^2}} \cap Z_2 \geq \frac{\delta_{NI}^B/r - (\bar{\mu}_R - \bar{\mu}_P)}{\sqrt{\tilde{\sigma}_R^2 + \tilde{\sigma}_P^2}} \right] \\
 &\geq p^*.
 \end{aligned} \tag{14}$$

Here, r is a fraction indicating the proportion of the NI margin to be used in the assay sensitivity, Z_1 and Z_2 stand for standard normal random variables, and p^* is a pre-specified value such as $p^* = 0.95, 0.975$ or any other clinically reasonable value chosen from the level (α) condition which we describe in further detail in Section 4.2.1. The equation (14) is our proposed Bayesian decision rule for jointly establishing the NI with AS.

3.2. Unknown Variance Case

For the historical placebo-controlled trial, we assume Jeffreys' non-informative prior for the variance σ_{IH}^2 given by, $\pi(\sigma_{IH}^2) \propto \sigma_{IH}^{-2}$, $I \in \{R, P\}$. The prior distributions for the means μ_{IH} , $I \in \{R, P\}$ remain the same as in Section 3.1. Under this setup, the joint posterior of $(\mu_{IH}, \sigma_{IH}^2)$ is a product of the Normal distribution and Jeffreys' prior given by the conditional posterior of σ_{IH}^2 and the marginal posterior of σ_{IH}^2 ; i.e.

$$\mu_{IH} | \sigma_{IH}^2, \bar{X}_l^H \sim N\left(\bar{X}_l^H, \frac{\sigma_{IH}^2}{n_l^H}\right); \sigma_{IH}^2 | \bar{X}_l^H \sim IG\left(\frac{n_l^H - 1}{2}, \frac{n_l^H - 1}{2} s_{lH}^2\right).$$

With some standard algebra it can be shown that the marginal posterior distribution of μ_{IH} , $I \in \{R, P\}$, is given by,

$$\mu_{IH} | \bar{X}_l^H \sim t_{n_l^H - 1}\left(\bar{X}_l^H, \frac{s_{lH}^2}{n_l^H}\right); s_{lH}^2 = \frac{1}{n_l^H - 1} \sum_{j=1}^{n_l^H} (X_{jl}^H - \bar{X}_l^H)^2 \text{ where } l \in \{R, P\}, \tag{15}$$

where $t_\nu(\mu, \frac{\sigma^2}{s})$ denotes the t -distribution with ν degrees of freedom, with μ and $\frac{\sigma^2}{s}$ as the location and scale parameters, respectively. To determine the NI and AS margins using

Bayesian approach it is essential to compute the distribution of $(\mu_{RH} - \mu_{PH} | Data)$ which is not available in a closed form. However, we can generate Markov chain Monte Carlo (MCMC) samples from (15) from the individual t -distributions which can then be used to compute the Bayesian NI margin (B) using (9). For the frequentist case, Gamalo et al. [15] recommended t -distribution for evaluation of F , which is only valid in asymptotic sense via Welch's approximation [20].

3.2.1. Posterior Distributions in Current Trial: We use the posterior distribution of means and variances of the reference and placebo arms from the historical trial as the informative priors for the parameters in the current trial. That is, the joint prior for (μ_l, σ_l^2) in the current trial be specified by,

$$\pi(\mu_l | \sigma_l^2) \propto N\left(\bar{X}_l^H, \frac{\sigma_l^2}{n_l}\right); \pi(\sigma_l^2) \propto IG\left(\frac{n_l^H - 1}{2}, \frac{n_l^H - 1}{2} s_{lH}^2\right) \text{ where } l \in \{R, P\}. \quad (16)$$

With some algebra it can be shown that the marginal posteriors for R and P -arms are given by,

$$\begin{aligned} \mu_l | \bar{X}_l &\propto t_{n_l^*} \left(\mu_l^*, \frac{\psi_l^2}{K_l} \right), l \in \{R, P\} \\ \psi_l^2 &= \frac{1}{n_l^*} \left[\frac{n_l^H - 1}{2} s_{lH}^2 + (n_l - 1) s_l^2 + \frac{n_l n_l^H}{n_l + n_l^H} (\bar{X}_l - \bar{X}_l^H)^2 \right], \\ \mu_l^* &= \frac{1}{K_l} (n_l^H \bar{X}_l^H + n_l \bar{X}_l), \\ K_l &= n_l^H + n_l, \\ n_l^* &= \frac{n_l^H - 1}{2} + n_l. \end{aligned} \quad (17)$$

For the experimental arm (E), since no prior data is available, we use the non-informative priors for the mean and variance. This yields the marginal posterior for mean (μ_E) in experimental arm as,

$$\mu_E | \bar{X}_E \sim t_{n_E - 1} \left(\bar{X}_E, \frac{s_E^2}{n_E} \right); s_E^2 = \frac{1}{n_E - 1} \sum_{j=1}^{n_E} (X_{jE} - \bar{X}_E)^2. \quad (18)$$

As in the known variance case the difference of t -distributions for NI and AS testing is not in a closed form, and we use a MCMC sampling based approach. Suppose we generate M

posterior samples in the current trial, then the general form of the Bayesian decision rule to declare NI with AS is,

$$\begin{aligned}
 p(\delta_{NI}^B, \lambda, r, 1 - \alpha/2) &= P(\mu_E - \mu_R \geq -\delta_{NI}^B \cap \mu_R - \mu_P \geq \delta_{NI}^B / r | Data) \\
 &= \frac{1}{M} \sum_{m=1}^M [I(\mu_E^m - \mu_R^m \geq -\delta_{NI}^B) \cap I(\mu_R^m - \mu_P^m \geq \delta_{NI}^B / r | Data)] \\
 &\geq p^*,
 \end{aligned} \tag{19}$$

where p^* is pre-specified. such as $p^* = 0.975, 0.95$ or any other clinically reasonable values. In the simulation study, we used 0.975. We shall refer to equation (19) as the proposed Bayesian decision criterion.

4. Simulation Study

In this section, we enumerate few simulation studies to benchmark the performance of the above Bayesian procedure in comparison to the frequentist approach. Numerical results are then presented in Tables 1–2.

4.1. Simulation Steps

The following simulation steps are used to calculate the power and type-I error for unknown variance case. Note that known variance case can be followed in similar fashion.

1. Specify $n_P^H, n_R^H, \mu_{PH}, \mu_{RH}, \sigma_{PH}, \sigma_{RH}, n_P, n_R, n_E, \mu_P, \mu_R, \mu_E, \sigma_P, \sigma_R, \sigma_E$ where $\mu_{PH} < \mu_{RH}$ in order to generate X_P^H, X_E^H .
2. Generate the data from $X_I^H \sim N(\mu_{IH}, \sigma_{IH}^2)$, for $I \in \{P, R\}$ and calculate the sample means $(\bar{X}_P^H, \bar{X}_R^H)$ and sample variances (s_{PH}^2, s_{RH}^2) .
3. Use Gibbs sampling method to generate posterior samples from the historical trial as specified in equation (15). Set the number of MCMC iterations to be N .
4. Calculate the difference $(\mu_R^H - \mu_P^H)$ for N pairs and sort the results from smallest to largest and then pick the $\alpha N/2$ -th sample point as the Bayesian threshold δ_{NI}^B .
5. Let μ_E be less than $\mu_R - \delta$ (for calculating type-I error) and generate data $X_I \sim N(\mu_I, \sigma_I^2)$ for $I \in \{P, R, E\}$. Note that for a specified preservation level λ , $\delta_{NI}^B = (1 - \lambda)\Delta^B$ for the Bayesian setting and $\delta_{NI}^F = (1 - \lambda)\Delta^F$ for the frequentist setting.
6. From the current trial generate M posterior samples using the conditional posterior specified in equation (16–18).

7. Compute the difference of posterior samples for $(\mu_E - \mu_R)$ and $(\mu_R - \mu_P)$ for M pairs.
8. Calculate the posterior probability

$$p(\delta_{NI}^B, \lambda, r, 1 - \alpha/2) = \frac{1}{M} \sum_{m=1}^M \left[I(\mu_E^m - \mu_R^m \geq -\delta_{NI}^B) \cap I(\mu_R^m - \mu_P^m \geq \delta_{NI}^B/r | X_P, X_E, X_R) \right]$$

9. Bayesian criterion: If $p(\delta_{NI}^B, \lambda, r, 1 - \alpha/2) \geq p^*$, increase the COUNTS by 1; otherwise 0.
10. Go back to to Step 2 and repeat the simulation n^* times. Calculate power by using the COUNTS divided by n^* .

Note that for the frequentist procedure, Step 9 needs to be replaced by the frequentist decision criteria such as the one based on comparison of p-values with α . Since multiple testing is involved, this step requires additional discussion. Kwong et al. [10] proposed controlling for FWER however we will elaborate against it in the next section.

4.2. Simulation Results

We first determine the number of MCMC samples required to have stable results prior to conducting simulation studies related to the type-I error. A trace plot of posterior estimator for each parameter suggests that $N = M = 10,000$ MCMC samplers with 1000 burn-ins are sufficient to produce a stable Bayesian estimate of average posterior probability.

In the subsequent experiment, we consider $\mu_P^H = 0$, $\mu_R^H = 1$, $\sigma_{P,H}^2 = 1$, $\sigma_{R,H}^2 = 1$, $\sigma_P^2 = 2$, $\sigma_R^2 = 1$ and $\sigma_E^2 = 1.5$. Throughout this simulation study, we set $\mu_E = \mu_R - \delta_{NI}^B + \xi_1$ and $\mu_P = \mu_R - \delta_{NI}^B/r - \xi_2$, where ξ_1 and ξ_2 (> 0) are constants. Following the discussion of Section 3, it is possible to consider various possible values of ξ_1 and ξ_2 , with the restriction that $\xi_1 \geq \xi_2$. However, to reduce the huge combinatorial burden of different possible values of ξ_1 and ξ_2 , we set $\xi_1 = \xi_2/2$ for all exposition. We choose $r = 1 - \lambda$. This specific choice of r is the most conservative AS margin obtained from the historical two-arm trial as discussed earlier in Section 3.1. If it clinically makes sense, it is possible to choose other values of $r \in [1 - \lambda, 1]$ which will make the AS margins less wider resulting in less conservative decisions procedures.

As pointed out by one of the referee, any NI trial (including both two-arm and three-arm) makes an important assumption that generally historical comparisons of active controls with placebo resulted in a significant advantage, as otherwise the reference drug would have never established. This in turn also assumes that the possibility of no trials being reported where such superiority failed to establish. This can happen due to many reason, however the most important being the publication bias against negative results. Similar problem also prevalent in the meta-analysis literature [21, 22]. Since only positive trials are used to set the NI margin, there is a chance that the NI margin is based on an over-estimate of the true (but unobserved) treatment effect. In the current approach instead using this (possibly biased)

treatment difference directly a lower 95% confidence interval (credible interval for Bayesian case) is chosen as NI margin. Additionally introduction of the preservation level (λ) in the NI margin also somewhat controls this overestimation bias. However to demonstrate this problem and following one of the referees suggestion we have done two different simulation studies. We first considered the situation when historical comparison that resulted significant advantage of reference to placebo with an average p-value $\ll 0.001$ (i.e. Average p-value: $5.072588e-07$ and SD: $2.019145e-05$). Next we consider another simulation when historical reference is not so strongly established compared to placebo with an average p-value $\in [0.001, 0.05]$ (i.e. Average p-value: 0.028 and SD: 0.068). Figure 1 shows the distribution of the width of NI margin under these two different scenarios. Note that the width of NI margin is proportional to the observed effect size in historical trial.

4.2.1. The Issue of Multiple Testing—Since we want to test both the NI and AS hypotheses simultaneously, this is a multiple testing situation. However NI with AS is only accepted when we reject both H_0 and K_0 . This clearly falls in the framework of Intersection-Union (IU) testing setup. An interesting feature of intersection-union tests is that no multiplicity adjustment is necessary to control the size of a test. In fact Berger [23] shown that IU test has size at-most α . However there exists some degree of reservation against IU testing. Dmitrienko et al. [24] mentioned that, “Intersection-union tests are sometimes biased in the sense that their power function can drop below their size in the alternative space”. They also point out that IU test tends to be very conservative, for further references on this we refer to Dmitrienko et al. [24] and the references therein. Nevertheless we would like to mention our main goal in this article is not to improve IU testing rather it comes as a byproduct of joint testing of NI with AS. Albeit in case a better approach is found for type-I error adjustment under IU testing setup, one can easily replace this to calculate alternative type-I error without altering the prescribed Bayesian (and frequentist) approach prescribed in this article.

4.2.2. Power Comparison: For the power comparison, we let ξ_2 vary from 0.2 to 0.8 by the increments of 0.2 for different choices of λ and set $\xi_1 = \xi_2/2$. Note that we also fixed $r = 1 - \lambda$ as indicated earlier. For Bayesian power calculations we obtained the MCMC samples from the joint distribution of (12) and (13) directly. No multiplicity adjustment is made for both Bayesian and frequentist approach a joint testing of NI and AS hypotheses follows IU ($H_0 \cup K_0$ vs. $H_1 \cap K_1$ in MHT and also in Stucke and Kieser [25]) testing, as we discussed previously. However, a pitfall of this non-adjustment is that the individual hypotheses cannot be tested at levels higher than the nominal significance level either (see Berger [23]). There exist considerable debate on this issue and moreover whether the non-adjusted test (MIN test of Laska and Meisner [26]) is more powerful is not well accepted. Dmitrienko et al. [24] also reported that the MIN test is very conservative because of the requirement that all hypotheses must be rejected at level α . As a result some adjustments to the level is recommended by many researchers (see relaxed IU test by Deng et al. [27]). As mentioned in the subsection 4.2 we did not proceed further in modifying the definition of type-I error in this article. For Bayesian decision rule we choose $p^* = 0.975$, to mimic its’ frequentist counterpart. As mentioned in Gamalo et al. [28] this traditional choice of p^* could be very conservative and may lead to very restricted type-I error. For this reason they

suggested doing Bayesian calibration, which generally selects a smaller value of p^* , albeit keeping type-I error still less than 0.025. This results in higher power of the calibrated Bayesian test. However following referee suggestion we did not calibrate p^* in the current article, which could be done if deemed necessary and will result in significant power gain in our experience. Table 1 shows the power for varying sample size, preservation level (λ), and effect size (ξ) when the historical trials strongly establishes Active Control against placebo which generally results in larger effect size. Table 2 shows the power when the historical trials not strongly establishes Active Control against placebo thus resulting in smaller effect size. For both cases the Bayesian approach is either equal or is more powerful than the frequentist approach. Though for larger effect size this difference is non-significant which is as expected. Simulation results for the known variance case are not shown here as both the methods perform similarly for the chosen sample sizes.

5. Application To Three Drug Datasets

We tested our proposed method on three published datasets coming from different disease areas. Each of these studies have been considered earlier in the non-inferiority context. Though we should mention the exact subsets of these datasets considered earlier may not be the same as ours, as different studies use different inclusion-exclusion criteria as well as number of other modifications suitable for their specific purposes. However, as depicted next the overall results obtained from our analysis are comparable with the previous published articles.

5.1. Study 1: Depression Trial

The first dataset is coming from a depression trial described in Higuchi et al. [29]. This specific study was also considered in the paper by Hida and Tango [9], where they proposed a frequentist version of the problem that we are considering here. The primary objective of the study is to compare the efficacy and safety of duloxetine (E) with those of paroxetine (R) and placebo (P). This is a double-blinded, randomized, active-controlled, parallel-group study of a 6-week treatment with duloxetine ($n_E = 147$) or paroxetine ($n_R = 148$) or placebo ($n_P = 145$). The primary endpoint is the change in HAMD-17 total score at the end of 6-week. The mean decrease in HAMD-17 total score is 10.2 ± 6.1 for duloxetine group, 9.4 ± 6.9 for those in paroxetine group, and 8.3 ± 5.8 for placebo group. In their article Hida and Tango [9] took slightly different approach by specifying NI margin beforehand from the clinical standpoint. To compare the Bayesian and frequentist approaches, we use a noninformative prior with unknown variance to obtain δ_{NI}^B and we also obtain δ_{NI}^F . Hida and Tango [9] also established the non-inferiority; however, showed the lack of assay sensitivity. We report our result in Table 3. We also report the average posterior probability of the simultaneously rejection of H_0 and K_0 for different Bayesian margin. Since non-informative prior is used in the historical trial, both the Bayesian and frequentist margins are very close. Note that the Bayesian results are dependent on the value of p^* in use. Table 3 used non-calibrated $p^* = 0.975$ which also results in very conservative type-I error (< 0.01). If calibrated, $p^* = 0.78$ would grantee overall type-I error < 0.025 . When compared to the reported Posterior probability in Table 3 this will change all “0” in AS Decision to “1” which will act as the major difference between the Bayesian and frequentist procedure. We

have also reported the parameter estimates and 95% credible intervals for the mean effects using the Bayesian approach in Table 4. The difference in decision in testing AS hypothesis between the Bayesian and frequentist approaches become clear when we compare the 95% confidence interval of $\mu_R - \mu_P$ reported as $[-0.07, 2.64]$, with the corresponding Bayesian credible interval for the same reported as $[0.72, 2.79]$. Since we have used $r = 1 - \lambda$ throughout for the AS margin, it is not affected by the changing λ values, as a result the frequentist margin ($\delta^F = 0.732$) is always inside the 95% confidence interval, while the Bayesian margin ($\delta^B = 0.716$) falls outside the credible interval as the joint posterior probability of rejecting $H_0 \cup K_0$ is always higher than the calibrated p^* .

5.2. Study 2: Home-Based Blood Pressure Intervention Study

This trial is reported first in Ghosh et al. [14]; however, apart from this being a general motivating example no serious data analysis was conducted. In their paper Pezzin et al. [30] reported the result of this trial albeit using a traditional frequentist approach. It is a three-arm comparative study on the effectiveness of organizational interventions at improving the blood pressure (BP). Two interventions to be tested are (i) a basic intervention delivering “just-in-time” information to the nurses, physicians, and patients while the patient is receiving traditional post-acute home health care, and (ii) an “augmented” intervention transitioning patients to a Home-based hypertension Support Program that extends the information, monitoring and feedback available to patients and primary care physicians for an 18-month period beyond an index home care admission. Usual care is included as a third arm. The primary goal is to see if the basic intervention (E) is at least as good as the augmented one (R), relative to the usual care (P). The unit of randomization is caregivers (nurses). The study involved a total of 845 newly admitted patients with uncontrolled hypertension (HTN). A complete follow-up sample size of 525 patients were divided roughly equally across groups. Primary study endpoint is at 3-months after the beginning. The primary outcome variable was the decrease from baseline in Systolic Blood Pressure (SBP). Note that unlike the rest of the paper, for this example a lower value of mean is considered as improvement. As a result for this specific example, the directions of NI and AS hypotheses are modified accordingly. Unadjusted SBP outcome at 3-Month follow-up are: For Usual care, 160.5 ± 25.3 ; for Basic care, 158.7 ± 25.4 ; and for Augmented care, 152.5 ± 26.2 . More details of the study design description can be found in [30]. Similar to the previous example, we used the noninformative prior with unknown variance to construct δ_{NI}^B and δ_{NI}^F . As can be seen from Table 5, in this specific example, both the Bayesian and frequentist decisions are in complete agreement irrespective of Bayesian calibration. Both the methods failed to declare the non-inferiority of the basic intervention though both the methods established the assay sensitivity.

5.3. Study 3: Mild-Asthma Study

Our third example is taken from Pigeot et al. [8] paper describing a “Mildly Asthmatic Patients”, where the primary outcome variable is FVC (forced vital capacity). The dataset consists of experimental ($n_E = 35$), reference ($n_R = 19$) and a placebo ($n_P = 20$) groups. The summary measure for the primary outcome variable, FVC, are as follows: for E group 4.32 ± 1.16 , for R group 4.86 ± 1.03 , and for placebo group 3.14 ± 0.97 . However, Pigeot et al.

[8] did not describe any further details about the study citing the fact that the experimental treatment was under review. In the absence of the actual trial data, we simulated 5,000 separate datasets from normal distribution each with specified parameter values and sample size. The results of the Bayesian and frequentist analyses are reported in Table 7. For the frequentist approach NI cannot be declared, though assay sensitive is established for all λ . This was also reported in [8], albeit no test for assay sensitivity was carried out. For the Bayesian method the NI testing is dependent on the λ values being used. It can be seen from the table that for smaller values of λ , which correspond to larger margin size (δ^B), NI is accepted. However, when $\lambda = 0.25$ the NI cannot be established. Like the frequentist approach, the assay sensitivity is accepted throughout. If Bayesian calibration is used the only difference is that for $\lambda = 0.25$ the NI decision would change from “0” to “1” under Bayesian setup (as $p^* = 0.78 < 0.89$). This data set shows an interesting example where non-inferiority is clearly dependent upon the choice of margin which in turn depends upon the clinically acceptable preservation level.

6. Discussion

The usage of historical information is critical in non-inferiority trial. Taking this into consideration, in this article we have developed a fully Bayesian approach for testing three-arm NI trials, which avoids the inherent ambiguity of assay sensitivity and external validity. We considered endpoints consisting of normal means, with both known and unknown variances. Non-informative priors are adopted for the mean and variance in historical trials, albeit our setup is flexible enough to accommodate informative priors if available from historical trial/trials. We compared our Bayesian approach with the corresponding frequentist setup in simulation studies. The Bayesian method performs favorably both with and without calibrated, and is flexible enough to accommodate heterogeneity. Another advantage of the Bayesian approach is that instead of reporting the p-values and confidence intervals, the full posterior distributions of $\mu_E - \mu_R$ and $\mu_R - \mu_P$ can be used to characterizing the behavior as a function of effect size. Though we have considered three-arm non-inferiority trials exclusively, our developed Bayesian method can be easily adapted for three-arm equivalence trials that deal with placebo along with experimental and reference drugs [31]. Equivalence trials that deal with generic drugs fall in this setup. However since the null and alternative hypotheses of the equivalence trial are slightly different than that for a non-inferiority trial, appropriate adjustments are necessary for Bayesian decision rule to reflect that. The Intersection-Union testing setup is common to both equivalence and three-arm NI trial that aims to test NI with AS jointly. A disadvantage of IU test is over-conservative type-I error. A remedy could be by adopting a modified definition of the error rate [24].

We also applied both the Bayesian and frequentist methods on three different clinical trial data. The results suggest that the Bayesian method perform favorably in all situations, and that this method does not depend on any asymptotic approximation like frequentist method. We are currently developing Bayesian methods for the binary and survival type endpoints where similar principle is applicable however their implementation will offer substantial new challenge.

Acknowledgement

The research of first author is partly supported by NIH grant number P30-ES020957. Authors would also like to thank an anonymous referee and the editorial team, whose comments provided additional insights and have greatly improved the scope and presentation of the paper.

References

- [1]. D'Agostino RB, Massaro JM, and Sullivan LM. Noninferiority trials: Design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine*, 22(2):169–186, 2003. [PubMed: 12520555]
- [2]. ICH E9. ICH Harmonised Tripartite Guideline. Statistical Principles for Clinical Trials. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2009.
- [3]. ICH E10. ICH Harmonised Tripartite Guideline. Choice of Control Group and Related Issues in Clinical Trials. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, 2009.
- [4]. FDA. Food and Drug Administration Guidance for Industry Noninferiority Clinical Trials., 2010.
- [5]. EMA. Guideline on the choice of the noninferiority margin (Doc. Ref. EMEA/CPMP/EWP/215)., 2005.
- [6]. Hung HMJ and Wang SJ. Multiple testing of noninferiority hypotheses in active controlled trials. *Journal of Biopharmaceutical Statistics*, 14(2):327–335, 2004. [PubMed: 15206530]
- [7]. Schumi J and Wittes JT. Through the looking glass: understanding non-inferiority. *Trials*, 12(2): 106–118, 2011. [PubMed: 21539749]
- [8]. Pigeot I, Schfer J, Rhmel J, and Hauschke D. Assessing noninferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, 22(6):883–899, 2003. [PubMed: 12627407]
- [9]. Hida E and Tango T. On the three-arm noninferiority trial including a placebo with a prespecified margin. *Statistics in Medicine*, 30(3):224–231, 2011. [PubMed: 21213340]
- [10]. Kwong KS, Cheung SH, Hayter AJ, and Wen MJ. Extension of three-arm non-inferiority studies to trials with multiple new treatments. *Statistics in Medicine*, 31(24):2833–2843, 2012. [PubMed: 22744965]
- [11]. Simon R. Bayesian design and analysis of active control clinical trials. *Biometrics*, 55(2):484–487, 1999. [PubMed: 11318204]
- [12]. Spiegelhalter DJ, Abrams KR, and Myles JP. Bayesian approaches to clinical trials and health-care evaluation. New York: Wiley, 2004.
- [13]. Liu J, Hsueh H, and Hsiao C. A bayesian non-inferiority approach to evaluation of bridging studies. *Journal of Biopharmaceutical Statistics*, 14(2):291–300, 2004. [PubMed: 15206527]
- [14]. Ghosh P, Nathoo F, Gnen M, and Tiwari RC. Assessing noninferiority in a three-arm trial using the bayesian approach. *Statistics in Medicine*, 30(15):1795–1808, 2011. [PubMed: 21520456]
- [15]. Gamalo MA, Wu R, and Tiwari RC. Bayesian approach to non-inferiority trials for normal means. *Statistical Methods in Medical Research*, Published online 21 5 2012.
- [16]. Gamalo MA, Tiwari RC, and LaVange LM. Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics*, 13(1):25–40, 2014. [PubMed: 23913880]
- [17]. Lewis RJ and Berry DA. Group sequential clinical trials: a classical evaluation of bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89(428):1528–1534, 1994.
- [18]. Wathen JK and Thall PF. Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine*, 27(27):5586–5604, 2008. [PubMed: 18752257]
- [19]. Koch A and Rohmel J. Hypothesis testing in the gold standard design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of biopharmaceutical statistics*, 14(2):315–325, 2004. [PubMed: 15206529]

- [20]. Welch BL. The generalization of “student’s” problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947. [PubMed: 20287819]
- [21]. Rothstein HR, Sutton AJ, and Borenstein M. *Publication bias in meta-analysis: Prevention, assessment and adjustments*. John Wiley & Sons, 2006.
- [22]. Sterne JA, Gavaghan D, and Egger M. Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*, 53(11):1119–1129, 2000. [PubMed: 11106885]
- [23]. Berger RL. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4): 295–300, 1982.
- [24]. Dmitrienko A, Tamhane AC, and Bretz F. *Multiple testing problems in pharmaceutical statistics*. CRC Press, 2009.
- [25]. Stucke K and Kieser M. A general approach for sample size calculation for the three-arm “gold standard” noninferiority design. *Statistics in Medicine*, 31(28):3579–3596, 2012. [PubMed: 22764073]
- [26]. Laska EM and Meisner MJ. Testing whether an identified treatment is best. *Biometrics*, pages 1139–1151, 1989. [PubMed: 2611321]
- [27]. Deng X, Xu J, and Wang C. Improving the power for detecting overlapping genes from multiple dna microarray-derived gene lists. *Statistics in Medicine*, 9(Suppl 6):S14, 2008.
- [28]. Gamalo MA, Wu R, and Tiwari RC. Bayesian approach to noninferiority trials for proportions. *Journal of biopharmaceutical statistics*, 21(5):902–919, 2011. [PubMed: 21830922]
- [29]. Higuchi T, Murasaki M, and Kamijima K. Clinical evaluation of duloxetine in the treatment of major depressive disorder-placebo and paroxetine-controlled double-blinded comparative study. *Japanese Journ of Clinical Psychopharmacology*, 12:1613–1634, 2009.
- [30]. Pezzin LE, Feldman PH, Mongoven JM, McDonald MV, Gerber LM, and Peng TR. Improving blood pressure control: Results of home-based post-acute care interventions. *Journ. of Gen Intern Med*, 26(3):280–286, 2011.
- [31]. Lange S, Freitag G, and Trampisch HJ. Practical experience with the design and analysis of a three-armed equivalence study. *Eur. Journ. of Clin. Pharmacol*, 54(7):535–540, 1998.

Figure 1. Density plot of NI margin over 10000 simulations for two different situations, evaluating the effect of historical trial over the width of NI margin.

Table 1.

Historical trials strongly establish Active Control (against placebo): For the power comparison, we set $\mu_E = \mu_R - \delta_{NI}^B + \xi_1$ and $\mu_P = \mu_R - \delta_{AS}^B - \xi_2$, where $\mu_P^H = 0$, $\mu_R^H = \mu_R = 1$, $\sigma_{P,H}^2 = 1$, $\sigma_{R,H}^2 = 1$, $\sigma_P^2 = 2$, $\sigma_R^2 = 1$ and $\sigma_E^2 = 1.5$, where $r = 1 - \lambda$, $p^* = 0.975$ in 50,000 replications and 5,000 samplers. To reduce the computational cost $\xi_1 = \xi_2/2$ with $n_{P,H} = n_{R,H} = 100$; and $n_R = n_P = n_E = 150,300$.

$n_R = n_p = n_E = 150$			$n_R = n_p = n_E = 300$		
$\lambda = 0$					
ξ_2	Bayesian Method	Frequentist Method	ξ_2	Bayesian Method	Frequentist Method
0.2	0.08	0.01	0.2	0.11	0.06
0.4	0.35	0.23	0.4	0.64	0.57
0.6	0.72	0.63	0.6	0.93	0.91
0.8	0.92	0.87	0.8	1.00	0.99
$\lambda = 0.25$					
0.2	0.08	0.01	0.2	0.11	0.05
0.4	0.34	0.23	0.4	0.64	0.57
0.6	0.71	0.62	0.6	0.94	0.90
0.8	0.92	0.87	0.8	1.00	0.99
$\lambda = 0.50$					
0.2	0.07	0.01	0.2	0.11	0.06
0.4	0.35	0.23	0.4	0.63	0.57
0.6	0.71	0.62	0.6	0.93	0.94
0.8	0.92	0.87	0.8	1.00	0.99
$\lambda = 0.75$					
0.2	0.08	0.01	0.2	0.11	0.05
0.4	0.35	0.24	0.4	0.62	0.57
0.6	0.71	0.62	0.6	0.93	0.91
0.8	0.92	0.87	0.8	1.00	0.99

Table 2.

Historical trials when Active Control is not strongly established (against placebo): For the power comparison, we set $\mu_E = \mu_R + \xi_1$ and $\mu_P = \mu_R - \xi_2$, where $\mu_P^H = 0, \mu_R^H = \mu_R = 0.3, \sigma_{P,H}^2 = 1, \sigma_{R,H}^2 = 1, \sigma_P^2 = 2, \sigma_R^2 = 1$ and $\sigma_E^2 = 1.5$, where $r = 1 - \lambda, p^* = 0.975$ in 50,000 replications and 5,000 samplers. To reduce the computational cost $\xi_1 = \xi_2/2$ with $n_{P,H} = n_{R,H} = 100$; and $n_R = n_P = n_E = 150,300$.

$n_R = n_p = n_E = 150$			$n_R = n_p = n_E = 300$		
$\lambda = 0$					
ξ_2	Bayesian Method	Frequentist Method	ξ_2	Bayesian Method	Frequentist Method
0.2	0.04	0.00	0.2	0.08	0.03
0.4	0.24	0.09	0.4	0.54	0.23
0.6	0.67	0.51	0.6	0.90	0.79
0.8	0.87	0.83	0.8	1.00	0.98
$\lambda = 0.25$					
0.2	0.05	0.00	0.2	0.07	0.01
0.4	0.24	0.08	0.4	0.55	0.22
0.6	0.66	0.50	0.6	0.91	0.78
0.8	0.88	0.83	0.8	1.00	0.98
$\lambda = 0.50$					
0.2	0.04	0.00	0.2	0.07	0.01
0.4	0.25	0.07	0.4	0.57	0.23
0.6	0.66	0.50	0.6	0.91	0.78
0.8	0.88	0.83	0.8	1.00	0.97
$\lambda = 0.75$					
0.2	0.04	0.00	0.2	0.09	0.02
0.4	0.25	0.07	0.4	0.58	0.22
0.6	0.67	0.51	0.6	0.92	0.77
0.8	0.86	0.83	0.8	1.00	0.97

Table 3.

Bayesian and frequentist decision in the Depression trial where “1” stands for rejection and “0” stands for acceptance. We reported individual testing of NI and AS hypothesis and also the joint decision. NI with AS is achieved when both H_0 and K_0 are rejected. The values are obtained using $p^* = 0.975$,

$n_P^H = 130, n_R^H = 140, \mu_P^H = 8.5, \sigma_{PH} = 6.0, \mu_R^H = 10, \sigma_{RH} = 6.0$ and $N = 50,000$.

λ	Posterior Probab.	$\mathcal{E}^B$ Margin	$\mathcal{E}^F$ Margin	Bayesian Decision	NI^B Decision	AS^B Decision	Frequentist Decision	NI^F Decision	AS^F Decision
0.00	0.95	0.716	0.732	0	1	0	0	1	0
0.25	0.93	0.537	0.549	0	1	0	0	1	0
0.50	0.90	0.358	0.366	0	1	0	0	1	0
0.75	0.86	0.179	0.183	0	1	0	0	1	0
0.90	0.83	0.071	0.073	0	1	0	0	1	0
1.00	0.80	0.000	0.000	0	1	0	0	1	0

Table 4.

Posterior Estimates and 95% credible interval.

Parameter	Mean	Interval
μ_P	8.12	(7.39, 8.85)
μ_R	9.87	(9.13, 10.61)
μ_E	10.48	(9.45, 11.51)
$\mu_E - \mu_R$	0.61	(-0.65, 1.88)
$\mu_R - \mu_P$	1.76	(0.72, 2.79)

Table 5.

Bayesian and frequentist decision in the Home-Based Blood Pressure Intervention Study where “1” stands for rejection and “0” stands for acceptance. We reported individual testing of NI and AS hypothesis and also the joint decision. Both the Bayesian and frequentist approaches fail to establish the non-inferiority though both the methods establish the assay sensitivity. The values are obtained using $p^* = 0.975$,

$n_P^H = 210, n_R^H = 205, \mu_P^H = 159, \sigma_{PH} = 26, \mu_R^H = 154, \sigma_{RH} = 26$ and $N = 50,000$.

λ	Posterior Probab.	δ^B Margin	δ^F Margin	Bayesian Decision	NI^B Decision	AS^B Decision	Frequentist Decision	NI^F Decision	AS^F Decision
0	0.11	1.792	1.591	0	0	1	0	0	1
0.25	0.08	1.344	1.193	0	0	1	0	0	1
0.5	0.05	0.896	0.795	0	0	1	0	0	1
0.75	0.04	0.448	0.397	0	0	1	0	0	1
0.9	0.03	0.179	0.159	0	0	1	0	0	1
1	0.02	0	0	0	0	1	0	0	1

Table 6.

Posterior Estimates and 95% credible interval.

Parameter	Mean	Interval
μ_P	160.98	(158.59, 163.39)
μ_R	154.21	(151.91, 156.61)
μ_E	158.77	(155.10, 162.48)
$\mu_E - \mu_R$	4.55	(0.15, 9.00)
$\mu_P - \mu_R$	6.77	(3.40, 10.12)

Table 7.

Bayesian and frequentist rejection decisions in the Mild-Asthma Study where “1” stands for rejection and “0” stands for acceptance. Along with the joint decision, individual testing of NI and AS hypothesis are reported. Both the Bayesian and frequentist approaches fail to establish non-inferiority though assay sensitivity is established. The values are obtained using $p^* = 0.975$, $n_P^H = 30$, $n_R^H = 30$, $\mu_P^H = 3.1$, $\sigma_{PH} = 1$, $\mu_R^H = 4.35$, $\sigma_{RH} = 1$ and $N = 5,000$ for each dataset.

λ	Posterior Probab.	δ^B Margin	δ^F Margin	Bayesian Decision	NI^B Decision	AS^B Decision	Frequentist Decision	NI^F Decision	AS^F Decision
0	0.98	0.836	0.835	1	1	1	0	0	1
0.25	0.89	0.627	0.626	0	0	1	0	0	1
0.5	0.67	0.418	0.417	0	0	1	0	0	1
0.75	0.36	0.209	0.208	0	0	1	0	0	1
0.9	0.2	0.083	0.083	0	0	1	0	0	1
1	0.13	0	0	0	0	1	0	0	1

Table 8.

Posterior Estimates and 95% credible interval.

Parameter	Mean	Interval
μ_P	2.91	(2.62,3.21)
μ_R	4.61	(4.33,4.91)
μ_E	4.31	(3.87, 4.75)
$\mu_E - \mu_R$	-0.31	(-0.83,0.22)
$\mu_R - \mu_P$	1.70	(1.29,2.11)