



A more powerful test for three-arm non-inferiority via risk difference: Frequentist and Bayesian approaches

Erina Paul ^a, Ram C. Tiwari^{†b}, Shrabanti Chowdhury^{a,c} and Samiran Ghosh^{a,c}

^aCenter of Molecular Medicine and Genetics, Wayne State University, Detroit, MI, USA; ^bDivision of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, FDA, Silver Spring, MD, USA; ^cFamily Medicine & Public Health Sciences, Wayne State University, Detroit, MI, USA

ABSTRACT

Necessity for finding improved intervention in many legacy therapeutic areas are of high priority. This has the potential to decrease the expense of medical care and poor outcomes for many patients. Typically, clinical efficacy is the primary evaluating criteria to measure any beneficial effect of a treatment. Albeit, there could be situations when several other factors (e.g. side-effects, cost-burden, less debilitating, less intensive, etc.) which can permit some slightly less efficacious treatment options favorable to a subgroup of patients. This often leads to non-inferiority (NI) testing. NI trials may or may not include a placebo arm due to ethical reasons. However, when included, the resulting three-arm trial is more prudent since it requires less stringent assumptions compared to a two-arm placebo-free trial. In this article, we consider both Frequentist and Bayesian procedures for testing NI in the three-arm trial with binary outcomes when the functional of interest is risk difference. An improved Frequentist approach is proposed first, which is then followed by a Bayesian counterpart. Bayesian methods have a natural advantage in many active-control trials, including NI trial, as it can seamlessly integrate substantial prior information. In addition, we discuss sample size calculation and draw an interesting connection between the two paradigms.

ARTICLE HISTORY

Received 1 September 2020
Accepted 16 October 2021

KEYWORDS


Assay sensitivity; conditional approach; Dirichlet prior; risk difference; non-inferiority margin

1. Introduction

Well-designed Randomized Control Trials (RCTs) are accepted standards for measuring an intervention's impact across many diverse disease areas and thus are considered gold-standard for establishing a new treatment regime. In the presence of clinically proven established treatments/therapies, it is not ethical justified to allocate patients in the placebo arm. Thus, this gives rise to trials that compare the experimental drug with one or more active comparators. These trials are considered as standard in Comparative Effectiveness Research (CER), which is designed for providing evidence on the effectiveness, benefits,

CONTACT Samiran Ghosh  sghos@med.wayne.edu

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

 Supplemental data for this article can be accessed here. <https://doi.org/10.1080/02664763.2021.1998391>

and risks of a broad range of interventions to the patients and clinicians. Although multiple treatment comparisons are possible [27], for the sake of simplicity, here we consider one experimental arm compared with one active reference. However, in certain situations superiority of the new drug might be in question. It may be reasonable to test if the experimental treatment is not worse than the reference by more than a pre-specified margin. These type of active-controlled trials known as the non-inferiority (NI) trials are intended to show if the new drug retains a substantial portion of the active control effect, thus making it more preferable to some patients due to its other desirable properties [8]. The choice of a pre-specified margin, also termed as NI margin (δ), is a critical issue in these trials. Although regulatory agencies provided broad guidelines on the choice of δ [9,10,22,23], it must be examined based on the past performance of the active control and is usually desirable to choose a margin that reflects the clinically acceptable largest loss of effect. Hence, the NI trials need to be administered with extreme caution [10,21,32].

In the last few decades, two-arm NI trials (experimental vs. reference) have been developed predominantly under the Frequentist paradigm. However, two-arm trials suffer some major challenges from the design, analysis, and possible interpretations point of view. One of the major concerns is that the two-arm NI trial may not support assay sensitivity (AS) directly and requires external validation [8]. This is because without a placebo arm no direct proof can be established about the efficacy of the reference drug over placebo. To compensate this, [9] recommends the inclusion of a placebo arm when ethically possible, resulting in three-arm 'gold-standard' design, that has greater confidence concerning AS and lesser concern related to external validity. For three-arm trials, [25,31] proposed the choice of NI margin as the pre-specified fraction of the unknown effect size of reference drug, instead of directly specifying a fixed margin. Later, this approach was extended first by [34] and then by [24] for binary end-point by considering difference of proportions (or risk difference). Pigeot *et al.* [31] suggested that the superiority of the reference drug over placebo should be established first to satisfy the AS assumption before carrying out such NI testing. An alternative to this is the fixed margin approach of [18] which requires joint testing of NI and AS albeit resulting in a rather conservative intersection-union type test [17]. In this article, we proposed an improved Frequentist test based on conditional principle following Pigeot's fraction margin approach for a binary outcome. Note, for binary end-points, risk difference is not the only function of interest. However, the nature of NI hypothesis, margin construction and the resulting methodological formulation for other types of functional will diverge significantly as shown in our recently published paper (see [4,5]). For related developments on those directions, please see the discussion section.

Clinical trials, particularly NI trials, have used Bayesian approaches since long past which can be found in the references, for example, see [11,13,16,33] among others. Gamalo *et al.* [12] considered Bayesian approach for the analysis of two-arm NI trials for binary outcomes. Ghosh *et al.* [16] also put forward a novel Bayesian analysis for a three-arm NI trial following Pigeot's fraction margin approach. The existence of prior information in the current NI trial is advantageous. The Bayesian paradigm delivers a natural route to obtain the prior information and help to reduce the sample size as well as cost by combining that information with the current trial. In this paper, we also propose an exact Bayesian procedure that is based on the conditional Frequentist principle to test NI. We also propose an approximation-based Bayesian approach that gives a closed-form solution of the Bayesian posterior probability, thus avoiding computational complexity of an exact

Bayesian approach with a slight loss of accuracy. All approaches are evaluated on simulated and one published dataset from mental health trial.

The rest of the article is organized as follows. In Section 2, we give the NI hypothesis and the existing and our proposed Frequentist method for testing it. In Section 3, we propose a novel Bayesian methodology to design and perform the analysis of a three-arm NI trial for binary outcomes. Along with conjugate priors, we consider two other prior scenarios incorporating the condition of AS. Section 4 presents an interesting connection between Frequentist and Bayesian posterior probabilities in the three-arm trial. In Section 5, we present the algorithm and results for simulation studies as well as sample size table. Finally, in Section 6, we apply our proposed methodology to a published clinical trial dataset. We conclude the article with discussions in Section 7.

2. Three-arm Frequentist NI testing

Following [4,5,24], we construct the three arm non-inferiority trial for the primary binary endpoints under the experimental (E), reference (R), and placebo (P) arms. Let X_l , $l \in \{E, R, P\}$ denote the number of successes with n_l number of subjects in l th arm. The random variable $X_l \sim \text{Bin}(n_l, \pi_l)$ with the corresponding probability $\pi_l (\in [0, 1])$. Without loss of generality, we assume the greater treatment benefits for the higher response probabilities in this scenario. In general, for two arm NI trial, the risk difference problem with pre-specified NI margin, $\delta < 0$, is stated as $H_0 : \pi_E - \pi_R \leq \delta$ vs. $H_1 : \pi_E - \pi_R > \delta$. Pigeot *et al.* [31] and Kieser and Friede [24] proposed the mathematical expression for δ as $f(\pi_R - \pi_P)$, where f is a negative fraction assuming the AS condition, that is, $\pi_R > \pi_P$. As discussed in [4,5], we can build the three arm NI hypothesis using δ and hence f as $H_0 : (\pi_E - \pi_P)/(\pi_R - \pi_P) \leq \theta$ vs. $H_1 : (\pi_E - \pi_P)/(\pi_R - \pi_P) > \theta$, where $\theta = 1 + f$ is a pre-specified fraction of the effect of the reference drug relative to the placebo. In this three-arm NI trial, the efficacy of the test drug when compared to placebo attains more than $\theta \times 100\%$ of the efficacy of the reference drug as compared to placebo. Although the choice of $\theta (\in [0, 1])$ as shown in [31] depends on the clinical approval, in this case θ is limited in $[0.5, 1)$ for the NI testing of the new drug to retain at least 50% effect of the active control. Hence, the NI hypothesis of the risk difference can be written as

$$H_0 : \pi_E - \theta\pi_R - (1 - \theta)\pi_P \leq 0 \quad \text{vs.} \quad H_1 : \pi_E - \theta\pi_R - (1 - \theta)\pi_P > 0. \quad (1)$$

For this NI test, the rejection of null hypothesis satisfies that a pre-defined proportion of the unknown effect of the reference over placebo is maintained by the experimental treatment.

2.1. Existing marginal approach

Kieser and Friede [24] developed statistical test procedures under Frequentist paradigm for the NI testing under three-arm trial for binary outcomes. They constructed the test statistic for testing the NI hypothesis in (1) by considering the maximum likelihood estimate (MLE) of the linear contrast $\pi_E - \theta\pi_R - (1 - \theta)\pi_P$, given by $T = \hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P$, where $\hat{\pi}_l = X_l/n_l$ is the MLE of π_l , $l \in \{E, R, P\}$. Different tests can be obtained by considering the maximum likelihood (ML) or restricted ML (RML) estimate of the variance of T , given by $\text{Var}(T) = \pi_E(1 - \pi_E)/n_E + \theta^2\pi_R(1 - \pi_R)/n_R + (1 - \theta)^2\pi_P(1 - \pi_P)/n_P$. RML estimates can be obtained subject to the constraint $\pi_E - \theta\pi_R - (1 - \theta)\pi_P = 0$. Under

asymptotic normality, the standardized statistic $T/\sqrt{\text{Var}(T)} \stackrel{H_0}{\sim} N(0, 1)$, since $\pi_E - \theta\pi_R - (1 - \theta)\pi_P = 0$. An asymptotic level α Wald-type test is obtained by rejecting the null hypothesis if T exceeds $100(1 - \alpha)\%$. In this case, the power of T can be written as $1 - \Phi(z_{1-\alpha}\sigma_T^{\text{null}}/\sigma_T^{\text{alt}} - \mu_T^{\text{alt}}/\sigma_T^{\text{alt}})$, where σ_T^{null} denotes the standard deviation under H_0 and $\mu_T^{\text{alt}}, \sigma_T^{\text{alt}}$ represent the mean and standard deviation, respectively, under H_1 .

2.2. A novel Frequentist proposal

It is important to note that a pretest for the superiority of the active control over the placebo should be performed before the NI is investigated (see [31]). NI testing thus then only carried out as a second step provided the AS condition ($\pi_R > \pi_P$) holds. However, it is often agreed [16,24,25,31] that if active control retains majority of the effect over placebo then in practice the statistical power to perform joint testing (NI and AS) will be very similar to that of testing NI only [35]. However, this may not always be true and traditionally the pre-tested AS condition has not been used further in NI testing in the marginal Frequentist effect-retention approach, except for margin construction. We introduce here a more powerful conditional approach for risk difference. Since NI and AS hypothesis are related, this leads to significant power gain in certain situations. Notably, [4,5] proposed a similar approach for risk ratio and odds ration albeit without any theoretical guarantee for power gain. A major point of this paper is to show that both theoretically as well as via simulation. For finding the MLE, we truncate the parameter space of (π_E, π_R, π_P) such that it belongs to $\{\pi_E, \pi_R, \pi_P : \pi_E \in [0, 1], \pi_R \in [0, 1], \pi_P \in [0, 1], \pi_R > \pi_P\}$. One may develop a likelihood ratio test based on the statistic

$$T = \hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P \tag{2}$$

the AS condition $\hat{\pi}_R > \hat{\pi}_P$ under null hypothesis via Wald-type test. Following [29], one can improve the convergence via the RML which requires solving under H_0

$$(\hat{\pi}_{E,\text{RML}}, \hat{\pi}_{R,\text{RML}}, \hat{\pi}_{P,\text{RML}}) = \underset{\pi_E - \theta\pi_R - (1-\theta)\pi_P \leq 0, \pi_R > \pi_P}{\text{arg max}} \log l(\pi_E, \pi_R, \pi_P), \tag{3}$$

where $\log l(\pi_E, \pi_R, \pi_P)$ is the log-likelihood of (π_E, π_R, π_P) . For the odds and risk ratios, [4,5] discussed a strategy using unrestricted MLE to reduce the computational difficulty. This strategy is well established in many practical applications as mentioned in [20,26]. Using similar concept, that is, using $T_{\text{ML}} = \hat{\pi}_{E,\text{ML}} - \theta\hat{\pi}_{R,\text{ML}} - (1 - \theta)\hat{\pi}_{P,\text{ML}}$, we can solve the optimization problem numerically for the risk difference. However, for our case, we consider the part restricted by the AS condition, $\hat{\pi}_{R,\text{ML}} > \hat{\pi}_{P,\text{ML}}$ and hence

$$T_{\text{RML}} \simeq T_{\text{ML}} * I[\hat{\pi}_{R,\text{ML}} > \hat{\pi}_{P,\text{ML}}], \tag{4}$$

where ‘ \simeq ’ represents the approximation. The distribution of the product of random variables can be formulated as $f(T_{\text{RML}}) \simeq f(T_{\text{ML}} | \hat{\pi}_{R,\text{ML}} > \hat{\pi}_{P,\text{ML}}) \times Pr[\hat{\pi}_{R,\text{ML}} > \hat{\pi}_{P,\text{ML}}]$. Hence, the test statistic can be written as $(T_{\text{ML}} | \hat{\pi}_R > \hat{\pi}_P) \propto (\hat{\pi}_{E,\text{ML}} - \theta\hat{\pi}_{R,\text{ML}} - (1 - \theta)\hat{\pi}_{P,\text{ML}} | \hat{\pi}_{R,\text{ML}} > \hat{\pi}_{P,\text{ML}})$. From now onwards, we denote the ML estimate $\hat{\pi}_{l,\text{ML}}$ by $\hat{\pi}_l$, $l \in \{E, R, P\}$. Now one can reformulate the test statistic for three-arm NI testing as $W = (\hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P) | \hat{\pi}_R > \hat{\pi}_P = ((\hat{\pi}_E - \hat{\pi}_P) - \theta(\hat{\pi}_R - \hat{\pi}_P)) | \hat{\pi}_R > \hat{\pi}_P \equiv (U - \theta V) | V > 0$. The exact small sample distribution of W is non-normal under

the current-setup, however, [1] proved that W has Normal distribution under continuous setting. Hence, for the binary case under asymptotic normality of W , we can similarly prove that $(W - \mu_w)/\sigma_w \sim AN(0, 1)$, where μ_w and σ_w^2 are the mean and variance of W , respectively. Chowdhury *et al.* [4,5] proved a similar lemma for calculating mean and variance under the conditional approach for risk and odds ratios. In this paper, we use the same approach to prove the lemma for risk difference.

Lemma 2.1: *Under conditional normal approach, the mean μ_w and variance σ_w^2 of $W = (\hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P | \hat{\pi}_R > \hat{\pi}_P)$ are given by*

$$\begin{aligned} \mu_w &= \mu_U + \sigma_U \frac{\rho}{c} \phi(d) - \theta \left(\mu_V + \sigma_V \frac{1}{c} \phi(d) \right), \\ \sigma_w^2 &= \sigma_U^2 \left[1 + \frac{\rho^2}{c} d\phi(d) - \left(\frac{\rho}{c} \phi(d) \right)^2 \right] + \theta^2 \sigma_V^2 \left[1 - \frac{\phi(d)}{c} \left(\frac{\phi(d)}{c} - d \right) \right] \\ &\quad - 2\theta \left[\sigma_U \sigma_V \frac{\rho}{c} (c + d\phi(d)) + \sigma_U \mu_V \frac{\rho}{c} \phi(d) + \sigma_V \mu_U \frac{1}{c} \phi(d) + \mu_U \mu_V \right. \\ &\quad \left. - \left(\mu_U + \sigma_U \frac{\rho}{c} \phi(d) \right) \left(\mu_V + \sigma_V \frac{1}{c} \phi(d) \right) \right], \end{aligned}$$

where $\mu_U = \pi_E - \pi_P$, $\mu_V = \pi_R - \pi_P$, $\sigma_l^2 = \pi_l(1 - \pi_l)/n_l$, $l \in \{E, R, P\}$, $\sigma_U^2 = \sigma_E^2 + \sigma_P^2$, $\sigma_V^2 = \sigma_R^2 + \sigma_P^2$, $\rho = \text{Var}(\hat{\pi}_P)/\sqrt{\text{Var}(U)\text{Var}(V)} = \sigma_P^2/\sqrt{\sigma_U^2\sigma_V^2}$, $d = -\mu_V/\sigma_V$, and $c = 1 - \Phi(d)$.

Proof: See Supplementary Material 1. ■

Now, $(W - \mu_w^{\text{null}})/\sigma_w^{\text{null}} \sim AN(0, 1)$ under H_0 and $(W - \mu_w^{\text{alt}})/\sigma_w^{\text{alt}} \sim AN(0, 1)$ under H_1 , where by μ_w^{null} and μ_w^{alt} are the means under null and alternative, respectively, and $\sigma_w^{2\text{null}}$ and $\sigma_w^{2\text{alt}}$ are the variances under null and alternative, respectively. Hence, the critical region of the test under the Frequentist approach is given by $W > k^*$, where k^* is obtained by assuming a test of size α : $P_{H_0}(W > k^*) = \alpha \Rightarrow k^* = \mu_w^{\text{null}} + z_{1-\alpha}\sigma_w^{\text{null}}$, where $z_{1-\alpha}$ is the $100(1 - \alpha)\%$ percentile point of the $N(0, 1)$ distribution. In general, the value of α is set to be 0.025. Based on the Lemma 2.1, it can be noted that μ_w^{null} , μ_w^{alt} , σ_w^{null} , and σ_w^{alt} depends on π_E , π_R , and π_P with π_E^{null} satisfies $\pi_E^{\text{null}} - \theta\pi_R - (1 - \theta)\pi_P = 0$ and π_E^{alt} satisfies $\pi_E^{\text{alt}} - \theta\pi_R - (1 - \theta)\pi_P > 0 \Rightarrow (\pi_E^{\text{alt}} - \pi_P) > \theta(\pi_R - \pi_P)$, where *null* and *alt* in the exponent represent the proportions under null and alternative, respectively. In simulation study, we followed the approach of [4,5,17] to generate π_R for a pre-defined θ , π_E , and π_P such that it satisfies null hypothesis of equality as mentioned in Equation (1).

Lemma 2.2: *At fixed α and sample size, our proposed conditional test statistic ($W = (\hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P | \hat{\pi}_R > \hat{\pi}_P)$) has equal or more power than the existing marginal test statistic ($T = \hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P$) for testing NI hypothesis in (1).*

Proof: See Supplementary Material 2. ■

This lemma shows that there is effective power gain in the conditional test or conversely speaking, to attain a fixed power, the conditional test requires smaller sample size. Though

for simplicity, the proof is given for equal allocation case, it can be easily extended for more general unequal allocation case. As observed in the simulation study (Section 5), this power gain is substantial when the gap between π_R and π_P is small and negligible when $\pi_R \gg \pi_P$. This is parallel to what was noted at the beginning of Section 2.2. Note that the above theoretical claim for power gain via conditional approach for a binary outcome is restricted to risk difference case only.

2.3. Sample size

Using our proposed approach, we can calculate sample size for the assessment of NI to attain a desired power for a point alternative $\pi_E = \pi_E^{\text{alt}}$: $P_{H_1}(W > k^*) = 1 - \Phi(k^* - \mu_w^{\text{alt}}/\sigma_w^{\text{alt}})$. The power function of the test is derived by specifying the fixed values of π_R , π_P , and θ and consider different values of π_E such that the ratio $(\pi_E - \pi_P)/(\pi_R - \pi_P) \in [0.5, 1.4]$. As described in [4,5], let r_1 and r_2 be the allocation ratio of the sample sizes corresponding to the reference and placebo arms, respectively, relative to the experimental arm with sample of size $n_E = n$. Hence, the total sample size can be expressed as $N = n(1 + r_1 + r_2)$ for the allocation ratio $n_E : n_R : n_P = 1 : r_1 : r_2$. To attain at least $100(1 - \beta)\%$ power, the sample size ‘ n ’ (of the arm E) is computed via the equation

$$P_{H_1}(W > k^*) \geq 1 - \beta \Rightarrow \Phi\left(\frac{k^* - \mu_w^{\text{alt}}}{\sigma_w^{\text{alt}}}\right) \leq \beta. \tag{5}$$

In this paper, we set β as 20% and vary π_E^{alt} to get the minimum sample size needed for 80% power.

3. Three-arm Bayesian NI testing

Any NI trial by design is an active control trial, where the availability of historical data on one or more arm/s is more or less guaranteed. Bayesian design [30] offers an interesting pathway to bring this additional information into play which can lead to substantial savings. Gamalo *et al.* [12] developed Bayesian procedures for NI testing in two-arm trial with binary end-point that allows the incorporation of the historical data on the active control via the use of informative priors. In this section, we propose an exact Bayesian and an approximate Bayesian test procedure under fraction margin approach for three-arm NI trial via risk difference.

3.1. Exact Bayesian approach

We consider three different prior choices, such as Conjugate Beta Prior (CBP), Proper Uniform Prior (PUP), and Dirichlet Prior (DP), where the AS condition is incorporated explicitly parallel to the proposed Frequentist approach described earlier. Among these three priors, the sampling procedure is easy to implement for CBP, whereas for DP it is computationally more intensive than other two procedures. For the PUP, the sampling has to be done from the restricted domain. Also, the posterior is not in the closed-form for DP. So in this section, for the illustration purpose, we provide the formal test procedure for NI testing and address the sample size calculation based on these three different prior settings.

3.1.1. Conjugate Beta prior (CBP)

Under the Binomial setting, the usual conjugate prior is the Beta distribution. In this three-arm NI trial, we assume the Beta prior with hyper-parameters $\alpha_l \in \mathbb{R}^+$ and $\beta_l \in \mathbb{R}^+$ with $l \in \{E, R, P\}$, for the proportion of successes (π_l) as proposed in [4]. For the three-arm NI trial with AS condition, the joint prior distribution of the proportions of successes can be defined as $f(\pi_E, \pi_R, \pi_P) = I(\pi_R > \pi_P) \prod_{l \in \{E, R, P\}} f(\pi_l | \alpha_l, \beta_l)$, where $f(\pi_l | \alpha_l, \beta_l)$ is the density of the standard Beta distribution. The joint posterior distribution of proportions given the number of successes can be written as $f(\pi_E, \pi_R, \pi_P | X_E, X_R, X_P, \alpha_l, \beta_l) \propto I(\pi_R > \pi_P) \prod_{l \in \{E, R, P\}} \pi_l^{\alpha_l + x_l - 1} (1 - \pi_l)^{\beta_l + n_l - x_l - 1}$, $0 < \pi_l < 1$. Under the conditional approach, the posterior samples can be generated from the joint posterior distribution satisfying the AS condition, that is, $\pi_R > \pi_P$. Now, based on the prior information of the placebo-controlled trial, we can choose the value of the hyper-parameters. For the informative prior, the hyper-parameters can be computed by equating the mean or mode (with smaller variance) with the success probabilities. If we do not have substantial ideas about the parameters, non-informative prior is a common choice in this situation and in case of Beta non-informative prior, the choice of hyper-parameters is $\alpha_l = \beta_l = 1$ with $l \in \{E, R, P\}$.

3.1.2. Proper uniform prior (PUP)

In this case, the prior distributions are assigned to the parameters π_E, π_R , and π_P so that the restriction $0 < \pi_P < \pi_R < 1$ is automatically satisfied. We give joint prior on (π_R, π_P) by putting Beta distribution on π_P and conditional on π_P , a truncated Uniform distribution on π_R , with the support on $(\pi_P, 1)$, so that $\pi_R > \pi_P$. We also put unrestricted prior $\text{Beta}(\alpha_E, \beta_E)$ on π_E . Thus, the joint distribution of (π_R, π_P) is $f(\pi_R, \pi_P) \propto \pi_P^{\alpha_P - 1} (1 - \pi_P)^{\beta_P - 2}$, $0 < \pi_P < \pi_R < 1$ and the joint distribution of (π_E, π_R, π_P) is given by $f(\pi_E, \pi_R, \pi_P) \propto \pi_E^{\alpha_E - 1} (1 - \pi_E)^{\beta_E - 1} \pi_P^{\alpha_P - 1} (1 - \pi_P)^{\beta_P - 2}$, $0 < \pi_E < 1$, $0 < \pi_P < \pi_R < 1$. The joint posterior distribution, obtained by multiplying the joint likelihood with the joint prior, is proportional to the product of two full Beta and one truncated Beta distribution: $f(\pi_E, \pi_R, \pi_P | \mathbf{X}) \propto \text{Beta}(\pi_E | \alpha_E + X_E, \beta_E + n_E - X_E) \times \text{Trunc Beta}(\pi_R | X_R + 1, n_R - X_R + 1) \times \text{Beta}(\pi_P | \alpha_P + X_P, n_P + \beta_P - 1 - X_P)$, where $\text{Trunc Beta}(\pi_R | X_R + 1, n_R - X_R + 1)$ is a truncated Beta distribution with support on $0 < \pi_P < \pi_R < 1$ and \mathbf{X} denotes the relevant data. The MCMC samples from the posterior for π_E and π_P can be generated from the updated Beta distributions. Given a draw for π_P , the MCMC samples for π_R can be generated from the truncated Beta distribution with the support $(\pi_P, 1)$.

3.1.3. Dirichlet prior (DP)

In this setup, we put a Dirichlet prior on (π_R, π_P) with support on $0 < \pi_P < \pi_R < 1$. We make the following transformation $(\pi_R, \pi_P) \Rightarrow (u_1, u_2, u_3)$ such that $u_1 = \pi_P$, $u_2 = \pi_R - \pi_P$, and $u_3 = 1 - \pi_R$. Assume $(u_1, u_2, u_3) \sim \text{Dirichlet}(\alpha_1, \alpha_2, \alpha_3)$, where $0 < u_j < 1$ and $\sum_{j=1}^3 u_j = 1$. Then, the joint distribution of (π_R, π_P) is given by $f(\pi_R, \pi_P) \propto \pi_P^{\alpha_1 - 1} (1 - \pi_R)^{\alpha_3 - 1} (\pi_R - \pi_P)^{\alpha_2 - 1}$, $0 < \pi_P < \pi_R < 1$. The joint prior of (π_E, π_R, π_P) can be obtained as before by multiplying $f(\pi_R, \pi_P)$ by $f(\pi_E)$ which is $\text{Beta}(\alpha_E, \beta_E)$ and then the joint posterior of $(\pi_E, \pi_R, \pi_P | \mathbf{X})$ can be obtained by multiplying the joint posterior of $(\pi_R, \pi_P) | X_R, X_P$ with $f(\pi_E | X_E) \equiv \text{Beta}(\alpha_E + X_E, \beta_E + n_E - X_E)$ and is given by $f(\pi_E, \pi_R, \pi_P | \mathbf{X}) \propto \pi_P^{\alpha_1 + X_P - 1} (1 - \pi_P)^{n_P - X_P} \pi_R^{X_R} (1 - \pi_R)^{n_R + \alpha_3 - X_R - 1} (\pi_R - \pi_P)^{\alpha_2 - 1} \times \pi_E^{\alpha_E + X_E - 1} (1 - \pi_E)^{\beta_E + n_E - X_E}$, $0 < \pi_P < \pi_R < 1$, $0 < \pi_E < 1$. This joint posterior is not

in any standard form and hence Metropolis-Hastings acceptance-rejection sampling is required with a proposal density to generate MCMC samples from the posterior [14]. A convenient proposal density could be the product of three Beta distributions with appropriately chosen parameters.

Remark 3.1: Following [16,31], we continue to assume that AS condition, that is, $\pi_R > \pi_P$ is true. As a result truncated priors are chosen. This assumption explicitly reflects the fact that active control still retains some of its effect over placebo. In a situation when this assumption is questionable, it is not advisable to carry out a three-arm NI trial, rather a superiority trial of new treatment over placebo is more realistic.

Remark 3.2: Among the three proposed priors under Bayesian exact approach, the CBP gives equal support to the three parameters which are treated independently and is the simplest form the computational point of view. On the contrary, under PUP, the parameters π_R and π_P are made to depend on each other and in the absence of any prior information an uniform distribution is an obvious choice for π_R with restricted support to incorporate AS condition. Under the DP, more flexibility can be achieved by considering the joint distribution of π_R and π_P . However, the choice of the Dirichlet parameters is an additional burden along with its computational complexity. While we have only considered proper priors, improper priors are also possible albeit when posterior propriety holds, however, not explored here for the brevity purpose.

3.1.4. Test procedure

We formulated the test procedure to determine the experimental drug compared to the active control for the risk difference similar to the [4] who proposed the same for risk and odds ratio type functional. Under the NI setup, the common acceptable range of the effect size (θ) is $[0.5, 1)$. Hence, we can claim the NI of the test drug relative to reference drug if the posterior probability under the alternative hypothesis as mentioned in (1) exceeds some pre-defined clinically meaningful threshold, say, $R_{NI} = p^*$. Borrowing the idea from [16] (Section 3.3), the Bayesian decision rule to claim NI in this setting is defined as

$$P \left(H_1 : \frac{\pi_E - \pi_P}{\pi_R - \pi_P} > \theta \mid \pi_R > \pi_P, \mathbf{X} \right) > R_{NI}. \tag{6}$$

The probability in (6) can be calculated empirically by generating M MCMC samples from the posterior distribution of $(\pi_l \mid X_l), l \in E, R, P$. The estimated probability is given by

$$\hat{P} \left(H_1 : \frac{\pi_E - \pi_P}{\pi_R - \pi_P} > \theta \mid \pi_R > \pi_P, \mathbf{X} \right) \approx \frac{1}{M} \sum_{m=1}^M I \left(\frac{\pi_E^m - \pi_P^m}{\pi_R^m - \pi_P^m} > \theta \mid \pi_R^m > \pi_P^m \right),$$

where π_E^m, π_R^m , and π_P^m denote the m th MCMC sample, $m = 1, \dots, M$, drawn from the posterior distribution, satisfying the AS condition ($\pi_R^m > \pi_P^m$) with sufficiently large M . Note, the slight distinction with [16] previous approach (which is the direct Bayesian version of [31]) is that the usage of AS condition in the conditioning statement, which not only acting as a gate-keeper but also being used to calculate the posterior probability, yielding greater power (as proved in Lemma 2.2).

3.1.5. Sample size

The power under this NI setup can be calculated by estimating the probability of the test drug out of n^* times. Let π_E^{alt} be the value of π_E under H_1 . Hence, mathematically, the estimated power is formulated as $\widehat{\text{Power}} = (\#\text{of times } P(\pi_E^{alt} - \theta\pi_R - (1 - \theta)\pi_P > 0 \mid \pi_R > \pi_P, \mathbf{X}) > p^*)/n^*$, where the value of π_E for known values of π_R and π_P such that $(\pi_E - \pi_P)/(\pi_R - \pi_P) \in [0.5, 1.4]$. For NI testing, this ratio depends on the choice of θ which equals to $[0.5, 1)$ under H_0 and exceeds under H_1 . As discussed in Section 2.3, the minimum sample size ‘ n ’ of the arm E and other two arms corresponding to reference and placebo arms by incorporating different allocation ratios can be obtained by setting the power to be at least $100(1 - \beta)\%$. Due to generating random samples from the posterior distribution, we can notice sampling fluctuation in the results.

3.2. Approximate Bayesian approach

We next propose an approximate Bayesian approach for NI testing that incorporates the AS condition and also explicitly derive the formula for sample size determination. Note, the approximation-based approach gives a closed form of the posterior probability and hence saves the computation time of the MCMC sample generation from the posterior distribution.

3.2.1. Test procedure

We consider the Beta prior for the proportions π_l in each arm, that is, $\pi_l \sim \text{Beta}(\alpha_l, \beta_l)$, and the responses are assumed to be Binomially distributed, that is, $X_l \sim \text{Bin}(n_l, \pi_l)$, $l \in \{E, R, P\}$. The Frequentist test statistic for testing the hypothesis in (1) is given by $T = (X_E/n_E - \theta X_R/n_R - (1 - \theta)X_P/n_P)$. Under asymptotic normality assumption, we have $T \mid \mu_T \sim AN(\mu_T, \sigma_T^2)$, where $\mu_T = \pi_E - \theta\pi_R - (1 - \theta)\pi_P = (\pi_E - \pi_P) - \theta(\pi_R - \pi_P)$ and $\sigma_T^2 = \pi_E(1 - \pi_E)/n_E + \theta^2\pi_R(1 - \pi_R)/n_R + (1 - \theta)^2\pi_P(1 - \pi_P)/n_P$. Putting Normal prior on μ_T , for large sample we can approximate $\mu_T \sim AN(\mu^*, \sigma^{*2})$, where $\mu^* = E(\hat{\mu}_T) = \mu_E - \theta\mu_R - (1 - \theta)\mu_P$ and $\sigma^{*2} = \sigma_E^2 + \theta^2\sigma_R^2 + (1 - \theta)^2\sigma_P^2$, where μ_l and σ_l^2 are the mean and variance of $\text{Beta}(\alpha_l, \beta_l)$, $l \in \{E, R, P\}$. Keeping in mind the condition of AS, that is, $\pi_R > \pi_P$, we take prior on $v_T \equiv (\mu_T \mid \pi_R > \pi_P)$. Assuming $v_T \sim AN(\mu_v^*, \sigma_v^{*2})$, the posterior $v_T \mid \mathbf{X} \sim AN(\tilde{\mu}_T, \tilde{\sigma}_T^2)$, where $\tilde{\mu}_T$ and $\tilde{\sigma}_T^2$ are given as

$$\tilde{\mu}_T = \frac{T}{\sigma_T^2} + \frac{\mu_v^*}{\sigma_v^{*2}}, \quad \tilde{\sigma}_T^2 = \frac{1}{\frac{1}{\sigma_T^2} + \frac{1}{\sigma_v^{*2}}}.$$

We refer to [1] for the detailed derivation of μ_v^*, σ_v^{*2} .

Lemma 3.1: Under conditional normal approximation, the mean μ_v^* and variance σ_v^{*2} of $v_T = \pi_E - \theta\pi_R - (1 - \theta)\pi_P \mid \pi_R > \pi_P$ are given by

$$\mu_v^* = \mu_{\eta EP} + \sigma_{\eta EP} \frac{\rho}{c} \phi(a) - \theta \left(\mu_{\eta RP} + \sigma_{\eta RP} \frac{1}{c} \phi(a) \right),$$

$$\sigma_v^{*2} = \sigma_{\eta EP}^2 \left[1 + \frac{\rho^2}{c} a \phi(a) - \left(\frac{\rho}{c} \phi(a) \right)^2 \right] + \theta^2 \sigma_{\eta RP}^2 \left[1 - \frac{\phi(a)}{c} \left(\frac{\phi(a)}{c} - a \right) \right]$$

$$\begin{aligned}
 & - 2\theta \left[\sigma_{\eta EP} \sigma_{\eta RP} \frac{\rho}{c} (c + a\phi(a)) + \sigma_{\eta EP} \mu_{\eta RP} \frac{\rho}{c} \phi(a) + \sigma_{\eta RP} \mu_{\eta EP} \frac{1}{c} \phi(a) + \mu_{\eta EP} \mu_{\eta RP} \right. \\
 & \left. - \left(\mu_{\eta EP} + \sigma_{\eta EP} \frac{\rho}{c} \phi(a) \right) \left(\mu_{\eta RP} + \sigma_{\eta RP} \frac{1}{c} \phi(a) \right) \right], \tag{7}
 \end{aligned}$$

where $\mu_{\eta RP} = \pi_R - \pi_P$, $\mu_{\eta EP} = \pi_E - \pi_P$, $\sigma_{\eta EP}^2 = \sigma_E^2 + \sigma_P^2$, $\sigma_{\eta RP}^2 = \sigma_R^2 + \sigma_P^2$, $\rho = \frac{\sigma_P^2}{\sigma_{\eta EP} \sigma_{\eta RP}}$, $a = -\frac{\mu_{\eta RP}}{\sigma_{\eta RP}}$, and $c = 1 - \Phi(a)$, μ_l and σ_l^2 being the mean and variances of $\text{Beta}(\alpha_l, \beta_l)$, $l \in \{E, R, P\}$.

Proof: See Supplementary Material 3. ■

The Bayesian decision rule for deciding that the experimental treatment is non-inferior to the active comparator is given by [11]: $P(v_T \geq 0 | \mathbf{X}) \geq p^*$, where p^* is a pre-specified constant usually chosen to be 0.975 or 0.95.

3.2.2. Sample size

The sample size ‘ n ’ of the arm E under approximate Bayesian approach can be calculated by satisfying the two conditions: (C1) $P[P(v_T \geq 0 | \mathbf{X}) \geq p^* | H_0] \leq \alpha$, (C2) $P[P(v_T \geq 0 | \mathbf{X}) \geq p^* | H_1] \geq 1 - \beta$, where the probability in (C1) is the estimated Bayesian version of average type-I error while that in (C2) is the estimated power of the test, β being the type-II error. The sample size ‘ n ’ is determined from (C2) by fixing β to have at least $100(1 - \beta)\%$ power of the test and simultaneously satisfying (C1). As in the Frequentist approach, we choose $\alpha = 0.025$. We note that

$$\begin{aligned}
 P(v_T \geq 0 | \mathbf{X}) &= P\left(\frac{v_T - \tilde{\sigma}_T^2 \tilde{\mu}_T}{\tilde{\sigma}_T} > \frac{-\tilde{\sigma}_T^2 \tilde{\mu}_T}{\tilde{\sigma}_T}\right) \geq p^* \\
 \Leftrightarrow -\tilde{\sigma}_T \tilde{\mu}_T < z_{1-p^*} &\Leftrightarrow T > -z_{1-p^*} \left(\frac{1}{\sigma_T^2} + \frac{1}{\sigma_v^{*2}}\right)^{1/2} \sigma_T^2 - \frac{\mu_v^*}{\sigma_v^{*2}} \sigma_T^2,
 \end{aligned}$$

where z_{1-p^*} is the $100(1 - p^*)\%$ of the $N(0, 1)$ distribution. Now the power function is obtained by varying π_E such that $0.5 \leq (\pi_E - \pi_P)/(\pi_R - \pi_P) \leq 1.4$, keeping the other proportions π_R , π_P , and θ fixed. Let us denote μ_T and σ_T^2 by μ_T^{null} and $\sigma_T^{2\text{null}}$, respectively, under H_0 , and similarly under H_1 denote the respective quantities by μ_T^{alt} and $\sigma_T^{2\text{alt}}$. Thus condition (C1) can be rewritten in terms of T as

$$\begin{aligned}
 P_{H_0} \left[T > -z_{1-p^*} \left(\frac{1}{\sigma_T^{2\text{null}}} + \frac{1}{\sigma_v^{*2}}\right)^{1/2} \sigma_T^{2\text{null}} - \frac{\mu_v^*}{\sigma_v^{*2}} \sigma_T^{2\text{null}} \right] &\leq \alpha, \\
 \Leftrightarrow P_{H_0} \left[\frac{T - \mu_T^{\text{null}}}{\sigma_T^{\text{null}}} > \left(-z_{1-p^*} \left(\frac{1}{\sigma_T^{2\text{null}}} + \frac{1}{\sigma_v^{*2}}\right)^{1/2} \sigma_T^{2\text{null}} - \frac{\mu_v^*}{\sigma_v^{*2}} \sigma_T^{2\text{null}} - \mu_T^{\text{null}}\right) / \sigma_T^{\text{null}} \right] &\leq \alpha, \\
 \Leftrightarrow \Phi \left(z_{1-p^*} \left(\frac{1}{\sigma_T^{2\text{null}}} + \frac{1}{\sigma_v^{*2}}\right)^{1/2} \sigma_T^{\text{null}} + \frac{\mu_v^*}{\sigma_v^{*2}} \sigma_T^{\text{null}} + \frac{\mu_T^{\text{null}}}{\sigma_T^{\text{null}}} \right) &\leq \alpha. \tag{8}
 \end{aligned}$$

Similarly, condition (C2) becomes

$$\Phi \left(z_{1-p^*} \left(\frac{1}{\sigma_T^{2alt}} + \frac{1}{\sigma_v^{*2}} \right)^{1/2} \sigma_T^{alt} + \frac{\mu_v^*}{\sigma_v^{*2}} \sigma_T^{alt} + \frac{\mu_T^{alt}}{\sigma_T^{alt}} \right) \geq 1 - \beta. \tag{9}$$

Now ‘ n ’ can be solved from (9) by setting $\beta = 20\%$ and simultaneously satisfying (8). We vary π_E^{alt} (which is included in μ_T^{alt}) to get minimum sample size satisfying at least 80% power for each π_E^{alt} . The sample size for the arms R and P can be obtained considering the allocation ratios r_1 and r_2 as discussed earlier.

4. Bayesian–Frequentist connection in three-arm trial

In this section, we connect the Bayesian and Frequentist approaches by transforming the Bayesian posterior probability of the tested hypothesis into the Frequentist probability of Bernoulli trial after adjusting the number of events and population sizes. This section is motivated from the work of [36] who showed similar connection for two-arm trial with integer-valued hyper-parameters, by linking Frequentist p -values and Bayesian conditional measure of evidence [2,7]. This work also offers additional insight about effective sample size gain in Bayesian set up under conjugate prior specification. We consider the CBP setting; that is, $X_l | \pi_l \sim \text{Bin}(n_l, \pi_l)$, prior $\pi_l \sim \text{Beta}(\alpha_l, \beta_l)$, and the posterior distribution $\pi_l | X_l \sim \text{Beta}(\alpha_l + X_l, n_l - X_l + \beta_l)$, $l \in \{E, R, P\}$, with the restriction that the hyper-parameters are integers. The Bayesian decision rule to declare NI of the test drug over the reference given the AS condition ($\pi_R > \pi_P$) holds, as given in Section 3.1.4, can be written as

$$P[\pi_E - \theta\pi_R - (1 - \theta)\pi_P > 0 | \pi_R > \pi_P, \mathbf{X}] > p^*. \tag{10}$$

Define, $\eta_{RP} = \pi_R - \pi_P$. Now, since the probability in (10) does not have a closed form, it is approximated by generating posterior samplers as in the following:

$$\begin{aligned} P(\pi_E - \theta\pi_R - (1 - \theta)\pi_P > 0 | \pi_R > \pi_P, \mathbf{X}) &= P((\pi_E - \pi_P) > \theta(\pi_R - \pi_P) | \pi_R > \pi_P, \mathbf{X}) \\ &= \int_0^\infty P(\pi_E - \pi_P > \theta c | \mathbf{X}) f_{\eta_{RP} | \eta_{RP} > 0}(c) dc \approx \frac{1}{M} \sum_{i=1}^M g(\theta c_i, \mathbf{X}), \end{aligned} \tag{11}$$

where $g(\theta c_i, \mathbf{X}) = P(\pi_E - \pi_P > \theta c_i | \mathbf{X})$, c_i being the i th sampled value of $\pi_R - \pi_P | (\pi_R > \pi_P)$, and \mathbf{X} denotes the relevant section of the data. To obtain $P(\pi_E - \pi_P > \theta c | \mathbf{X})$, we refer to [36] and present the following two theorems that link the Frequentist and Bayesian approaches and can be used to estimate the probability in (11).

Theorem 4.1: *Let $P_F(\cdot)$ be Fisher’s exact one-sided probability for testing $H_0 : \theta_1 \leq \theta_2$ versus $H_1 : \theta_1 > \theta_2$, θ_i , $i = 1, 2$ being the Binomial rates for the two populations involving n_1 and n_2 individuals, respectively, x_i , $i = 1, 2$ are the total number of events in two respective populations; (a_i, b_i) , $i = 1, 2$ are the hyper-parameters of the Beta priors on θ_i , $i = 1, 2$, and let $P_B(\cdot)$ be the probability of the same hypothesis under the Bayesian paradigm, then the*

following holds true:

$$P_B(\theta_1 \leq \theta_2 | x_1, n_1, x_2, n_2) = P_F(\theta_1 \leq \theta_2 | x_1 + a_1, n_1 + a_1 + b_1 - 1, x_2 + a_2 - 1, n_2 + a_2 + b_2 - 1).$$

Theorem 4.2: Let $x_1 + a_1 > 0, x_2 + a_2 > 0, n_1 + b_1 - x_1 - 1 > 0$. Then, for a sufficiently small δ ,

$$P_B(\theta_1 + \delta \leq \theta_2 | x_1, n_1, x_2, n_2) = P_F(\theta_1 \leq \theta_2 | x_1 + a_1, n_1 + a_1 + b_1 - 1, x_2 + a_2 - 1, n_2 + a_2 + b_2 - 1) \delta \times h(\alpha_1, \beta_1, \alpha_2, \beta_2, n_1, n_2, \mathbf{X}) + o(\delta),$$

where $h(\alpha_1, \beta_1, \alpha_2, \beta_2, n_1, n_2, \mathbf{X}) = \Gamma(a_1 + b_1 + n_1)\Gamma(a_2 + b_2 + n_2) / \Gamma(a_1 + x_1)\Gamma(b_1 + n_1 - x_1)\Gamma(a_2 + x_2)\Gamma(b_2 + n_2 - x_2) \times \Gamma(a + x - 1)\Gamma(b + n - x - 1) / \Gamma(a + b + n - 2)$, (\cdot) stands for summation. The probabilities $P_F(\cdot)$ and $P_B(\cdot)$, θ_i , and (a_i, b_i) , $i = 1, 2$ remain same as defined in Theorem 4.1.

We give the following proposition using the identities in the above two theorems, which can be used to obtain the probability $P_B(\pi_E - \pi_P > \theta c | \mathbf{X})$.

Proposition 4.1: Taking $\theta_1 = \pi_P, \theta_2 = \pi_E, a_1 = \alpha_P, a_2 = \alpha_E, b_1 = \beta_P, b_2 = \beta_E, x_1 = x_P, x_2 = x_E, n_1 = n_P, \text{ and } n_2 = n_E$, for a fixed value of $c, g(\theta c, \mathbf{X}) = P(\pi_E - \pi_P > \theta c | \mathbf{X})$ can be estimated using Theorems 4.1 and 4.2 as

$$\begin{aligned} P_B(\pi_E - \pi_P > \theta c | x_P, n_P, x_E, n_E) &= P_B(\pi_P + \theta c < \pi_E | x_P, n_P, x_E, n_E) \\ &= P_F(\pi_P + \theta c < \pi_E | x_P + \alpha_P, n_P + \alpha_P + \beta_P - 1, x_E + \alpha_E - 1, n_E + \alpha_E + \beta_E - 1) - \theta ch(\alpha_E, \beta_E, \alpha_P, \beta_P, n_P, n_E, \mathbf{X}), \end{aligned} \tag{12}$$

$h(\cdot)$ is as given in Theorem 4.2. This can be repeated for each $c_i, i = 1, \dots, M$ to obtain the probability in (11).

Another way of linking the Frequentist and Bayesian approach can also be found from the following identities ([36]) which can be used to approximate the incomplete Beta integral by sum:

$$\frac{\Gamma(n + a + b)}{\Gamma(a + x)\Gamma(n + b - x)} \int_p^1 k^{a+x-1}(1 - k)^{b+n-x-1} dk = \sum_{i=0}^{x^F-1} \binom{n^F}{i} p^i (1 - p)^{n^F-i}, \tag{13}$$

$x^F = x + a$ and $n^F = n + a + b - 1$. The identity in (13) can be used to approximate $g(\theta c | \mathbf{X}) = P_B(\pi_E - \pi_P > \theta c | \mathbf{X})$ in (11) as given in the following proposition.

Proposition 4.2: Taking $p = \pi_E, a = \alpha_P, b = \beta_P, x = x_P, \text{ and } n = n_P, P_B(\pi_P < \pi_E | x_P, n_P, x_E, n_E)$ can be approximated by the sum of gamma functions using the identity

in (13) as

$$\begin{aligned}
 &P_B(\pi_P < \pi_E \mid x_P, n_P, x_E, n_E) \\
 &= 1 - \frac{\Gamma(n_E + \alpha_E + \beta_E)}{\Gamma(\alpha_E + x_E)\Gamma(n_E - x_E + \beta_E)} \\
 &\quad \times \sum_{i=0}^{\alpha_P + x_P - 1} \frac{\Gamma(\alpha_E + x_E + i)\Gamma(n_E + \beta_E - x_E - 1 + n_P + \alpha_P + \beta_P - i)}{\Gamma(n_E + \alpha_E + \beta_E + n_P + \alpha_P + \beta_P - 1)}.
 \end{aligned}$$

Thus, for a fixed c , $P(\pi_E - \pi_P > \theta c \mid \mathbf{X})$ can be calculated using Theorem 4.1 and 4.2 as $P_B(\pi_E - \pi_P > \theta c \mid \mathbf{X}) = P_B(\pi_P < \pi_E \mid n_P, n_E, \mathbf{X}) - \theta c \times h(\alpha_E, \beta_E, \alpha_P, \beta_P, n_E, n_P, \mathbf{X})$, where the function $h(\cdot)$ is given in Theorem 4.2.

Proof: See Supplementary Material 4. ■

Repeating the calculation of $P_B(\pi_E - \pi_P > \theta c \mid \mathbf{X})$ for each $c_i, i = 1, \dots, M$, one can obtain the posterior probability of NI hypothesis given AS condition from (11). Similar to [36], the Bayesian test of significance is equivalent to Fisher’s exact test with adjusted value of the parameters. This is characterized as the effective sample size change in the literature [3,28].

5. Simulation and sample size calculation

In this section, we enumerate simulation studies to evaluate the performance of the Bayesian as well as Frequentist procedures presented above. The power curves are generated for the test considering three different priors under exact Bayesian, under Frequentist as well as Bayesian approximation procedures. For the exact Bayesian approach, power curves are compared under the informative and non-informative Beta priors. In the latter part of the section focuses on sample size calculation for the assessment of NI to attain the desired power under three approaches: (1) Frequentist normal approximation, (2) Bayesian normal approximation, and (3) Bayesian exact approach for the three-arm NI testing.

5.1. Steps for simulation

The following simulation steps are used to calculate the type-I error and power for the three different prior scenarios described earlier: (1) Conjugate Beta-Binomial, (2) PUP, and (3) DP. For the CBP setting, we assume a non-informative prior for the proportions in each of the three-arms; that is, $\pi_l \sim \text{Beta}(1, 1), l \in \{E, R, P\}$. We also consider an informative Beta prior so that the mode of the Beta distribution equals the parameter and compared the power between non-informative with the informative prior. For PUP, we consider the non-informative Beta priors for the experimental arm (π_E) and the placebo arm (π_P), while π_R is generated from truncated Beta with the support on $(\pi_P, 1]$. Finally, for the DP, we put non-informative Beta prior on π_E and choose suitable values for the Dirichlet parameters. We consider a randomized trial with the sample size allocation ratio as $n_E : n_R : n_P = 1 : r_1 : r_2$. In following we give the steps for the simulation as discussed in [4]:

- S1: Specify n_E, n_R, n_P (or, the allocation ratios), $\pi_l, l \in \{E, R, P\}$ with $\pi_R > \pi_P$, and θ and vary π_E such that $\pi_P + 0.5(\pi_R - \pi_P) \leq \pi_E \leq \pi_P + 1.4(\pi_R - \pi_P)$ to generate $\mathbf{X} = \{X_E, X_R, X_P\}$.
- S2: Generate $X_l \sim \text{Binomial}(n_l, \pi_l), l \in \{E, R, P\}$ for given values of the ratio $(\pi_E - \pi_P)/(\pi_R - \pi_P)$ or π_E .
- S3: For exact Bayesian approach, M many MCMC samples are generated from the posterior distribution based on the priors as mentioned in Section 3 satisfying the AS condition $\pi_R > \pi_P$. For Frequentist and approximate Bayesian cases, we disregard this step. For the PUP and the DP, the posterior sample values satisfy $\pi_R > \pi_P$ automatically because of the in-built restriction.
- S4: For each posterior samples, compute the ratio $(\pi_E - \pi_P)/(\pi_R - \pi_P)$ and estimate the posterior probability is as follows:

$$P\left(\frac{\pi_E - \pi_P}{\pi_R - \pi_P} > \theta \mid \pi_R > \pi_P, \mathbf{X}\right) \approx \frac{1}{M} \sum_{m=1}^M I\left(\frac{\pi_E^m - \pi_P^m}{\pi_R^m - \pi_P^m} > \theta \mid \pi_R^m > \pi_P^m, \mathbf{X}\right).$$

- S5: Set Count = 0 and increase Count by 1 if the posterior probability $> p^*$, otherwise, Count = 0.
- S6: Repeat the steps S2 to S5 for a large number say n^* times. Calculate type-I error and power by using COUNTS divided by n^* . For type-I error calculation π_E should satisfy $(\pi_E - \pi_P)/(\pi_R - \pi_P) = \theta$, and for power calculation, π_E should be $(\pi_E - \pi_P)/(\pi_R - \pi_P) > \theta$.
- S7: Based on the estimated power from the step S6, the power curve can be plotted for sequence of π_E satisfying the condition $0.5 \leq (\pi_E - \pi_P)/(\pi_R - \pi_P) \leq 1.4$.

Note that for Frequentist and Bayesian approximation approaches, S4 and S5 are replaced by the corresponding decision rule as mentioned in Sections 2.2 and 3.2.1, respectively.

5.2. Simulation result

For the CBP and PUP, since the posterior is available in closed form, we chose the number of posterior samplers M to be 1000. For the DP, we have determined the number of MCMC samples to be $M = 1000$ taking every 50th value of 50,000 MCMC samples. We assume non-informative Beta(1, 1) prior for the three-arms for the CBP setting. Throughout the simulation study, we consider the following specification of the parameters: $\pi_R = 0.7, \pi_P = 0.1$, and we set π_E such that $(\pi_E - \pi_P)/(\pi_R - \pi_P) \in [0.5, 1.4]$. We consider several values for $n_l, l \in \{E, R, P\}$ such that $n_E : n_R : n_P = n : nr_1 : nr_2, 'n'$ being the common sample size. Unequal allocation is also possible as will be described in Section 5.3. Another important criterion is the choice of p^* which we fixed at 0.975. However, as reported in [13] this choice could give too restrictive type-I error in the Bayesian context. One way to alleviate this problem is to perform Bayesian calibration, but is not pursued to reduce the computational burden.

In Figure 1(a), we present four power curves corresponding to different values of θ : 0.8, 0.7, 0.6, and 0.5 and $n = 100$ under Bayesian conjugate non-informative Beta(1, 1) prior for each arm. The three values of θ correspond to the three choices of NI margin which are

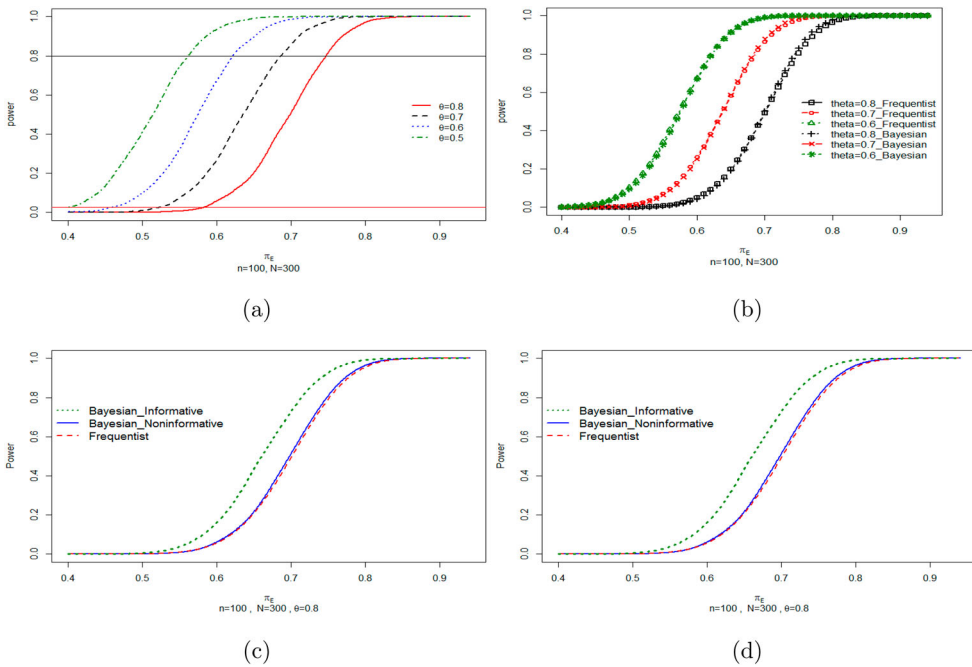


Figure 1. Power curves for different θ under Bayesian conjugate prior on the top left (a) and under Frequentist and Bayesian normal approximation on the top right (b). Comparison of Frequentist, Bayesian non-informative, and informative power curves at the bottom left (c). Comparison of power curves under CBP, PUP, and DP at the bottom right (d).

the maximum clinically relevant difference: $\delta = -0.2(\pi_R - \pi_P)$, $\delta = -0.3(\pi_R - \pi_P)$, and $\delta = -0.4(\pi_R - \pi_P)$, that is, f is chosen as -0.2 , -0.3 , and -0.4 , respectively. This indicates that in order to be non-inferior, the experimental drug with respect to placebo, must attain more than 80%, 70%, and 60%, respectively, of the effect of the reference drug as compared to the placebo. Also, for the proposed test, we can infer that the smaller value of θ is more powerful than that of higher θ because of the easier declaration of NI of the experimental drug. In Figure 1(b), we plot the power curves for a balanced study design with a common sample size $n = 100$ for the Frequentist approach and approximate Bayesian approach under non-informative prior. From Figure 1(b), we observe that both the methods produce almost similar power curves. Although the Bayesian power curve is slightly above the Frequentist one, from our experience, this should be the case under a flat prior. In Figure 1(c), we do a comparison among the power curves obtained under the Frequentist approach, Bayesian exact approach under non-informative prior, and the same for informative prior with the common sample size $n = 100$ in each arm. For the informative prior, we put the prior in each arm as follows: E : Beta(40, 17.71), R : Beta(40, 17.71), and P : Beta(2, 10) and for the non-informative prior, we consider the same Beta(1, 1) prior as earlier. We note that these priors are chosen so that the mode of Beta distribution is equal to the value of the corresponding proportion parameter. We observe that Bayesian power curve under non-informative prior is almost similar to that of the Frequentist power curve, while the Bayesian power curve under informative prior is much higher.

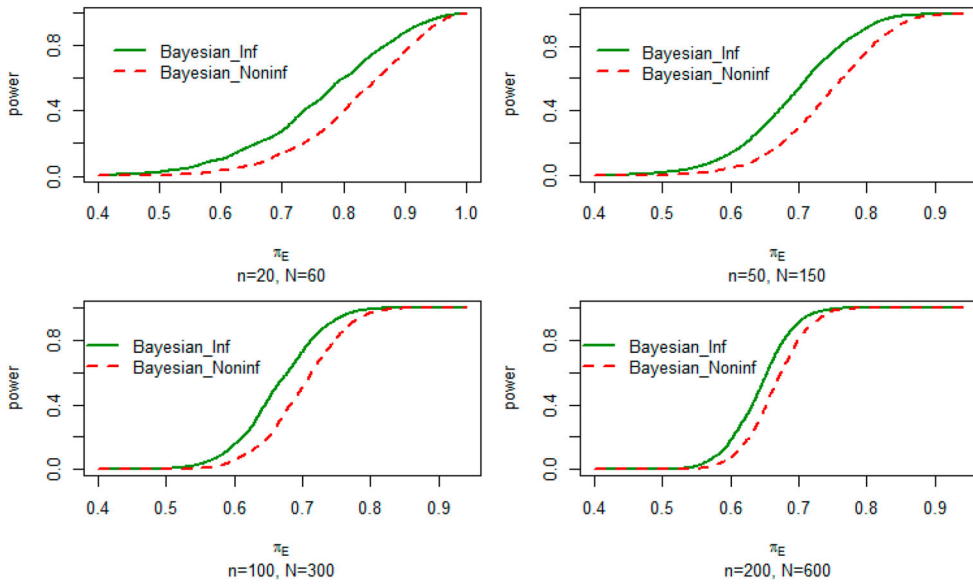


Figure 2. Comparison of informative vs. non-informative power curves under conjugate Beta prior, $\theta = 0.8$.

The gain in power by using informative priors is also depicted in Figure 2, where we give four plots for $n = 20, 50, 100$, and 200 to compare the informative with non-informative power curves under CBP setting. For the DP, with Dirichlet parameters $(1, 1, 1)$ and Beta(1, 1) prior for each arm, the test is too conservative for $(\pi_E - \pi_P)/(\pi_R - \pi_P) \leq \theta$ and yields type-I error close to 0. However, in our experience it is possible to choose the Dirichlet parameters so that type-I error becomes close to 0.025, thus yielding better power as compared to CBP and PUP. This is depicted in Figure 1(d). We have chosen the Dirichlet parameters $(\alpha_1, \alpha_2, \alpha_3)$ as $(2, 3, 6)$ which gives the marginal distribution for π_R and π_P as: $\pi_P \sim \text{Beta}(2, 9)$ and $\pi_R \sim \text{Beta}(5, 6)$. To make the power curves comparable we chose the same priors under CBP and PUP setup. In all three cases, the arm E is given Beta(1, 1) prior. From Figure 1(d), we see that the DP outperforms the CBP and PUP. The latter yields the least power among the three.

5.3. Sample size

We refer to the Sections 2.3, 3.1.5, and 3.2.2, respectively, for the sample size determination under Frequentist, exact Bayesian, and approximate Bayesian approaches. We determine the sample size $n_l, l \in \{E, R, P\}$ setting the power at $(1 - \beta)$ with β as the pre-specified type-II error. Let us consider $n_P = n, n_R = r_1 n$, and $n_E = r_2 n$ with $r_1, r_2 > 0$. To calculate the sample size for each arm, we explore three possible allocations for experimental, reference, and placebo arms, (i) (1:1:1) with $r_1 = r_2 = 1$; (ii) (2:2:1) with $r_1 = 1, r_2 = 1/2$; and (iii) (3:2:1) with $r_1 = 2/3, r_2 = 1/3$ of the total sample size N . The sample size is calculated as the smallest 'n' which satisfies $power \geq 1 - \beta$. To make a comparison of the existing Frequentist approach with the proposed conditional one, first we present the sample sizes under both the approaches in Table 2. For simplicity, we only consider equal allocation to the three treatment arms.

Table 1. Sample size for marginal vs. conditional Frequentist approach.

θ	π_E	Existing ($\pi_R = 0.7, \pi_P = 0.1$)				Conditional ($\pi_R = 0.6, \pi_P = 0.55$)			
		n_P	N	n_P	N	n_P	N	n_P	N
0.8	0.90	26	78	26	78	30	90	28	84
	0.85	38	114	38	114	43	129	41	123
	0.80	58	174	58	174	68	204	64	192
	0.75	99	297	99	297	120	360	114	342
	0.70	203	609	203	609	257	771	248	744
	0.65	604	1812	604	1812	875	2625	866	2598
0.7	0.90	17	51	17	51	27	81	26	78
	0.85	24	72	24	72	39	117	38	114
	0.80	34	102	34	102	61	183	60	180
	0.75	51	153	51	153	106	318	104	312
	0.70	85	255	85	255	222	666	218	654
	0.65	165	495	165	495	703	2109	698	2094

We determine the sample size under the two approaches for $\theta = \{0.8, 0.7\}$ with $(\pi_R = 0.7, \pi_P = 0.1)$ and $(\pi_R = 0.6, \pi_P = 0.55)$. From Table 1, we observe that for $\pi_R = 0.7$ and $\pi_P = 0.1$ the sample size under the conditional approach is identical to that calculated under the marginal approach, while for $\pi_R = 0.6$ and $\pi_P = 0.55$, the sample size under the conditional approach is smaller than the existing one to achieve a power of 80%. This observation implies for smaller difference between π_R and π_P , the proposed conditional approach is more powerful, while for larger difference both the approaches behave similarly. This fact supports the claim proved in Lemma 2.2. In rest of the sample size calculation, only conditional Frequentist approach is considered as it is the more powerful than the marginal approach. In Table 2, we demonstrate the sample sizes under our proposed approaches with $\pi_R = 0.7$ and $\pi_P = 0.1$. Similar to [4,5], we assign $\alpha = 0.025$ for Frequentist and the sample sizes satisfying $power \geq 1 - \beta$ also allow estimated type-I error of at most $\alpha = 0.025$ for Bayesian exact and approximate methods.

In Table 2, total sample sizes for three allocations are calculated based on the sample size of the placebo arm, n_P . For example, the total sample size corresponding to the allocation ratio 1:1:1 is $3n_P$ whereas for 2:2:1 and 3:2:1, the total sample sizes are $5n_P$ and $6n_P$, respectively. As discussed in [4,5], one might not consider balanced design due to the ethical reason and the smaller difference between E and R compared to the difference from placebo. From Table 2, we observe smaller sample size for the unbalanced allocation (2 : 2 : 1) as compared to the balanced design (1 : 1 : 1). Similarly, we notice a minor reduction in sample size for the unbalanced case (3 : 2 : 1) as compared to (2 : 2 : 1). A similar interpretation can be found in Figure 3 where the power curves show three different allocations under Frequentist and exact Bayesian approaches with non-informative prior with $N = 300$. We note that the type-I error rate is exactly 0.025 for the Frequentist approach and always maintained below 0.025 for the Bayesian approaches (Table 2). In cases where $\alpha \ll 0.025$, Bayesian calibration can be performed to improve sample size, but is not explored in the current paper.

6. Application in a real data

To illustrate the real data application, we revisited major depressive disorder data described in [19]. This dataset has been analyzed in many articles, including [15,18]. Chowdhury

Table 2. Frequentist and Bayesian sample sizes to achieve a power of 80% for $\theta = \{0.8, 0.7\}$, $\alpha = 0.025$, and $\pi_E \in [0.65, 0.9]$, keeping $\pi_R = 0.7$ and $\pi_P = 0.1$ under three different allocations.

E	Allocation		θ	π_E	Frequentist normal		Bayesian normal			Bayesian exact			
	R	P			n_P	N	n_P	N	$\hat{\alpha}$	n_P	N	$\hat{\alpha}$	
1	1	1	0.8	0.90	26	78	20	60	0.014	21	63	0.020	
				0.85	38	114	32	96	0.017	33	99	0.022	
				0.80	58	174	52	156	0.019	52	156	0.025	
			0.7	0.75	99	297	93	279	0.021	93	279	0.024	
				0.70	203	609	196	588	0.022	184	552	0.025	
				0.85	24	72	20	60	0.015	21	63	0.025	
	2	2	1	0.8	0.80	34	102	30	90	0.018	29	87	0.021
					0.75	51	153	48	144	0.020	47	141	0.023
					0.70	85	255	82	246	0.021	80	240	0.024
				0.7	0.65	165	495	162	486	0.023	158	474	0.023
					0.90	13	65	10	50	0.014	11	55	0.023
					0.85	19	95	16	80	0.017	17	85	0.015
2	2	1	0.8	0.80	30	150	27	135	0.019	27	135	0.025	
				0.75	50	250	47	235	0.021	47	235	0.021	
				0.70	103	515	99	495	0.022	94	470	0.024	
			0.7	0.85	12	60	10	50	0.015	11	55	0.023	
				0.80	18	90	16	80	0.018	17	85	0.025	
				0.75	26	130	25	125	0.020	26	130	0.020	
	3	2	1	0.8	0.70	44	220	42	210	0.021	43	215	0.023
					0.65	85	425	83	415	0.023	83	415	0.022
					0.90	11	66	9	54	0.015	9	54	0.022
				0.7	0.85	16	96	14	84	0.017	14	84	0.023
					0.80	24	144	22	132	0.019	22	132	0.025
					0.75	40	240	39	234	0.021	37	222	0.024
3	2	1	0.8	0.70	81	486	80	480	0.022	79	474	0.024	
				0.85	10	60	9	54	0.016	9	54	0.014	
				0.80	14	84	13	78	0.018	13	78	0.019	
			0.7	0.75	21	126	20	120	0.020	20	120	0.020	
				0.70	34	204	33	198	0.021	34	204	0.016	
				0.65	66	396	65	390	0.023	64	384	0.018	

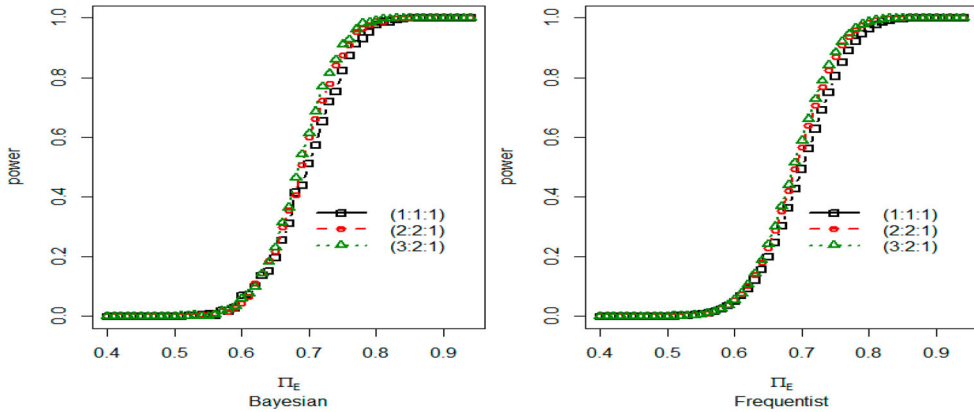


Figure 3. Power curves for different allocations, $\theta = 0.8$.

et al. [4,5] used this dataset for binary outcome with risk and odds ratio type functional. In this analysis, we implement our proposed methods for risk difference purpose. Briefly, the primary endpoint, HAMD-17 total score, was a continuous scale explaining the change from baseline at the end of sixth week with three arms: duloxetine ($n_E = 147$), paroxetine

($n_R = 148$), and placebo ($n_P = 145$). We consider two binary outcomes, Response and Remission which are presented in Table 3. As described in [18], Response is defined as the reduction of more than 50% change of the total score at the end of six week. Remission is defined as maintaining HAMD-17 score of less than 17 at the same end-point. For the existing Frequentist approach, the p -value of the test is calculated as

$$p\text{-value} = P_{H_0}(T > T_{\text{obs}}) = 1 - \Phi\left(\frac{T_{\text{obs}}}{\sqrt{\sigma_T^{2\text{null}}}}\right), \tag{14}$$

where $T = \hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P$ is the Frequentist statistic under the existing approach, T_{obs} is the observed value of T , and $\sigma_T^{2\text{null}}$ is the variance of T under null hypothesis. For the conditional Frequentist approach, we calculate the p -value as

$$p\text{-value} = P_{H_0}(W > W_{\text{obs}}) = 1 - \Phi\left(\frac{W_{\text{obs}} - \mu_w^{\text{null}}}{\sigma_w^{\text{null}}}\right), \tag{15}$$

where $W = (\hat{\pi}_E - \theta\hat{\pi}_R - (1 - \theta)\hat{\pi}_P) | \hat{\pi}_R > \hat{\pi}_P$ is the Frequentist test statistic for the conditional testing, W_{obs} is the observed value of W , and μ_w^{null} and $\sigma_w^{2\text{null}}$ are the mean and variance of W under null hypothesis as given in Section 2. For the Bayesian approach, we start with non-informative priors and then consider informative priors to compare the results. We use $p^* = 0.975$ to determine NI of duloxetine over paroxetine. The Frequentist p -values are compared with $\alpha = 0.025$ to deduce the decision. Assuming non-informative Beta(1, 1) prior for $\pi_l, l \in \{E, R, P\}$ the samplers are generated for the three rates from Beta distributions as in Step 3 of the simulation. We calculate the posterior probability $P(H_1 | X)$ for the rejection of H_1 which is the quantity estimated in Step 4 of the simulation. This is reported in Table 4 for different values of $\theta \in [0.5, 1)$, in order to ensure that the test drug has meaningful clinical effect retention. These posterior probabilities are compared with p^* to deduce the Bayesian decision. We also checked that the AS condition holds with probability close to 1 for both the Response and the Remission outcome. From Table 4, we observe that the Frequentist p -values decrease while the posterior probabilities increase as θ decreases implying greater chance of declaring NI for smaller values of θ , which is compatible with the simulation results observed in Section 5. Also, we observe that the p -values under the conditional approach is smaller or at most equal to that under the marginal approach which is consistent to the Lemma 2.2. However, since none of the p -values is smaller than $\alpha = 0.025$, NI hypothesis cannot be rejected and hence non-inferiority cannot be claimed for any θ . As evident, the Remission data has lower posterior probabilities than Response data. Using non-informative Beta prior, the posterior probabilities are less than the pre-specified cutoff $p^* = 0.975$ and hence the NI of E relative to R cannot be claimed. However, when we choose an informative Beta prior, E : Beta(40, 34), R : Beta(40, 36), and P : Beta(40, 64), NI is established for $\theta \leq 0.55$, for Response data. Similarly, taking the priors as E : Beta(40, 77), R : Beta(40, 80), and P : Beta(40, 141), NI is claimed for $\theta = 0.5$ for the Remission data. For the PUP on the arm R and with Beta(1, 1) prior on the arms E and P , we obtain results very similar to the CBP set-up. However, choosing informative priors for the arms E and P as in the CBP one can claim NI for $\theta = 0.5$. Finally considering DP, with parameters (1, 1, 1) along with non-informative Beta prior for E , the posterior probabilities are found to be too small, even smaller than the CBP or PUP, to claim NI. However,

Table 3. Remission and Response as outcome in the depression trial [19].

Outcome	Duloxetine	Paroxetine	Placebo
Remission	50	49	32
Response	80	78	56
Total	$n_E = 147$	$n_R = 148$	$n_P = 145$

Table 4. Frequentist p -values and Bayesian posterior probabilities under different informative (Info) and non-informative (Non-Info) priors.

θ	Freq-Marg- p	Freq-Cond- p	CBP Non-info	CBP Info	PUP Non-info	PUP Info	DP Non-info	DP Info
Response data								
0.80	0.198	0.195	0.810	0.845	0.687	0.832	0.370	0.910
0.75	0.159	0.157	0.836	0.879	0.746	0.864	0.397	0.928
0.70	0.125	0.124	0.871	0.911	0.791	0.907	0.437	0.947
0.65	0.097	0.096	0.908	0.942	0.839	0.938	0.468	0.956
0.60	0.073	0.073	0.933	0.962	0.872	0.954	0.506	0.970
0.55	0.055	0.055	0.944	0.976*	0.892	0.967	0.546	0.981*
0.50	0.040	0.040	0.955	0.985*	0.923	0.984*	0.592	0.985*
Remission data								
0.80	0.265	0.259	0.723	0.778	0.718	0.789	0.158	0.809
0.75	0.225	0.220	0.776	0.832	0.758	0.831	0.168	0.861
0.70	0.188	0.184	0.811	0.872	0.797	0.874	0.183	0.896
0.65	0.154	0.154	0.832	0.905	0.829	0.908	0.198	0.924
0.60	0.124	0.122	0.872	0.935	0.859	0.946	0.212	0.950
0.55	0.098	0.097	0.899	0.953	0.888	0.964	0.228	0.971
0.50	0.077	0.076	0.916	0.976*	0.914	0.979*	0.249	0.985*

Denotes the posterior probability is greater than $p^ = 0.975$.

if the Dirichlet parameters are chosen to be (60, 22, 73) with $Beta(40, 36)$ for the arm E for the Response data, NI can be claimed for $\theta \leq 0.55$. Similarly, with Dirichlet parameters (150, 75, 450) with $Beta(80, 160)$ for the arm E , NI can be claimed for $\theta = 0.5$ for the Remission data. We note, here, that the Dirichlet parameters as well as the informative priors under CBP or PUP are so chosen that the mean of the Beta distribution coincides with the estimates of proportion parameters. A point to note, the choice of informative priors cannot be set arbitrarily in practice to claim NI, rather it must be guided from available and verifiable sources.

7. Conclusion

In this paper, we have presented new Frequentist and Bayesian test procedures for the ‘gold standard’ three-arm NI trial which includes a placebo arm. We focused primarily on binary outcome with risk difference being the metric of comparison. In the Frequentist setup, we introduce a more powerful conditional test of NI which makes more intuitive sense with a reduction in sample size requirement under certain situations. In our proposed methods, we explored the fraction margin approach with unknown NI margin, δ , which can be fluctuating based on the effect size of the treatment. On the other hand, the three-arm fixed margin approach of [18] is based on joint testing which requires additional attention for decision making as it may result in a biased test (see [6] for Intersection Union test and [17]). We provided sample size estimation for the three-arms of NI trial under

three types of allocation (in E, R, P) using all three approaches. We have seen that even with the non-informative prior Bayesian normal approximation, as well as Bayesian exact approach yields greater or equal power as compared to Frequentist approach. The sample sizes using Bayesian approaches are smaller than that of the Frequentist approach for the desired power of 80%. From our investigation it is evident that an unbalanced allocation of the total sample size in NI trial results in the reduction of the required number of patients to achieve a fixed power. According to [31] an unbalanced allocation of the total sample size in a NI trial is desirable from an ethical point of view. Besides these technical aspects, NI trial has to be reflected in several substantive respects. The concerns include the choice of δ , the question of whether a placebo can be included as an additional arm of the study, AS, to give a few examples among others.

The results of the real clinical trial data suggest that the exact Bayesian methods perform favorably in all situations, and that these methods do not rely on any asymptotic approximation. Notably, with binary end-points, risk difference is not the only function of interest. One may also frame both a two-arm as well as a three-arm hypothesis in terms of log odds and/or relative risk ratios. For two-arm trial [30] proposed a fully Bayesian method for such metrics. Their method is based on a fixed margin-based approach, where margin construction was not the priority. Our group recently published (see [4,5]) conditional Frequentist and Bayesian test for risk ratio, odds ratio and number needed to treat which uses a similar approach as in the current paper, albeit, without any direct mathematical proof of power gain. The effect of prior miss-specification in the NI context is also an open area of research. Robust prior in the form of mixture distribution could lead to more stable and less sensitive result. Another interesting extension could be semi or non-parametric extensions of our approach. Ghosh *et al.* [16] proposed a semi-parametric extension of the Bayesian test procedure for continuous outcomes, which can be further extended for the binary responses.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The research of last author is partly supported by PCORI [contract number ME-1409-21410] and NIH [grant number P30-ES020957].

ORCID

Erina Paul  <http://orcid.org/0000-0003-2172-5644>

References

- [1] B.C. Arnold, R.J. Beaver, R.A. Groeneveld, and W.Q. Meeker, *The nontruncated marginal of a truncated bivariate normal distribution*, *Psychometrika* 58 (1993), pp. 471–488.
- [2] J.O. Berger and T. Sellke, *Testing a point null hypothesis: The irreconcilability of p values and evidence*, *J. Am. Stat. Assoc.* 82 (1987), pp. 112–122.
- [3] H.C. Brunier and J. Whitehead, *Sample sizes for phase II clinical trials derived from Bayesian decision theory*, *Stat. Med.* 13 (1994), pp. 2493–2502.

- [4] S. Chowdhury, R.C. Tiwari, and S. Ghosh, *Bayesian approach for assessing non-inferiority in three-arm trials for risk ratio and odds ratio*, Stat. Biopharm. Res. 11 (2019), pp. 34–43.
- [5] S. Chowdhury, R.C. Tiwari, and S. Ghosh, *Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial*, Comput. Stat. Data Anal. 132 (2019), pp. 70–83.
- [6] C. Chuang-Stein, P. Stryszak, A. Dmitrienko, and W. Offen, *Challenge of multiple co-primary endpoints: A new approach*, Stat. Med. 26 (2007), pp. 1181–1192.
- [7] D.R. Cox, [*Testing precise hypotheses*]: Comment, Stat. Sci. 2 (1987), pp. 335–336.
- [8] R.B. D’Agostino Sr, J.M. Massaro, and L.M. Sullivan, *Non-inferiority trials: Design concepts and issues—the encounters of academic consultants in statistics*, Stat. Med. 22 (2003), pp. 169–186.
- [9] EMA, *Guideline on the choice of the noninferiority margin*, Doc. Ref. EMEA/CPMP/EWP/2158/99, European Medicines Agency: Pre-authorisation Evaluation of Medicines for Human Use, 2005.
- [10] FDA, *Non-inferiority clinical trials to establish effectiveness: Guidance for industry*, US Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), Silver Spring, MD, 2016.
- [11] M.A. Gamalo, R.C. Tiwari, and L.M. LaVange, *Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products*, Pharm. Stat. 13 (2014), pp. 25–40.
- [12] M.A. Gamalo, R. Wu, and R.C. Tiwari, *Bayesian approach to noninferiority trials for proportions*, J. Biopharm. Stat. 21 (2011), pp. 902–919.
- [13] M.A. Gamalo, R. Wu, and R.C. Tiwari, *Bayesian approach to non-inferiority trials for normal means*, Stat. Methods Med. Res. 25 (2016), pp. 221–240.
- [14] A. Gelman, H.S. Stern, J.B. Carlin, D.B. Dunson, A. Vehtari, and D.B. Rubin, *Bayesian Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC, 2013.
- [15] S. Ghosh, S. Ghosh, and R.C. Tiwari, *Bayesian approach for assessing non-inferiority in a three-arm trial with pre-specified margin*, Stat. Med. 35 (2016), pp. 695–708.
- [16] P. Ghosh, F. Nathoo, M. Gönen, and R.C. Tiwari, *Assessing noninferiority in a three-arm trial using the Bayesian approach*, Stat. Med. 30 (2011), pp. 1795–1808.
- [17] S. Ghosh, R.C. Tiwari, and S. Ghosh, *Bayesian approach for assessing noninferiority in a three-arm trial with binary endpoint*, Pharm. Stat. 17 (2018), pp. 342–357.
- [18] E. Hida and T. Tango, *On the three-arm non-inferiority trial including a placebo with a prespecified margin*, Stat. Med. 30 (2011), pp. 224–231.
- [19] T. Higuchi, M. Murasaki, and K. Kamijima, *Clinical evaluation of duloxetine in the treatment of major depressive disorder placebo-and paroxetine-controlled double-blind comparative study*, Japanese J. Clin. Pharmacol. 12 (2009), pp. 1613–1634.
- [20] L. Huang, J. Zalkikar, and R.C. Tiwari, *A likelihood ratio test based method for signal detection with application to FDA’s drug safety data*, J. Am. Stat. Assoc. 106 (2011), pp. 1230–1241.
- [21] H.M.J. Hung and S.J. Wang, *Multiple testing of noninferiority hypotheses in active controlled trials*, J. Biopharm. Stat. 14 (2004), pp. 327–335.
- [22] ICH Steering Committee, *ICH harmonised tripartite guideline: Statistical principles for clinical trials*, February 1998.
- [23] ICH Steering Committee, *ICH harmonised tripartite guideline: Choice of control group and related issues in clinical trials*, July 2000.
- [24] M. Kieser and T. Friede, *Planning and analysis of three-arm non-inferiority trials with binary endpoints*, Stat. Med. 26 (2007), pp. 253–273.
- [25] A. Koch and J. Röhmle, *Hypothesis testing in the ‘gold standard’ design for proving the efficacy of an experimental treatment relative to placebo and a reference*, J. Biopharm. Stat. 14 (2004), pp. 315–325.
- [26] M. Kulldorff, *A spatial scan statistic*, Commun. Stat. - Theory Methods 26 (1997), pp. 1481–1496.
- [27] K.S. Kwong, S.H. Cheung, A.J. Hayter, and M.J. Wen, *Extension of three-arm non-inferiority studies to trials with multiple new treatments*, Stat. Med. 31 (2012), pp. 2833–2843.
- [28] R.V. Lenth, *Some practical guidelines for effective sample size determination*, Am. Stat. 55 (2001), pp. 187–193.

- [29] T. Mütze, A. Munk, and T. Friede, *Design and analysis of three-arm trials with negative binomially distributed endpoints*, Stat. Med. 35 (2016), pp. 505–521.
- [30] M. Osman and S.K. Ghosh, *Novel Bayesian methods for non-inferiority tests based on relative risk and odds ratio for dichotomous data*, J. Stat. Theory Pract. 4 (2010), pp. 433–452.
- [31] I. Pigeot, J. Schäfer, J. Röhmel, and D. Hauschke, *Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo*, Stat. Med. 22 (2003), pp. 883–899.
- [32] J. Schumi and J.T. Wittes, *Through the looking glass: Understanding non-inferiority*, Trials 12 (2011), pp. 106.
- [33] R. Simon, *Bayesian design and analysis of active control clinical trials*, Biometrics 55 (1999), pp. 484–487.
- [34] M.-L. Tang and N.-S. Tang, *Tests of noninferiority via rate difference for three-arm clinical trials with placebo*, J. Biopharm. Stat. 14 (2004), pp. 337–347.
- [35] Y. Wu, Y. Li, Y. Hou, K. Li, and X. Zhou, *Study duration for three-arm non-inferiority survival trials designed for accrual by cohorts*, Stat. Methods Med. Res. 27 (2018), pp. 507–520.
- [36] B.G. Zaslavsky, *Bayesian hypothesis testing in two-arm trials with dichotomous outcomes*, Biometrics 69 (2013), pp. 157–163.