Biostatistics (2022) **23**, 1, *pp.* 136–156
doi: 10.1093/biostatistics/kxaa014
Advance Access publication on May 8, 2020

New approaches for testing non-inferiority for three-arm trials with Poisson distributed outcomes

SAMIRAN GHOSH*

Family Medicine & Public Health Sciences and Center of Molecular Medicine and Genetics, Wayne State University

sghos@med.wayne.edu

ERINA PAUL, SHRABANTI CHOWDHURY

Center of Molecular Medicine and Genetics, Wayne State University

RAM C. TIWARI

Division of Biostatistics, Center for Devices and Radiological Health, Office Surveillance and Biometrics, FDA, USA

Summary

With the availability of limited resources, innovation for improved statistical method for the design and analysis of randomized controlled trials (RCTs) is of paramount importance for newer and better treatment discovery for any therapeutic area. Although clinical efficacy is almost always the primary evaluating criteria to measure any beneficial effect of a treatment, there are several important other factors (e.g., side effects, cost burden, less debilitating, less intensive, etc.), which can permit some less efficacious treatment options favorable to a subgroup of patients. This leads to non-inferiority (NI) testing. The objective of NI trial is to show that an experimental treatment is not worse than an active reference treatment by more than a pre-specified margin. Traditional NI trials do not include a placebo arm for ethical reason; however, this necessitates stringent and often unverifiable assumptions. On the other hand, three-arm NI trials consisting of placebo, reference, and experimental treatment, can simultaneously test the superiority of the reference over placebo and NI of experimental treatment over the reference. In this article, we proposed both novel Frequentist and Bayesian procedures for testing NI in the three-arm trial with Poisson distributed count outcome. RCTs with count data as the primary outcome are quite common in various disease areas such as lesion count in cancer trials, relapses in multiple sclerosis, dermatology, neurology, cardiovascular research, adverse event count, etc. We first propose an improved Frequentist approach, which is then followed by it's Bayesian version. Bayesian methods have natural advantage in any active-control trials, including NI trial when substantial historical information is available for placebo and established reference treatment. In addition, we discuss sample size calculation and draw an interesting connection between the two paradigms.

*To whom correspondence should be addressed.

Keywords: Assay sensitivity; Count outcome; Fraction margin; Markov chain Monte Carlo; Non-inferiority margin; Sample size.

1. Introduction

The randomized controlled trials (RCTs) are traditionally the gold standard for judging the benefits of treatment for a disease under idealistic condition. According to the requirements set by the regulatory agencies (e.g., FDA, 2016; EMA, 2005), drug developers need to demonstrate evidence of efficacy and safety of an intervention through well-designed randomized placebo-controlled trial/s (RCTs). However, in the presence of an established treatment regime conducting placebo-controlled trial is unethical. Instead, experimental treatment is compared with an active control or reference treatment/drug/intervention. Most of such active comparator trials are superiority trials. However, when superiority of the experimental intervention is not clear, yet it poses certain attractive properties, one may resort to a non-inferiority (NI) trial. The objective of efficacy seeking NI trial is to establish that an experimental treatment is non-inferior to an active comparator within a small, pre-specified margin (or NI margin), and at the same time retains a substantial portion of the active controls effect in the current trial (D'Agostino Sr *and others*, 2003). The choice of this NI margin has been an area of major concern and some broad outlines have been provided by regulatory agencies (ICH Steering Committee, 1998, 2000; FDA, 2016; EMA, 2005). The following references (FDA, 2016; Hung and Wang, 2004; Schumi and Wittes, 2011) provide a detailed discussion on NI margin that must be constructed based on the past performance of active control. Sometime an intervention after passing superiority test for efficacy, additionally also tested for NI for safety, however that is not the main focus of this article (Tsong and Zhang, 2007; Lu *and others*, 2018).

For the ethical reason discussed above, NI trials mostly lack a placebo arm. Hence, such two-arm NI trials make some important assumptions regarding assay sensitivity and constancy. The validity of the resulting inference depends heavily on external validation. Assay sensitivity (AS) of a clinical trial is defined in its ability to distinguish an effective treatment from a less effective/ineffective one as defined by the ICH guideline (ICH Steering Committee, 1998, 2000). Moreover, it is also required that the effect size of the active control over placebo in the historical placebo-controlled trial holds in the current NI trial (i.e., constancy), otherwise efficacy of the experimental treatment over putative placebo cannot be shown. Kieser and Stucke (2016) mentioned several other factors that often plague two-arm NI trials. To alleviate some of these issues and if ethically acceptable and practically feasible, it is recommended by EMA (2005) to include a placebo arm in the current trial, resulting in a three-arm trial often known as "gold-standard" design. In Frequentist setup, Pigeot *and others* (2003) first proposed an approach where NI margin is adaptively formulated as the pre-specified negative fraction of the unknown effect size of the reference treatment over placebo in the current three-arm trial. Such formulation of the NI margin, called the fraction margin approach, is also known as "effect retention." This approach was extended by Kieser and Friede (2007) and Chowdhury *and others* (2018b) for the binary outcome, Mielke *and others* (2008) for censored exponentially distributed outcome, Ghosh *and others* (2017) for non-normal continuous outcome, Stucke and Kieser (2013) and Mütze *and others* (2016) for the count outcome to name a few.

1.1. *Background and motivating example*

A three-arm trial via fraction margin approach is a two-step process. In the first step, one must show that superiority of the reference over placebo, i.e., the AS condition holds. If this is successful, one proceeds to test NI of the experimental treatment. Due to the hierarchical structure of the multiple testing problem (AS and NI) though one do not need to adjust for type-I error, however, the pretest for assay sensitivity may lead to a reduction in power when testing for NI. This fact was demonstrated clearly in

Kieser and Friede (2007). A possible alternative is proposed by Hida and Tango (2013), where they suggested joint testing of AS and NI. This approach requires joint rejection of AS and NI null hypotheses to claim NI (with AS) leading to Intersection–Union testing (IUT). Though logical and can be tightly controlled (for type-I error) the IUT under Frequentist setup may lead to biased test (Berger, 1997). This is discussed in detail by Chuang-Stein *and others* (2007) in the superiority trial context. In this article based on Frequentist approach we have developed first, a more powerful test based on conditional principle for fraction margin based NI testing.

Since NI trials involve active control that has been well established in the past, the availability of historical information is almost guaranteed. Bayesian paradigm provides a natural path to combine information from various historical trials as prior and then to combine them with the current trial. This has the possibility of reducing sample size and cost burden. Bayesian approaches have been predominantly used in clinical trials, particularly in the NI trials since long past (e.g., see Simon, 1999; Ghosh *and others*, 2011, 2016; Gamalo *and others*, 2016, 2014). Bayesian approach for NI trial for two-arm can be found in Gamalo *and others* (2011), Chowdhury *and others* (2018a) and for three-arm can be found in Ghosh *and others* (2011), Ghosh *and others* (2016). To the best of our knowledge no literature exists for NI testing for count outcomes under Bayesian paradigm. Even from the Frequentist approach the papers are only handful. Albeit as mentioned in Stucke and Kieser (2013), the existence of count type outcome is not very uncommon. Examples include relapse-remitting count in multiple sclerosis (MS) trial (Friede and Schmidli, 2010; Noseworthy, 2003; Sormani *and others*, 2001); lesion count in cancer trial (McIntosh, 2001; Xie and Aickin, 1997); number of attacks in migraine trial (Silcocks *and others*, 2010); number of manic episodes in bipolar trial (Soeiro-de Souza *and others*, 2013); post-discharge adverse events (Tsilimingras *and others*, 2015) to name a few. Along with a novel and more powerful conditional Frequentist approach in this article, we also propose both exact and an approximate Bayesian approach for testing the NI hypothesis for count outcomes in three-arm trial. Effective sample size is calculated using all procedures to make a comparative analysis.

Our motivation for developing both Frequentist and Bayesian tests comes from a clinical trial dataset (Calabrese *and others*, 2012) on MS where primary outcome is counts of cortical inflammatory lesions (CLs). Recent studies have shown that CLs may have a major role in determining disability in patients with MS. Some of the critical symptoms of CLs include epilepsy, memory loss in relapsing-remitting multiple sclerosis (RRMS) patients, cortical atrophy in primary progressive MS, etc. The original dataset comprises of patients who were untreated or treated with different disease modifying drugs (DMDs). The objective of the study was to assess the effects of the DMDs compared with no therapy (placebo). The patients with RRMS were enrolled in a 2-year prospective, randomized, single-centric study. There were 50 patients who did not receive any therapy for the 2 years of the follow-up period, while the remaining patients who were part of the clinical study, received the respective DMDs. These patients were evaluated at baseline, at 12 and 24 months after the treatment. Table 1 shows the frequencies of the new CLs developed by these patients with MS after 1 year and 2 years of their treatment with the DMDs. As evident the primary response is count type, with good fit to Poisson distribution via Index of Dispersion test (Gbur, 1981) for all arms. We have applied our proposed test procedures to determine NI of the DMD glatiramer acetate (GA) (E) over subcutaneous interferon (IFN) beta-1a (R) for both 1- and 2-year data. The detailed illustrations are in Section 6.

The rest of the article is organized as follows. In Section 2, we give the NI hypothesis and existing Frequentist methods for testing the count data. We also introduce our proposed more powerful conditional testing in this section. In Section 3, we propose a novel Bayesian methodology for the same. We consider both conjugate and non-conjugate priors incorporating the condition of AS. In Section 4, the power and sample size calculations are discussed in detail for Frequentist approach, Bayesian normal approximation, and exact Bayesian methods. Section 5 presents the simulation results along with the power curves. Finally in Section 6, we apply our proposed Bayesian methodology for NI testing on this clinical trial dataset.

Table 1. *Frequencies of new MRI CLs over 1 and 2 years. N denotes sample size.*

1 year				2 years			
Arm	Counts	Frequencies	N, Mean, Var	Arm	Counts	Frequencies	N, Mean, Var
P	0	13	50, 1.6, 1.2	P	0	9	50, 3.0, 2.6
	≥ 1	37			≥ 1	41	
R	0	34	46, 0.4, 0.49	R	0	22	46, 0.8, 0.6
	≥ 1	12			≥ 1	24	
E	0	24	48, 0.8, 1.0	E	0	18	48, 1.3 , 1.21
	≥ 1	24			≥ 1	30	

The article concludes with discussion and future direction in Section 7. All proofs and additional simulation results are provided in Supplementary Appendix available at *Biostatistics* online.

2. Frequentist approach for NI testing

We adopt the notation used in Stucke and Kieser (2013) to illustrate the fraction margin approach for three-arm NI trial. We denote the experimental treatment by E, the reference by R, and the placebo by P. The sample size corresponding to the three arms are denoted by n_E, n_R, and n_P, respectively, which are not necessarily equal. Let X_{kE}, X_{kR}, and X_{kP}, $k = 1, \ldots, n_l$ denote the primary count type independent random variable corresponding to the kth individual in the respective treatment arms. The X_{kl} is distributed as Poisson $(\lambda_l t_l)$ with $\lambda_l (> 0)$ represents the rate parameter and t_l denotes the fixed follow-up times for $l \in \{E, R, P\}$. Hence, $\lambda_l t_l$ denotes the expected number of counts per-patient in the lth group. We assume that these random variables are mutually independent. Without loss of generality, we assume that higher the values of the Poisson rates λ_l, greater is the treatment benefits. Again, we denote the total number of counts for all n_l patients in the lth treatment arm by $X_l = \sum_{k=1}^{n_l} X_{kl}$ which is distributed as Poisson $(\lambda_l t_l n_l)$, $l \in \{E, R, P\}$. Later on for our analysis, we will consider homogeneous Poisson distributions for the treatment arms; that is, we take $t_l = 1$ for $l \in \{E, R, P\}$. Modeling non-homogeneous Poisson distribution is a possibility which we did not explore in this article. The usual NI hypothesis for a two-arm trial (without placebo) is

$$H_0 : \lambda_E - \lambda_R \leq \delta \text{ vs. } H_1 : \lambda_E - \lambda_R > \delta, \tag{2.1}$$

where $\delta < 0$ denotes the pre-specified amount of NI margin. In the current three-arm trial, the construction of δ via fraction margin approach (Pigeot *and others*, 2003) can be mathematically expressed as $\delta = f(\lambda_R - \lambda_P)$, where $-1 < f < 0$ assuming the condition of AS $(\lambda_R > \lambda_P)$. Hence, the hypothesis in (2.1) can be rewritten using the expression for δ as follows: $H_0 : \lambda_E - \lambda_R \leq f\ (\lambda_R - \lambda_P)$ vs. $H_1 : \lambda_E - \lambda_R > f(\lambda_R - \lambda_P)$. Now, putting $\theta = 1 + f$, the above hypothesis becomes

$$H_0 : \frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} \leq \theta \text{ vs. } H_1 : \frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} > \theta, \tag{2.2}$$

where θ is the pre-specified fraction of the effect of the reference drug relative to the placebo. Clearly, rejection of the null hypothesis ensures that the experimental treatment retains a portion of the unknown effect of the reference over placebo under the fraction margin approach (Kieser and Stucke, 2016) and would support NI of the experimental drug over the active control. Different choices of $\theta (\in [0, 1])$ have been proposed in Pigeot *and others* (2003). Particularly, for NI testing of the experimental drug, θ is

allowed to vary in the interval [0.5, 1), indicating at least 50% or more effect retention. The hypothesis in (2.2) can be expressed in the following form which is used later for deriving the statistical test procedures

$$H_0 : \lambda_E - \theta\lambda_R - (1-\theta)\,\lambda_P \leq 0 \text{ vs. } H_1 : \lambda_E - \theta\lambda_R - (1-\theta)\,\lambda_P > 0. \tag{2.3}$$

2.1. *Existing Frequentist approaches*

Mütze *and others* (2016) developed NI hypothesis testing where count outcome assumed to follow a negative binomial distribution. They constructed the test statistic for testing NI hypothesis by considering the maximum likelihood (ML) estimate of the linear contrast in H_0 (in 2.3) given by, $T = \hat{\lambda}_E - \theta\hat{\lambda}_R - (1-\theta)\,\hat{\lambda}_P$, where $\hat{\lambda}_l = X_l/n_l t_l$ is the maximum likelihood estimate (MLE) of λ_l, $l \in \{E, R, P\}$. The variance of the test statistic is given as $\text{Var}\,(T) = \lambda_E/n_E t_E + \theta^2\lambda_R/n_R t_R + (1-\theta)^2\,\lambda_P/n_P t_P$. Both ML and restricted maximum likelihood (RML) estimation techniques can be adopted to estimate $\text{Var}(T)$. The RML estimator can be obtained subject to the constraint $\lambda_E - \theta\lambda_R - (1-\theta)\,\lambda_P = 0$. Mütze *and others* (2016) also derived asymptotic sample size formulae along with optimal sample size allocation considering both balanced and unbalanced designs albeit in the Frequentist set up. For a two-arm trial, Stucke and Kieser (2013) derived the statistical test procedure using RML estimator and obtained approximate sample size formulae under the Frequentist set up. Note, this approach of NI testing is valid provided the AS null hypothesis has already been rejected first. Hence, the step for NI testing is always a conditional test. However, this AS conditioning is not used in any of the existing approach of Frequentist test. We have shown mathematically that if the pretested AS condition is used properly in the second step (i.e., in NI testing), this could lead to a more powerful test with considerable savings in sample size.

2.2. *Proposed Frequentist approach*

As mentioned in Section 1.1, it is often argued (Pigeot *and others*, 2003; Koch and Röhmel, 2004; Ghosh *and others*, 2011; Wu *and others*, 2018) that if active control has not lost all of its effect over placebo then the statistical power to perform joint testing (NI and AS) will be very similar to NI testing only. This may not be true in all situation as shown in Kieser and Friede (2007), except when power of the pretest is close to unity. Nevertheless, NI testing only happens provided the AS condition ($\lambda_R > \lambda_P$) holds. However, this pretested AS condition has not been used further, though NI and AS test statistics are related. We introduce here a new conditional approach for NI hypothesis testing by incorporating the pretested AS condition ($\lambda_R > \lambda_P$) directly. We have shown that this approach will perform better or as good as the existing approach. For finding the MLE, we truncate the parameter space of $(\lambda_E, \lambda_R, \lambda_P)$ such that it belongs to $\{\lambda_E, \lambda_R, \lambda_P : \lambda_E, \lambda_R, \lambda_P \in [0, \infty), \lambda_R > \lambda_P\}$. One may develop a likelihood ratio test based on the statistic $T = \hat{\lambda}_E - \theta\hat{\lambda}_R - (1-\theta)\hat{\lambda}_P = (\hat{\lambda}_E - \hat{\lambda}_P) - \theta(\hat{\lambda}_R - \hat{\lambda}_P) = U - \theta V$ under null hypothesis subject to the imposed condition, $\hat{\lambda}_R > \hat{\lambda}_P$ via Wald-type test. Following Mütze *and others* (2016) argument, one can improve the convergence of Wald-type test via the RML which requires solving under H_0, $\left(\hat{\lambda}_{E,\text{RML}}, \hat{\lambda}_{R,\text{RML}}, \hat{\lambda}_{P,\text{RML}}\right) = \arg\max_{\lambda_E - \theta\lambda_R - (1-\theta)\lambda_P \leq 0, \lambda_R > \lambda_P} \log l(\lambda_E, \lambda_R, \lambda_P)$, where $\log l(\lambda_E, \lambda_R, \lambda_P)$ is the log-likelihood of $(\lambda_E, \lambda_R, \lambda_P)$ to estimate T by T_{RML}. This optimization problem can be solved only numerically as no closed form expression is possible. To reduce computational burden one practical strategy that is often recommended is to work with unrestricted MLE which is $T_{ML} = \hat{\lambda}_{E,\text{ML}} - \theta\hat{\lambda}_{R,\text{ML}} - (1-\theta)\hat{\lambda}_{P,\text{ML}}$, however, only considering the part restricted by $\hat{\lambda}_{R,\text{ML}} > \hat{\lambda}_{P,\text{ML}}$, which is

$$T_{\text{RML}} \simeq T_{\text{ML}} * I[\hat{\lambda}_{R,\text{ML}} > \hat{\lambda}_{P,\text{ML}}]. \tag{2.4}$$

This strategy is proved to be quite useful in many practical applications (Huang *and others*, 2011; Kulldorff, 1997). Since working with product of random variables in (2.4) is little cumbersome, one can further show that $f(T_{\mathrm{RML}}) \simeq f(T_{\mathrm{ML}}|\hat{\lambda}_{R,\mathrm{ML}} > \hat{\lambda}_{P,\mathrm{ML}}) \times Pr[\hat{\lambda}_{R,\mathrm{ML}} > \hat{\lambda}_{P,\mathrm{ML}}]$. It is easy to prove that $Pr[\hat{\lambda}_{R,\mathrm{ML}} > \hat{\lambda}_{P,\mathrm{ML}}]$ is a constant value which can be absorbed as a proportionality constant. Hence, for all practical purposes, one can consider the distribution of the test statistic, $f(T_{\mathrm{ML}}|\hat{\lambda}_R > \hat{\lambda}_P) \propto f(\hat{\lambda}_{E,\mathrm{ML}} - \theta\hat{\lambda}_{R,\mathrm{ML}} - (1-\theta)\hat{\lambda}_{P,\mathrm{ML}}|\hat{\lambda}_{R,\mathrm{ML}} > \hat{\lambda}_{P,\mathrm{ML}})$. For notational simplicity from now onwards, we denote the ML estimate $\hat{\lambda}_{l,\mathrm{ML}}$ by $\hat{\lambda}_l$, $l \in \{E,R,P\}$. This leads to the modified test statistic for NI testing: $W = (\hat{\lambda}_E - \theta\hat{\lambda}_R - (1-\theta)\hat{\lambda}_P|\hat{\lambda}_R > \hat{\lambda}_P) = (U - \theta V|V > 0)$. Under the asymptotic normality of W, we have $(W-\mu_w)/\sigma_w \sim AN(0,1)$, where μ_w and σ_w^2 are the mean and variance of W, respectively.

LEMMA 2.2.1 Under conditional normal approximation, the mean μ_w and variance σ_w^2 of $W = \hat{\lambda}_E - \theta\hat{\lambda}_R - (1-\theta)\hat{\lambda}_P|\hat{\lambda}_R > \hat{\lambda}_P$ are given by $\mu_w = \mu_U + \sigma_U\frac{\rho}{c}\phi(d) - \theta\left(\mu_V + \sigma_V\frac{1}{c}\phi(d)\right), \sigma_w^2 = \sigma_U^2\left[1 + \frac{\rho^2}{c}d\phi(d) - \left(\frac{\rho}{c}\phi(d)\right)^2\right] + \theta^2\sigma_V^2\left[1 - \frac{\phi(d)}{c}\left(\frac{\phi(d)}{c} - d\right)\right] - 2\theta\left[\sigma_U\sigma_V\frac{\rho}{c}(c + d\phi(d)) + \sigma_U\mu_V\frac{\rho}{c}\phi(d) + \sigma_V\mu_U\frac{1}{c}\phi(d) + \mu_U\mu_V - \left(\mu_U + \sigma_U\frac{\rho}{c}\phi(d)\right)\left(\mu_V + \sigma_V\frac{1}{c}\phi(d)\right)\right]$, where $\mu_U = \lambda_E - \lambda_P$, $\mu_V = \lambda_R - \lambda_P$, $\sigma_l^2 = \frac{\lambda_l}{n_l}$ for $l \in \{E,R,P\}$, $d = -\frac{\mu_V}{\sigma_V}, c = 1 - \Phi(d)$, $\sigma_U^2 = \sigma_E^2 + \sigma_P^2$, $\sigma_V^2 = \sigma_R^2 + \sigma_P^2$, and $\rho = \frac{Var(\hat{\lambda}_P)}{\sqrt{Var(U)Var(V)}} = \frac{\sigma_P^2}{\sqrt{\sigma_U^2\sigma_V^2}}$.
Proof: See Supplementary Appendix A available at *Biostatistics* online.

Now under H_0, let us denote λ_E by $\lambda_E^{\mathrm{null}}$ and under H_1 denote λ_E by λ_E^{alt} as point alternative. Since $\lambda_E^{\mathrm{null}}$ satisfies $\lambda_E^{\mathrm{null}} - \theta\lambda_R - (1-\theta)\lambda_P = 0$, the expression of $\lambda_E^{\mathrm{null}}$ can be obtained via $\lambda_E^{\mathrm{null}} = \lambda_P + \theta(\lambda_R - \lambda_P)$. Under H_1, λ_E^{alt} satisfies $\lambda_E^{\mathrm{alt}} - \theta\lambda_R - (1-\theta)\lambda_P > 0 \Rightarrow (\lambda_E^{\mathrm{alt}} - \lambda_P) > \theta(\lambda_R - \lambda_P)$. Since λ_E is involved in the expression of the mean and variance of W, we denote $E(W)$ and Var (W) under H_0 by μ_w^{null} and $\sigma_w^{2\mathrm{null}}$ and under H_1, by μ_w^{alt} and $\sigma_w^{2\mathrm{alt}}$, respectively. Thus, we have $(W - \mu_w^{\mathrm{null}})/\sigma_w^{\mathrm{null}} \sim AN(0,1)$ under H_0 and $(W - \mu_w^{\mathrm{alt}})/\sigma_w^{\mathrm{alt}} \sim AN(0,1)$ under H_1. In Frequentist approach, the critical region of the test is given by $W > k_\alpha^*$, where k_α^* is obtained by assuming a test of size α: $P_{H_0}\left(W > k_\alpha^*\right) = \alpha \Rightarrow k_\alpha^* = \mu_w^{\mathrm{null}} + z_{1-\alpha}\sigma_w^{\mathrm{null}}$, where $z_{1-\alpha}$ is the $100(1-\alpha)\%$ percentile point of the $N(0,1)$ distribution. Traditionally, the value of α is chosen to be 0.025 (other choices are possible too). The expression of the power of the test is given by $P_{H_1}\left(W > k_\alpha^*\right) = 1 - \Phi((k_\alpha^* - \mu_w^{\mathrm{alt}})/\sigma_w^{\mathrm{alt}})$.

LEMMA 2.2.2 At a fixed α and sample size $N(= n_E + n_R + n_P)$, proposed conditional test statistic (W) has more power than the existing marginal test statistic (T) for testing NI hypothesis in (2.3).
Proof: See Supplementary Appendix B available at *Biostatistics* online.

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 Section 4.5, this power difference is substantial when the gap between λ_R and λ_P is small. In Supplementary section available at *Biostatistics* online, we have provided additional simulation result to demonstrate this fact. It should be also noted that this lemma is generalizable for continuous as well as for binary outcome with slightly different algebra, indicating the fact that our proposed conditional test should be *de facto* standard for Pigeot's fraction margin approach irrespective of the outcome types.

3. BAYESIAN APPROACHES FOR NI TESTING

As indicated in Section 1, availability of considerable prior information is almost guaranteed in any active control trial and NI RCT is not an exception. Albeit, the usage of these historical information via

the Frequentist approaches is rather limited. Bayesian approach provides a natural path to leverage this historical data which may result in substantial effective sample size gain. However, to the best of our knowledge no Bayesian methodology paper exists for any three-arm trial with count type endpoints. In this section, we discuss an exact Bayesian and an approximation-based Bayesian method for NI testing involving Poisson rates. Note, we did not develop here Bayesian approach for existing Frequentist approach of Mütze *and others* (2016). However, as we proposed a more powerful Frequentist test in Section 2.2, our Bayesian development closely follows that procedure.

As stated earlier, the NI margin is constructed as the negative fraction of the unknown difference of the count rate of responses in the reference and the placebo arm. We consider $\theta \geq 0.5$ to test for the NI of E relative to R with two different prior scenarios, including the conjugate prior where the AS condition $(\lambda_R > \lambda_P)$ is directly incorporated. This restriction reflects that the NI study is being carried out under the similar condition as that of the former studies in which the efficacy of the active control was proved, and it still retains its effect over placebo. This is a very realistic assumption because if the current trial is similar to the historical trial then the effect of reference drug over placebo should be constant in both the current and the historical trial (constancy assumption). In the following section, we discuss both the conjugate and non-conjugate prior settings. In case there is no available prior information, flat non-informative prior is assigned to λ_l which includes Jeffreys prior and other priors with adjusted parameters yielding large variance.

3.1. *Exact Bayesian approach*

3.1.1. *Conjugate Gamma prior* In the conjugate prior setting, we use a Gamma distribution as the prior for the Poisson rate in each arm of the trial; that is, we assume $\lambda_l \sim$ Gamma $(\alpha_l, \beta_l), l \in \{E, R, P\}$, where we assume α_l, β_l to be fixed hyper-parameters. After incorporating the assumption of AS $(\lambda_R > \lambda_P)$, the joint prior distribution of the Poisson rates in the three-arms becomes $f(\boldsymbol{\lambda}) = I(\lambda_R > \lambda_P) \prod_{l\in\{E,R,P\}} f(\lambda_l|\alpha_l, \beta_l)$, where $f(\lambda_l|\alpha_l, \beta_l)$ is the density of Gamma(α_l, β_l) distribution given as

$$f(\lambda_l|\alpha_l, \beta_l) \propto \lambda_l^{\alpha_l - 1} \exp\{-\lambda_l \beta_l\},\ \lambda_l > 0.$$

Since the number of counts, X_l, in each arm, follows a Poisson distribution with parameter $n_l\lambda_l t_l$, $l \in \{E, R, P\}$, the posterior distribution for $\lambda_l|X_l$ is Gamma $(\alpha_l + X_l, \beta_l + n_l t_l)$ satisfying AS condition is given by

$$f(\lambda_E, \lambda_R, \lambda_P|\text{Data}, \alpha_l, \beta_l) \propto I(\lambda_R > \lambda_P) \prod_{l\in\{E,R,P\}} \lambda_l^{\alpha_l + x_l - 1} \exp\{-\lambda_l(\beta_l + n_l t_l)\},\ \lambda_l > 0,\ l \in \{E, R, P\}.$$

The Markov chain Monte Carlo (MCMC) samples can be easily generated from this joint posterior distribution. The hyper-parameters α_l and β_l, $l \in \{E, R, P\}$ can be chosen depending on how much prior information is available. In the absence of prior information from historical placebo-controlled trial, they are chosen to be vague. The mean (μ), mode (μ^0), and variance (σ^2) of Gamma (α, β) are given as $\mu = \alpha/\beta$, $\mu^0 = (\alpha - 1)/\beta$, and $\sigma^2 = \alpha/\beta^2$. For the informative priors, the variance is made smaller making priors to be more specific.

3.1.2. *Non-conjugate prior* In this case, the prior distributions are so assigned to the parameters λ_E, λ_R, and λ_P that satisfy the restriction $0 < \lambda_P < \lambda_R$. We give joint prior on (λ_R, λ_P) by putting a Gamma prior on λ_R and a Beta prior on λ_P/λ_R which ensures $\lambda_R > \lambda_P$. We put unrestricted prior Gamma (α_E, β_E) on λ_E. The following transformation is made from (λ_R, λ_P) to (u_1, u_2): $u_1 = \lambda_P/\lambda_R \sim$ Beta (a, b), $u_2 = \lambda_R \sim$ Gamma(p, r). So, we have $0 < u_1 < 1$ (satisfies the AS

condition $(\lambda_R > \lambda_P)$) and $u_2 > 0$. The joint distribution of (u_1, u_2) is given by $f(u_1, u_2) = \text{Beta}(a, b) \times \text{Gamma}(p, r) \propto u_1^{a-1}(1-u_1)^{b-1}\exp\{-ru_2\}u_2^{p-1}$, which gives the joint distribution of (λ_R, λ_P) as $f(\lambda_R, \lambda_P) \propto \frac{1}{\lambda_R}(\lambda_P/\lambda_R)^{a-1}(1-\lambda_P/\lambda_R)^{b-1}\exp\{-r\lambda_R\}\lambda_R^{p-1}$, $0 < \lambda_P < \lambda_R$. The joint prior distribution of $(\lambda_E, \lambda_R, \lambda_P)$ can be obtained by multiplying $f(\lambda_R, \lambda_P)$ with $f(\lambda_E) \equiv \text{Gamma}(\alpha_E, \beta_E)$, which is given as

$$f(\lambda_E, \lambda_R, \lambda_P) \propto \lambda_E^{\alpha_E - 1}\exp\{-\beta_E\lambda_E\}\lambda_P^{a-1}(\lambda_R - \lambda_P)^{b-1}\exp\{-r\lambda_R\}\lambda_R^{p-a-b},$$

$0 < \lambda_E < \infty,\ 0 < \lambda_P < \lambda_R < \infty$. The joint posterior distribution $f(\lambda_E,\ \lambda_R,\ \lambda_P|\text{Data})$ is proportional to the multiplication of the joint likelihood and the joint prior as

$$f(\lambda_E,\ \lambda_R,\ \lambda_P|\text{Data}) \propto \text{Gamma}(\lambda_E|\alpha_E + X_E, \beta_E + n_E t_E) \times \exp\{-n_P\lambda_P t_P\}\lambda_P^{a+x_P-1} \times$$
$$\exp\{-\lambda_R(r + n_R t_R)\}\lambda_R^{x_R+p-a-b}(\lambda_R - \lambda_P)^{b-1},\ 0 < \lambda_P < \lambda_R < \infty,\ 0 < \lambda_E < \infty.$$

The posterior is not in the closed form and a Metropolis–Hastings acceptance–rejection sampling is required with a proposal density to generate posterior samples (Gelman *and others*, 2014). A convenient proposal density could be Gamma distribution with appropriately chosen priors. In our simulation, we use "rjags" (R-package; Plummer *and others*, 2016) to generate the samplers from the posterior density.

Remark 1: Following Pigeot *and others* (2003) and Ghosh *and others* (2011), we continue to assume that AS condition $(\lambda_R > \lambda_P)$ is tested in Step 1, before proceeding to test for NI. As a result truncated priors are chosen in Step 2, i.e., at NI testing. This assumption explicitly reflects the fact that active control still retains some of its effect over placebo. In a situation where this assumption is questionable, it is not advisable to carry out a three-arm NI trial, rather a superiority trial of the new treatment over placebo is more realistic.

3.1.3. *Test procedure* For NI testing, the value of θ is so chosen that is clinically acceptable to claim that an experimental drug is non-inferior to an active control. Usually, θ is chosen in the range $[0.5, 1)$ and NI of the test drug relative to the reference is claimed if the posterior probability of the alternative hypothesis given in (2.2) exceeds a pre-specified cutoff p^*. Following Ghosh *and others* (2011) (Section 3.3) thus, the Bayesian decision rule to claim NI of the test drug over the reference is given as

$$P\left(H_1 : \frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} > \theta|\lambda_R > \lambda_P, \text{Data}\right) > p^*. \tag{3.1}$$

The value of p^* is usually chosen to be 0.975 or 0.95. The above probability can be calculated empirically by generating the samplers from the posterior distribution of $\lambda_l|X_l, l \in \{E, R, P\}$. The estimated probability is given by

$$\hat{P}\left(H_1 : \frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} > \theta|\lambda_R > \lambda_P, \text{Data}\right) \approx \frac{1}{M}\sum_{m=1}^{M} I\left(\frac{\lambda_E^m - \lambda_P^m}{\lambda_R^m - \lambda_P^m} > \theta|\lambda_R^m > \lambda_P^m, \text{Data}\right), \tag{3.2}$$

where $\lambda_E^m,\ \lambda_R^m$, and λ_P^m denote the mth sample value drawn from the posterior distributions.

3.2. *Approximate Bayesian approach*

Note that in the exact Bayesian approach, the posterior sample generation is necessary to carry out the Bayesian inference which is often computationally intensive. Here, we propose an approximation-based

Bayesian approach for NI testing incorporating the AS condition that gives closed form of the posterior probability and hence, saves the computation time of the MCMC sample generation from posterior distribution. Consider the Gamma prior for the rate λ_l in each arm, that is, $\lambda_l \sim$ Gamma (α_l, β_l) and assume that the responses are distributed as Poisson; that is, $X_l \sim$ Poisson $(n_l\lambda_l t_l)$ for $l \in \{E, R, P\}$. The Frequentist test statistic for testing the hypothesis in equation (2.3) is given by $T = X_E/n_E t_E - \theta X_R/n_R t_R - (1-\theta)X_P/n_P t_P$. For the sake of simplicity, we take $t_l = 1$, $l \in \{E, R, P\}$. Under asymptotic normality assumption, we have $T|\mu_T \sim AN\left(\mu_T, \sigma_T^2\right)$, where $\mu_T = \lambda_E - \theta\lambda_R - (1-\theta)\lambda_P = (\lambda_E - \lambda_P) - \theta(\lambda_R - \lambda_P)$ and $\sigma_T^2 = \lambda_E/n_E + \theta^2\lambda_R/n_R + (1-\theta)^2\lambda_P/n_P$. Putting Normal prior on μ_T, we have $\mu_T \sim AN\left(\mu^*, \sigma^{*2}\right)$, where $\mu^* = E(\mu_T) = \mu_E - \theta\mu_R - (1-\theta)\mu_P$ and $\sigma^{*2} = \sigma_E^2 + \theta^2\sigma_R^2 + (1-\theta)^2\sigma_P^2$, μ_l and σ_l^2, $l \in \{E, R, P\}$ are the respective mean and variance of the Gamma prior for the Poisson rates. Next, we bring in the condition of AS $(\lambda_R > \lambda_P)$. So instead of taking prior on μ_T, we take prior on $\nu_T \equiv (\mu_T|\lambda_R > \lambda_P)$. Assume that $\nu_T \sim AN\left(\mu_\nu^*, \sigma_\nu^{*2}\right)$ and the posterior, $\nu_T|\text{Data} \sim AN\left(\widetilde{\mu}_T\widetilde{\sigma}_T^2, \widetilde{\sigma}_T^2\right)$. We refer to Arnold and Beaver (1993) for the detailed derivation of μ_ν^*, σ_ν^{*2}, $\widetilde{\mu}_T$, and $\widetilde{\sigma}_T^2$ (see also Supplementary Appendix C available at *Biostatistics* online). The Bayesian decision rule for the experimental treatment to be non-inferior to the active comparator is given by Gamalo *and others* (2014): $P(\nu_T \geq 0|\text{Data}) > p^*$, where p^* is the pre-specified clinically reasonable constant.

4. POWER AND SAMPLE SIZE DETERMINATION

We address the problem of calculating the sample size for the assessment of NI to attain a desired power using three approaches described in Sections 2 and 3. The normal approximation-based approaches do not require any simulation for the estimation of the power function as it can be expressed in a closed form (as presented in the following subsections). However, the exact Bayesian approach requires the simulation technique to obtain the empirical power which is then set to a desired level to calculate the corresponding sample size. In our sample size calculation, we consider $t_l = 1$, $l \in \{E, R, P\}$. We want to determine the sample size n_l, $l \in \{E, R, P\}$ setting the power at $(1-\beta)$, β is the pre-specified type-II error. We assume $n_E = n$, $n_R = r_1 n$, and $n_P = r_2 n$, where r_1 and r_2 determine the allocation of the sample sizes in the reference and the placebo arms, respectively, relative to the experimental arm. The total sample size in that case would be $N = n(1 + r_1 + r_2)$. In the following sub-section, we discuss the power and sample size calculation under the proposed Frequentist, approximation-based Bayesian, and exact Bayesian approaches.

4.1. *Frequentist approach*

To obtain the empirical power function of the NI testing in (2.3) using the test procedure described in Section 2.2, we fix λ_R, λ_P, and θ and vary λ_E such that the ratio $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) \in [0.5,\ 1.5]$. The ratio $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P)$ is so chosen that for NI testing under H_0 it equals $\theta \in [0.5, 1)$ and exceeds θ under H_1. Under the null hypothesis, denote λ_E by λ_E^{null} which satisfies $(\lambda_E^{\text{null}} - \lambda_P) = \theta(\lambda_R - \lambda_P)$ and under H_1, denote λ_E by λ_E^{alt} which satisfies $(\lambda_E^{\text{alt}} - \lambda_P) > \theta(\lambda_R - \lambda_P)$. The empirical type-I error is obtained for $\lambda_E = \lambda_E^{\text{null}}$ and the power is obtained for $\lambda_E = \lambda_E^{\text{alt}}$. The sample size is calculated from the following equation, to achieve a power of at least $100(1-\beta)\%$

$$P_{H_1}\left(W > k_\alpha^*\right) \geq 1 - \beta \Rightarrow \Phi\left(\frac{k_\alpha^* - \mu_w^{alt}}{\sigma_w^{alt}}\right) \leq \beta. \tag{4.1}$$

Setting β at 20%, that is, to have at least 80% power and at fixed $\alpha(= 0.025)$, n is determined from (4.1). We vary λ_E^{alt} to get the minimum sample size satisfying at least 80% power for each λ_E^{alt}.

4.2. *Exact Bayesian approach*

The Bayesian decision rule to declare NI as given in (3.1) can be written as:

$$P\left(\lambda_E - \theta\lambda_R - (1-\theta)\lambda_P > 0 | \lambda_R > \lambda_P,\ \text{Data}\right) > p^*, \tag{4.2}$$

Define $\eta_{RP} = \lambda_R - \lambda_P$. Since the probability in (4.2) does not have a closed form, it is approximated by generating samples from the posterior distribution and estimating the probability as

$$\begin{aligned} P\left(\lambda_E - \theta\lambda_R - (1-\theta)\lambda_P > 0 | \lambda_R > \lambda_P, \text{Data}\right) &= P\left(\left(\lambda_E - \lambda_P\right) > \theta\left(\lambda_R - \lambda_P\right) | \lambda_R > \lambda_P,\ \text{Data}\right) \\ = \int_0^\infty P\left(\lambda_E - \lambda_P > \theta c | \eta_{RP} = c, \text{Data}\right) f_{\eta_{RP}|\eta_{RP}>0}(c)\, dc &\approx \frac{1}{M}\sum_{i=1}^{M} g(\theta c_i, \text{Data}), \end{aligned} \tag{4.3}$$

where $g\left(\theta c_i, \text{Data}\right) = P\left(\lambda_E - \lambda_P > \theta c_i | \eta_{RP} = c_i, \text{Data}\right)$ and c_i being the ith sample value of $(\lambda_R - \lambda_P | \lambda_R > \lambda_P)$. We repeat the calculation of the estimated probability given in the left-hand side of the equation (4.2) for n^* times and obtain the proportion of times it exceeds the cutoff p^*. In simulation, the value of n^* is usually chosen to be 1000. As in the previous two approaches, we keep λ_R, λ_P, and θ fixed and vary λ_E such that $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P)$ varies within the range $[0.5, 1.5]$. For $\lambda_E^{\text{alt}} > \lambda_E^{\text{null}}$, the estimated power of the test can be calculated as

$$\hat{\text{Power}} = \frac{\text{No. of times } P\left(\lambda_E^{\text{alt}} - \theta\lambda_R - (1-\theta)\lambda_P > 0 | \lambda_R > \lambda_P, \text{Data}\right) > p^*}{n^*}.$$

The sample size can be obtained by setting the estimated power to be at least $100(1-\beta)\%$, β is usually chosen to be 0.2. We note here that since under the exact Bayesian approach the estimation of power involves generating samples from posterior distributions, there could be minor fluctuation in the estimated sample size.

4.3. *Approximate Bayesian approach*

For sample size determination under the approximation-based Bayesian approach, we choose "n" that satisfies the two conditions (Gamalo *and others*, 2014): (C1) $P\left[P\left(\nu_T \geq 0 | \text{Data}\right) > p^* | H_0\right] \leq \alpha$, (C2) $P\left[P\left(\nu_T \geq 0 | \text{Data}\right) > p^* | H_1\right] \geq 1-\beta$, where the probability in (C1) is the estimated average type-I error while that in (C2) is the estimated power of the test, β being the type-II error. The sample size is determined from condition (C2) by fixing β to have at least $100\,(1-\beta)\,\%$ power and simultaneously satisfying condition (C1). As in the Frequentist approach, we choose $\alpha = 0.025$. We note that $P\left(\nu_T \geq 0 | \text{Data}\right) = P\left(\left(\nu_T - \tilde{\sigma}_T^2\widetilde{\mu}_T\right)/\widetilde{\sigma}_T \geq -\tilde{\sigma}_T^2\widetilde{\mu}_T/\widetilde{\sigma}_T\right) > p^* \Leftrightarrow -\tilde{\sigma}_T\widetilde{\mu}_T \leq z_{1-p^*} \Leftrightarrow T \geq -z_{1-p^*}\left(1/\sigma_T^2 + 1/\sigma_\nu^{*2}\right)^{1/2}\sigma_T^2 - \mu_\nu^*/\sigma_\nu^{*2}\sigma_T^2$, where $\widetilde{\mu}_T = T/\sigma_T^2 + \mu_\nu^*/\sigma_\nu^{*2}$ and $\widetilde{\sigma}_T^2 = 1/(1/\sigma_T^2 + 1/\sigma_\nu^{*2})$ (see Supplementary Appendix C available at *Biostatistics* online). Here, z_{1-p^*} is the $100\,(1-p^*)$ % of the $N(0,1)$ distribution. Now, the power function is obtained by varying λ_E such that $0.5 \leq (\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) \leq 1.5$, keeping the other rates 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

$$P_{H_0}\left[T > -z_{1-p^*}\left(\frac{1}{\sigma_T^{2\text{null}}} + \frac{1}{\sigma_\nu^{*2}}\right)^{1/2}\sigma_T^{2\text{null}} - \frac{\mu_\nu^*}{\sigma_\nu^{*2}}\sigma_T^{2\text{null}}\right] \leq \alpha$$

$$\Leftrightarrow P_{H_0}\left[\frac{T-\mu_T^{\text{null}}}{\sigma_T^{\text{null}}} > \left(-z_{1-p^*}\left(\frac{1}{\sigma_T^{2\text{null}}}+\frac{1}{\sigma_v^{*2}}\right)^{1/2}\sigma_T^{2\text{null}} - \frac{\mu_v^*}{\sigma_v^{*2}}\sigma_T^{2\text{null}} - \mu_T^{\text{null}}\right)\Big/\sigma_T^{\text{null}}\right] \le \alpha \qquad (4.4)$$

$$\Leftrightarrow \Phi\left(z_{1-p^*}\left(\frac{1}{\sigma_T^{2\text{null}}}+\frac{1}{\sigma_v^{*2}}\right)^{1/2}\sigma_T^{\text{null}} + \frac{\mu_v^*}{\sigma_v^{*2}}\sigma_T^{\text{null}} + \frac{\mu_T^{\text{null}}}{\sigma_T^{\text{null}}}\right) \le \alpha.$$

Similarly, condition (C2) becomes

$$\Phi\left(z_{1-p^*}\left(\frac{1}{\sigma_T^{2\text{alt}}}+\frac{1}{\sigma_v^{*2}}\right)^{1/2}\sigma_T^{\text{alt}} + \frac{\mu_v^*}{\sigma_v^{*2}}\sigma_T^{\text{alt}} + \frac{\mu_T^{\text{alt}}}{\sigma_T^{\text{alt}}}\right) \ge 1-\beta. \qquad (4.5)$$

A similar derivation albeit for two-arm NI trial for binary outcome can be found in Gamalo *and others* (2014). Now, "n" can be solved from (4.5) by setting $\beta = 20\%$ and simultaneously satisfying condition (C1) for each λ_E^{alt} (which is included in μ_T^{alt}).

4.4. *Sample size under different allocation*

We determine the approximate sample size to attain a power of $1-\beta = 0.8$ under three different allocation scenarios for (E,R,P): (1:1:1), (2:2:1), and $(3:2:1)$ of the total sample size $N(= n(1+r_1+r_2))$. We express the sample sizes in the reference and the placebo group as proportions r_1 and r_2 of the sample size n_E in the experimental group. Hence, for the allocation $(1:1:1)$, $r_1 = r_2 = 1$; for (2:2:1), $r_1 = 1$ and $r_2 = \frac{1}{2}$; and for (3:2 : 1) the values are $r_1 = \frac{2}{3}$ and $r_2 = \frac{1}{3}$. Type-I error or $\alpha = 0.025$ is kept fixed for the Frequentist approach, while for Bayesian approach we also made sure the equations (4.4) and (4.5) hold simultaneously for fixed (α,β). In practice, θ is chosen in $[0.5,1)$, to ensure retention of at least 50% effect of the active control. The sample sizes are presented for $\theta = 0.8$ and 0.75 and for a range of λ_E keeping $\lambda_R = 21$ and $\lambda_P = 7$ in Table 2. Other values of λ's are also possible satisfying the restriction $\lambda_R > \lambda_P$.

We present the sample size for the placebo group (n_P). The sample sizes n_R and n_E for the arms R and E can be obtained from the allocation ratios. The total sample size for (1:1:1) is $N = 3n_P^{(1)}$, that for (2:2:1) is $N = 5n_P^{(2)}$, while for (3:2:1) it is $N = 6n_P^{(3)}$, where $n_P^{(1)}$, $n_P^{(2)}$, and $n_P^{(3)}$ are the respective sample sizes for the placebo group under the three allocations. Although appealing at first glance, one may not want to use a balanced study design because of two aspects: (i) due to ethical reasons in case an effective treatment exists, the number of patients receiving the placebo should be kept as small as possible and (ii) as pointed out by Koch and Tangen (1999), the difference between E and R should be expected to be much smaller than their respective difference from placebo so that the latter are easier to detect. From Table 2, we note that the necessary sample size is remarkably smaller for the unbalanced allocation (2:2:1) as compared to a balanced design and a minor reduction is again obtained for the unbalanced allocation (3:2:1) as compared to (2:2:1). Some additional results on this are also provided in Supplementary Appendix available at *Biostatistics* online.

4.5. *Sample size for marginal vs. conditional Frequentist approach*

To make a comparison of the existing marginal Frequentist approach with the proposed conditional Frequentist approach one, we present the sample sizes under both the approaches in Table 3. For simplicity, we only consider equal allocation to the three treatment arms. We determine the sample size under the two approaches for $\theta = \{0.9, 0.8\}$ with $(\lambda_R = 21, \lambda_P = 7)$, $(\lambda_R = 18, \lambda_P = 17.5)$, and $(\lambda_R = 7.5, \lambda_P = 7)$. From Table 3, we observe that for $\lambda_R = 21$ and $\lambda_P = 7$ the sample size under the conditional approach is identical to that calculated under the marginal approach, while for $\lambda_R = 18$ and $\lambda_P = 17.5$ or $\lambda_R = 7$

Table 2. *Sample sizes based on exact and approximate approaches to achieve a power of* 80% *for* $\theta = 0.8$ *and* 0.75, $\alpha = 0.025$ *and keeping* $\lambda_R = 21$ *and* $\lambda_P = 7$ *under three different allocations. The simulated power* ($\hat{\phi}$) *and estimated average type-I error* ($\hat{\alpha}$) *for exact Bayesian approach under non-informative Gamma prior are also reported to show that calculated sample size is adequate to guarantee* 80% *power except for minor numerical fluctuation. Note, Frequentist type-I error is always strictly maintained at* $\alpha = 0.025$ *by equation 4.1.*

					Frequentist normal			Approximate Bayesian			Exact Bayesian			
E	R	P	θ	λ_E	n_P	N	$\hat{\phi}$	n_P	N	$\hat{\phi}$	n_P	N	$\hat{\phi}$	$\hat{\alpha}$
1	1	1	0.80	20.0	79	237	0.802	78	234	0.801	79	237	0.802	0.0215
				19.7	113	339	0.802	112	336	0.798	112	336	0.808	0.0222
				19.4	176	528	0.790	175	525	0.795	175	525	0.796	0.0225
				19.1	312	936	0.795	310	930	0.797	302	906	0.789	0.0229
				18.8	700	2100	0.803	697	2091	0.799	685	2055	0.802	0.0224
			0.75	20.0	39	117	0.810	38	114	0.813	38	114	0.807	0.0173
				19.7	50	150	0.806	48	144	0.798	48	144	0.786	0.0208
				19.4	66	198	0.805	65	195	0.804	65	195	0.804	0.0217
				19.1	93	279	0.803	91	273	0.804	88	264	0.790	0.0185
				18.8	140	420	0.801	138	414	0.806	133	399	0.787	0.0184
2	2	1	0.80	20.0	40	200	0.805	40	200	0.808	37	185	0.794	0.0219
				19.7	57	285	0.801	57	285	0.803	52	260	0.798	0.0208
				19.4	89	445	0.799	89	445	0.802	81	405	0.783	0.0217
				19.1	158	790	0.798	157	785	0.800	153	765	0.805	0.0193
				18.8	353	1765	0.804	352	1760	0.802	351	1755	0.797	0.0189
			0.75	20.0	20	100	0.813	19	95	0.800	18	90	0.821	0.0210
				19.7	26	130	0.816	25	125	0.809	24	120	0.811	0.0201
				19.4	34	170	0.808	33	165	0.801	33	165	0.815	0.0181
				19.1	48	240	0.813	46	230	0.801	45	225	0.831	0.0181
				18.8	72	360	0.808	70	350	0.804	64	320	0.810	0.0180
3	2	1	0.80	20.0	33	198	0.819	32	192	0.813	31	186	0.795	0.0216
				19.7	47	282	0.804	46	276	0.799	44	264	0.805	0.0210
				19.4	72	432	0.798	71	426	0.796	71	426	0.786	0.0199
				19.1	128	768	0.803	127	762	0.799	125	750	0.795	0.0205
				18.8	287	1722	0.800	284	1704	0.797	277	1662	0.782	0.0209
			0.75	20.0	16	96	0.814	15	90	0.799	15	90	0.802	0.0225
				19.7	21	126	0.821	20	120	0.813	18	108	0.787	0.0201
				19.4	27	162	0.799	27	162	0.809	26	156	0.802	0.0216
				19.1	38	228	0.807	37	222	0.804	37	222	0.796	0.0193
				18.8	58	348	0.807	56	336	0.800	54	324	0.811	0.0188

and $\lambda_P = 7.5$, the sample size under the conditional approach is smaller than the existing one to achieve a power of 80%. This observation points out that for smaller difference between λ_R and λ_P, the proposed conditional approach is more powerful than the existing marginal approach, while for larger difference both the approaches behave similarly. This in line with the theoretical result we have proven in Lemma 2.2.2.

5. Simulation studies

We enumerate few simulation studies to evaluate the performance of the Frequentist as well as Bayesian procedures presented above. We generate the power curves for different values of θ, under both the

Table 3. *Sample size for marginal vs. conditional Frequentist approach*

		Marginal ($\lambda_R = 21, \lambda_P = 7$)		Conditional			Marginal ($\lambda_R = 18, \lambda_P = 17.5$)		Conditional			Marginal ($\lambda_R = 7.5, \lambda_P = 7$)		Conditional	
θ	λ_E	n_P	N	n_P	N	λ_E	n_P	N	n_P	N	λ_E	n_P	N	n_P	N
0.9	23.0	26	78	26	78	20.3	48	144	44	132	10.0	18	54	16	48
	22.7	31	93	31	93	20.0	63	189	57	171	9.7	23	69	21	63
	22.4	38	114	38	114	19.7	86	258	79	237	9.4	30	90	27	81
	22.1	47	141	47	141	19.4	124	372	115	345	9.1	41	123	38	114
	21.8	61	183	61	183	19.1	197	591	185	555	8.8	61	183	57	171
	21.5	81	243	81	243	18.8	359	1077	345	1035	8.5	100	300	91	273
0.8	23.0	12	36	12	36	20.3	43	129	40	120	10.0	16	48	15	45
	22.7	13	39	13	39	20.0	55	165	52	156	9.7	20	60	19	57
	22.4	15	45	15	45	19.7	75	225	71	213	9.4	26	78	25	75
	22.1	18	54	18	54	19.4	107	321	102	306	9.1	36	108	34	102
	21.8	20	60	20	60	19.1	167	501	160	480	8.8	52	156	50	150
	21.5	24	72	24	72	18.8	295	885	287	861	8.5	84	252	80	240

conjugate and non-conjugate priors and make a comparison among the informative and relatively non-informative Gamma priors under the conjugate set up. We consider a randomized trial with the sample size allocation ratio $n_E{:}n_R{:}n_P = 1{:}1{:}1$. Unequal sample size allocation is also possible and shown in Table 2 from the sample size perspective. However, to maintain brevity for the current power comparisons only equal allocation is described in detail.

5.1. *Simulation steps*

The following simulation steps are used to calculate the type-I error and power for the two different prior scenarios described earlier: (i) conjugate Gamma prior and (ii) a non-conjugate prior. For the conjugate prior setting, we choose two sets of hyper-parameters, one of which is relatively informative with respect to the other. Note that the priors are so chosen that the mean of the Gamma distribution equals the Poisson rates and shrinking the variance for the informative priors compared to the non-informative ones. For the non-conjugate prior, we put non-informative Gamma prior on λ_E and suitable values are chosen for the Beta and Gamma hyper-parameters. In the following, we give the formal steps of the simulation:

Step 1. Specify n_E, n_R, n_P (or, the allocation ratios), λ_l, $l \in \{E, R, P\}$ with $\lambda_R > \lambda_P$, and θ so that $\lambda_E \in [\lambda_P + 0.5(\lambda_R - \lambda_P),\ \lambda_P + 1.5(\lambda_R - \lambda_P)]$ to generate $\{X_E, X_R, X_P\} = \text{Data}$.

Step 2. For a given value of $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P)$ or equivalently λ_E, generate the data X_l from Poisson distribution Poisson $(n_l\lambda_l)$, $l \in \{E, R, P\}$.

Step 3. Generate M many posterior samples from the posterior distribution under the two priors given under Section 3.1. For the conjugate prior, we keep only those posterior values in the sample for which $\lambda_R > \lambda_P$. For the non-conjugate prior, the posterior sample values satisfy $\lambda_R > \lambda_P$ automatically because of the in-built restriction. For the mth posterior sample, calculate the ratio $(\lambda_E^m - \lambda_P^m)/(\lambda_R^m - \lambda_P^m)$.

Step 4. Calculate the posterior probability:

$$P\left(\frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} > \theta | \lambda_R > \lambda_P, \text{Data}\right) \approx \frac{1}{M}\sum_{m=1}^{M} I\left(\frac{\lambda_E^m - \lambda_P^m}{\lambda_R^m - \lambda_P^m} > \theta | \lambda_R^m > \lambda_P^m, \text{Data}\right).$$

Step 5. Bayesian decision criterion: If $P((\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) > \theta | \lambda_R > \lambda_P, \text{Data}) > p^*$, increase COUNTS by 1; otherwise 0.

Step 6. Go back to step 2 and repeat the simulation n^* (a large number chosen *a priori*) number of times:

i. Calculate the type-I error by using COUNTS divided by n^* for λ_E satisfying $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) = \theta$.

ii. Calculate the power by using COUNTS divided by n^* for λ_E satisfying $(\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) > \theta$.

Step 7. The power curve is generated for a range of λ_E such that $0.5 \leq (\lambda_E - \lambda_P)/(\lambda_R - \lambda_P) \leq 1.5$.

Note that under the Frequentist and approximation-based Bayesian approaches, Step 3 is not needed and Step 5 needs to be replaced by the corresponding decision criterion given in Section 2.2 and Section 3.2, respectively.

5.2. *Simulation results*

For the conjugate prior, we chose the number of posterior samplers, M, to be 1000. For the non-conjugate prior, a trace plot of posterior estimate for each parameter suggests $M = 1000$ MCMC samplers, where every 50th value of 50,000 MCMC samples taken as a value in the sample with 1000 burn-ins, are more than sufficient to produce stable estimate of the parameters. We consider $\lambda_R = 21$, $\lambda_P = 7$, and varying λ_E as in Table 2 for generating the power curves. Additionally, we also consider another specification of the parameters: $\lambda_R = 7$, $\lambda_P = 1$, and set λ_E in the range [4, 9], to see the behavior of the power curves for smaller values of λ_l, $l \in \{E, R, P\}$. The choice of p^* is an important criteria. Following the Frequentist set up, we choose $p^* = 0.975$. However, as reported in Gamalo *and others* (2011), this choice of p^* could give too restrictive type-I error. One way to alleviate this problem is to perform Bayesian calibration; however, it is not pursued in the present paper. In Figure 1, we present four power curves corresponding to four different values of $\theta = \{0.8, 0.75, 0.7, 0.65\}$ with $n = 100$ for parameter specification ($\lambda_R = 21$, $\lambda_P = 7$) and ($\lambda_R = 7$, $\lambda_P = 1$) under the Frequentist and Bayesian conjugate prior. We see that as θ increases, the power curve shifts to the right as the proposed NI test is more powerful for smaller values of θ. This is because for smaller values of θ, it is easier to declare NI. Note that for the exact Bayesian approach, we chose the Jeffreys prior as Gamma(0.5, 0.00001) which is a flat prior having large variance. The Jeffreys prior is obtained by equating $\sqrt{(I(\lambda))} = c\lambda^{-0.5}$ with the density of Gamma(α, β) and thus solving for α, β, and the constant c, $I(\lambda)$ is the Fisher information of λ. This prior is also used in computing Table 2. For interested reader an excellent discussion on choosing other neutral and non-informative priors on Gamma distribution is given in Kerman *and others* (2011). The horizontal red line in the Figure 1 corresponds to $\alpha = 0.025$. The type-I error rate under the Frequentist approach is always maintained at $\alpha = 0.025$, while that under the exact Bayesian approach is maintained at or below $\alpha = 0.025$ (see Table 2). Additional results on simulation for comparing conjugate vs. non-conjugate as well informative vs. non-informative priors are provided in Supplementary Appendix available at *Biostatistics* online.

6. Application

We illustrate our proposed Frequentist and Bayesian methodology with a MS (Calabrese *and others*, 2012) example described in Section 1.1. The lesions of MS typically arise within the optic nerves, spinal cord, brain stem, and the periventricular white matter of the cerebral hemispheres. Neuropathological techniques and magnetic resonance imaging (MRI) are used to identify the relationship of lesions to cortical veins. Further details of the data are presented in Table 1. For our NI testing, we consider GA as the experimental treatment (E), subcutaneous (sc) IFN beta-1a as the reference drug (R), and no therapy as the placebo

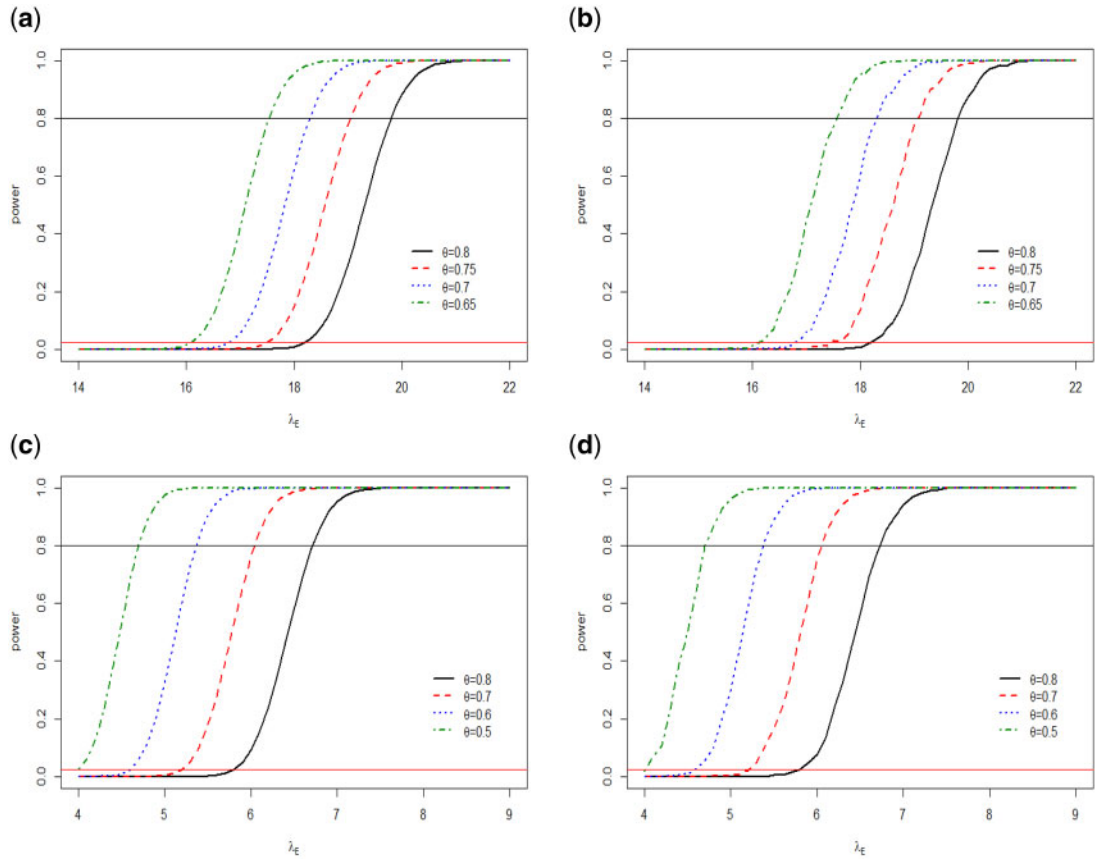

Fig. 1. Power curves for different θ under two sets of Poisson distribution parameter values (1) $\lambda_R = 21$, $\lambda_P = 7$ (top row) and (2) $\lambda_R = 7$, $\lambda_P = 1$ (bottom row). (a and c, left column) for Frequentist approach, while (b and d, right column) for exact Bayesian conjugate prior.

(P) for both 1-year and 2-year data. As indicated before, Poisson model provides satisfactory result in goodness of fit test for each arm. For our illustration, we analyze both 1-year and 2-year CLs count data separately using the exact Bayesian method under different priors, as well as using the Frequentist method. However, for formulating the original NI hypothesis, we assumed higher rate indicates greater treatment benefit, but here, for the lesion count data, smaller rate indicates higher treatment benefit. So, we reformulated the hypothesis in (2.3) as

$$H_0 : \lambda_E - \theta\lambda_R - (1-\theta)\,\lambda_P \geq 0 \text{ vs. } H_1 : \lambda_E - \theta\lambda_R - (1-\theta)\,\lambda_P < 0. \tag{6.1}$$

The required AS condition will be $(\lambda_P > \lambda_R)$. Now, from H_0 given in (6.1) we have the following equivalent condition: $(\lambda_E - \lambda_P) - \theta\,(\lambda_R - \lambda_P) \geq 0 \Rightarrow (\lambda_P - \lambda_E) \leq \theta\,(\lambda_P - \lambda_R) \Rightarrow (\lambda_P - \lambda_E)/(\lambda_P - \lambda_R) \leq \theta$. Hence, the alternative hypothesis H_1 in (6.1) becomes $H_1 : (\lambda_P - \lambda_E)/(\lambda_P - \lambda_R) > \theta$. From this, the Bayesian decision criteria is

$$P\left(H_1 : \frac{\lambda_E - \lambda_P}{\lambda_R - \lambda_P} > \theta | \text{Data}\right) > p^*. \tag{6.2}$$

This shows that the Bayesian decision rule remains unchanged as in the previous case. We use $p^* = 0.975$ to determine NI of GA over sc-IFN beta-1a.

From Table 1, we observe that after 12 months 37/50 (74%) of the patients who did not receive any therapy developed $\geq$1 new CLs counts, 12/46 (26%) patients treated with sc-IFN beta-1a and 24/48 (50%) treated with GA respectively developed at least one lesian count. These figures after 24 months came out as 41/50 (82%) for the patients with no therapy, 24/46 (52%) for those treated with sc-IFN beta-1a, and 30/48 (62%) for the GA-treated patients. So, we observe that the percentages of at least one lesian count increased from 1 year to 2 years for all treatment arms. Also, the calculated rates of occurrence of the CLs for no therapy, reference, and test drug, are respectively, 1.53, 0.37, and 0.79 after 1-year and 2.94, 0.72, and 1.29 after 2 years. The rate of the untreated (placebo) group is much higher than those of the treated groups, which indicates that the treatments have beneficial effect in lowering the new CLs development. We first carry out the analysis under the Frequentist approach and calculate the p-value for testing the hypothesis in equation (6.1) as p-value $= P_{H_0}(W < W_{\text{obs}})$, where W is the test statistic given by $W = (\hat{\lambda}_E - \theta\hat{\lambda}_R - (1-\theta)\,\hat{\lambda}_P | \hat{\lambda}_P > \hat{\lambda}_R)$ and W_{obs} is the observed value of W. The p-value is then compared with $\alpha = 0.025$ to deduce the Frequentist decision of NI. For Bayesian conjugate prior, we carry out the analysis assuming both non-informative and informative priors. For the non-informative case, we assume Jeffreys prior as Gamma $(0.5, 0.00001)$ for λ_l, $l \in \{E, R, P\}$ and generate posterior samplers for the three rates from Gamma distributions as in Step 3 of simulation studies, but keeping samples satisfying $\lambda_P > \lambda_R$ to account for the AS condition. We calculate the posterior probability for the rejection of H_1 as given in the left hand side of (6.2). We report $P(H_1|\text{Data})$ in Table 4 for different values of θ in the range $[0.5, 1)$, in order to ensure that the test drug has a meaningful effect. These posterior probabilities are compared with p^* to deduce the Bayesian decision. In Table 4, we also report the decisions: 1 (if NI is claimed) or 0 (otherwise) for Frequentist as well as Bayesian analyses. From Table 4, we observe that the posterior probabilities increase as the values of θ decrease implying higher chance of declaring NI for smaller values of θ. For the 1-year data, under Jeffreys prior, we see that the average posterior probability $P(H_1|\text{Data})$ are small for all values of θ meaning that NI of GA cannot be claimed for $\theta \in [0.5, 0.8]$. This is due to the fact that the rate of lesion count occurrence for GA-treated patients is higher than those treated with sc-IFN beta-1a, which gives an indication that GA is possibly inferior to sc-IFN beta-1a since its effect is not within the NI margin. But if we choose an informative prior with suitable parameters, then NI can be claimed for $\theta \leq 0.55$. In this case, we chose the following priors for the three arms respectively: E: Gamma$(8, 10)$, R: Gamma$(20, 5)$, and P: Gamma$(12, 8)$. For the 2-year data also, the rate of lesian count for GA is higher than that of the reference group; however, the difference between the rates is within the NI margin δ, to claim NI of GA for small values of θ, even under the Jeffreys prior. Considering informative prior, we can still improve on the posterior probabilities. Choosing the following priors: E: Gamma$(64, 49.6)$, R: Gamma$(12, 17)$, P: Gamma$(60, 20.4)$, NI is claimed for $\theta \leq 0.6$. Finally, considering the non-conjugate prior, for the reformulated hypothesis in (6.1), we assume the following: $u_1 = \lambda_R/\lambda_P \sim$ Beta(a, b),$u_2 = \lambda_P \sim$ Gamma(p, r), where $0 < u_1 < 1$, which satisfies the AS condition $(\lambda_R > \lambda_P)$ and $u_2 > 0$. For the 1-year data, assuming a relatively non-informative prior: Gamma $(0.8, 1)$ for E; Gamma $(1.5, 1)$ for P; and Beta $(1, 3.1)$ for λ_R/λ_P; we observe that NI cannot be claimed for any $\theta \in [0.5, 0.8]$. However, if the following priors are chosen: Gamma $(160, 200)$ for E; Gamma $(600, 400)$ for P; and Beta $(200, 630)$ for λ_R/λ_P; then NI can be claimed for $\theta = 0.5$. Also for the 2-year data, the observations are similar for the non-informative prior in the non-conjugate setting. NI cannot be claimed for the priors: Gamma $(0.8, 0.62)$ for E; Gamma $(0.75, 0.255)$ for P; and Beta $(1.2, 3.7)$ from the ratio λ_R/λ_P. However, for the relatively informative priors: Gamma $(80, 62)$ for E; Gamma $(75, 25.5)$ for P; and Beta $(12, 37)$ for $\frac{\lambda_R}{\lambda_P}$; NI can be claimed for $\theta = 0.5$. We note that the hyper-parameters for both conjugate and non-conjugate priors are so chosen that the mean of the Gamma distribution equals the estimated count rate in the respective arms. Also, we observed that for the 1-year data, more informative

Table 4. Bayesian and Frequentist decision in the lesian count data where "1" stands for the rejection and "0" stands for acceptance of the null hypothesis. Also posterior probabilities are reported under different priors.

	Posterior probabilities					Decision			
	Conjugate		Non-conjugate			Conjugate		Non-conjugate	
θ	Non-informative	Informative	Non-informative	Informative	Frequentist decision	Non-informative	Informative	Non-informative	Informative
					1-year data				
0.80	0.102	0.717	0.220	0.152	0	0	0	0	0
0.75	0.198	0.786	0.275	0.220	0	0	0	0	0
0.70	0.322	0.859	0.310	0.291	0	0	0	0	0
0.65	0.461	0.925	0.350	0.398	0	0	0	0	0
0.60	0.590	0.957	0.384	0.615	0	0	0	0	0
0.55	.00717	0.976	0.413	0.845	0	0	1	0	0
0.50	0.829	0.988	0.445	0.976	0	0	1	0	1
					2-year data				
0.80	0.262	0.226	0.266	0.280	0	0	0	0	0
0.75	0.481	0.456	0.302	0.456	0	0	0	0	0
0.70	0.705	0.729	0.336	0.650	0	0	0	0	0
0.65	0.847	0.902	0.355	0.821	0	0	0	0	0
0.60	0.930	0.979	0.381	0.913	0	0	1	0	0
0.55	0.983	0.998	0.404	0.963	1	1	1	0	0
0.50	0.994	0.999	0.431	0.990	1	1	1	0	1

priors are needed to claim NI, as compared to the 2-year data. This indicates that the present trial data for 1-year end-point does not support NI strongly, while for 2-year endpoint, NI can be claimed if we choose $\theta < 0.6$.

7. Discussion

According to several guidelines, the NI margin should be pre-specified in the protocol, while some allows flexibility of pre-specifying a fixed amount of effect retention (e.g., FDA, 2016; ICH Steering Committee, 1998, 2000; EMA, 2005; Wangge *and others*, 2013). Thus the value of the NI margin can vary greatly depending on the estimated effect size of the reference treatment ($\lambda_R - \lambda_P$). In this article, we presented novel Frequentist and Bayesian test procedures for three-arm NI trial under fraction margin approach. We proposed more powerful conditional test (Lemma 2.2.2) based on Frequentist principle which directly incorporates the AS condition in NI testing. We believe this is a better usage of available information. Under AS assumption, conditional principle is more realistic and more powerful than the traditional marginal NI testing and it does not result in a biased test (e.g., joint testing of NI and AS). In the conditional Frequentist approach, we conditioned the NI test statistic on $\hat{\lambda}_R > \hat{\lambda}_P$; however, it is very much possible to condition it based on the AS test statistic. However, this is not done in the current paper as that will make Bayesian (conditioned on $\lambda_R > \lambda_P$) and Frequentist approach incomparable, since then each approach will use slightly different conditioning statement. In the Bayesian context, we explored conjugate prior and also specified more flexible non-conjugate prior choices. In Section 4.2, for integer-valued parameters we have also shown an interesting connection between Bayesian posterior probability to Frequentist exact probability. This could be further exploited to connect Bayesian and Frequentist sample size in the line of Zaslavsky (2013). Since Bayesian power calculation requires additional computation, we tabulated the sample size in Table 2 under three different types of allocation. We hope that the clinicians will find this readily useful in designing such NI trial.

We have observed that the Bayesian normal approximation and the exact Bayesian approach yield greater power and hence require smaller sample size compared to the Frequentist approach. Albeit, we would like caution an user about the control of type-I error in Bayesian context as pointed out in recent papers by Kopp-Schneider *and others* (2019) and Psioda and Ibrahim (2018). It is reported that with informative prior strict type-I error control in the Frequentist sense is not possible under Bayesian setup. In this article, all reported type-I errors are "average type-I error" as defined in Gravestock *and others* (2017), which is essentially an average over all possible outcomes under null distribution. We thank an anonymous reviewer for pointing this out. Also, it is evident that an unbalanced allocation of the sample size in NI trial results in the reduction of the required number of patients to achieve a certain power. According to Pigeot *and others* (2003), an unbalanced allocation of the total sample size in a NI trial is desirable from ethical and substantial point of view. We also applied our proposed Bayesian test procedure on a clinical trial data on MS. The results suggest that the Bayesian methods perform favorably in all situations and that these methods do not depend on any asymptotic approximation as the Frequentist method.

Notably, with Poisson distributed outcomes, rate/count difference is not the only function of interest. In the binary context, apart from risk difference similar methods for risk ratio and odds ratio has been developed in both Frequentist (Chowdhury *and others*, 2018b) and Bayesian context Chowdhury *and others* (2018a) very recently. In a similar line one may frame two-arm NI trial using the ratio of Poisson rates as done in Stucke and Kieser (2013) for two-arm trial. However, for a three-arm trial, defining such a functional (in ratio form) is non-trivial. We are currently developing both the Frequentist and Bayesian methods for these types of functionals. Also for the count type outcome over-dispersion (and under-dispersion) is a frequent issue and Poisson model is not an ideal choice. However, given the dearth of Bayesian article for count data, we did not consider those issues in the current paper. One could use negative binomial (Mütze *and others*, 2016) or generalized Poisson distribution instead, however the

resulting Bayesian (and Frequentist) procedure will be much more involved and as a result left as a future work.

SOFTWARE

The open source R codes for all the simulation studies and real data analyses performed in this manuscript are available at https://github.com/erina633/Poisson3armNI. Also, there is a README.md file which describes the contents of the R files and all the source functions. All the proofs and additional results are placed in Supplementary Appendix available at *Biostatistics* online.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared. The article reflects the views of the author and should not be construed to represent FDA's views or policies.

FUNDING

The research of first author is partly supported by PCORI (contract number ME-1409-21410); and NIH (P30-ES020957).

REFERENCES

ARNOLD, B. C. AND BEAVER, R. J. (1993). The nontruncated marginal of a truncated bivariate normal distribution. *Psychometrika* **58**, 471–488.

BERGER, R. L. (1997). Likelihood ratio tests and intersection-union tests. In *Advances in Statistical Decision Theory and Applications*. Boston: Birkhäuser, pp. 225–237.

CALABRESE, M., BERNARDI, V., ATZORI, M., MATTISI, I., FAVARETTO, A., RINALDI, F., PERINI, P., AND GALLO, P. (2012). Effect of disease-modifying drugs on cortical lesions and atrophy in relapsing-remitting multiple sclerosis. *Multiple Sclerosis Journal* **18**, 418–424.

CHOWDHURY, S., TIWARI, R., AND GHOSH, S. (2018a). Approaches for testing non-inferiority in two-arm trial for risk ratio and odds ratio. *Journal of Biopharmaceutical Statistics* **29**, 425–445.

CHOWDHURY, S., TIWARI, R. C., AND GHOSH, S. (2018b). Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial. *Computational Statistics & Data Analysis*.

CHUANG-STEIN, C., STRYSZAK, P., DMITRIENKO, A., AND OFFEN, W. (2007). Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine* **26**, 1181–1192.

D'AGOSTINO SR, R. B., MASSARO, J. M., AND SULLIVAN, L. M. (2003). Non-inferiority trials: design concepts and issues—the encounters of academic consultants in statistics. *Statistics in Medicine* **22**, 169–186.

EMA (2005). *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.

FDA (2016). *Non-inferiority Clinical Trials to Establish Effectiveness: Guidance for Industry.* Silver Spring, MD: 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).

FRIEDE, T. AND SCHMIDLI, H. (2010). Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Statistics in Medicine* **29**, 1145–1156.

GAMALO, M. A., TIWARI, R. C., AND LAVANGE, L. M. (2014). Bayesian approach to the design and analysis of non-inferiority trials for anti-infective products. *Pharmaceutical Statistics* **13**, 25–40.

GAMALO, M. A., WU, R., AND TIWARI, R. C. (2011). Bayesian approach to noninferiority trials for proportions. *Journal of Biopharmaceutical Statistics*, **21**, 902–919.

GAMALO, M. A., WU, R., AND TIWARI, R. C. (2016). Bayesian approach to non-inferiority trials for normal means. *Statistical Methods in Medical Research* **25**, 221–240.

GBUR, E. E. (1981). On the poisson index of dispersion: On the poisson index. *Communications in Statistics-Simulation and Computation* **10**, 531–535.

GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. (2014). *Bayesian Data Analysis*, Volume 2. Boca Raton, FL: Chapman.

GHOSH, P., NATHOO, F., GONEN, M., AND TIWARI, R. C. (2011). Assessing noninferiority in a three-arm trial using the Bayesian approach. *Statistics in Medicine* **30**, 1795–1808.

GHOSH, S., CHATTERJEE, A., AND GHOSH, S. (2017). Non-inferiority test based on transformations for non-normal distributions. *Computational Statistics & Data Analysis* **113**, 73–87.

GHOSH, S., GHOSH, S., AND TIWARI, R. (2016). Bayesian approach for assessing non-inferiority in a three-arm trial with pre-specified margin. *Statistics in Medicine* **35**, 695–708.

GRAVESTOCK, I., HELD, L.; COMBACTE-Net Consortium. (2017). Adaptive power priors with empirical Bayes for clinical trials. *Pharmaceutical Statistics* **16**, 349–360.

HIDA, E. AND TANGO, T. (2013). Three-arm noninferiority trials with a prespecified margin for inference of the difference in the proportions of binary endpoints. *Journal of Biopharmaceutical Statistics* **23**, 774–789.

HUANG, L., ZALKIKAR, J., AND TIWARI, R. C. (2011). A likelihood ratio test based method for signal detection with application to FDA's drug safety data. *Journal of the American Statistical Association* **106**, 1230–1241.

HUNG, H. M. J. AND WANG, S. J. (2004). Multiple testing of noninferiority hypotheses in active controlled trials. *Journal of Biopharmaceutical Statistics* **14**, 327–335.

ICH STEERING COMMITTEE (1998). ICH harmonised tripartite guideline: statistical principles for clinical trials.

ICH STEERING COMMITTEE (2000). ICH harmonised tripartite guideline: choice of control group and related issues in clinical trials.

KERMAN, J. (2011). Neutral noninformative and informative conjugate beta and gamma prior distributions. *Electronic Journal of Statistics* **5**, 1450–1470.

KIESER, M. AND FRIEDE, T. (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine* **26**, 253–273.

KIESER, M. AND STUCKE, K. (2016). Assessing additional benefit in noninferiority trials. *Biometrical Journal* **58**, 154–169.

KOCH, A. AND RÖHMEL, J. (2004). 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**, 315–325.

KOCH, G. G. AND TANGEN, C. M. (1999). Non parametric analysis of covariance and its role in non-inferiority clinical trials. *Drug Information Journal* **33**, 1145–1159.

KOPP-SCHNEIDER, A., CALDERAZZO, S., AND WIESENFARTH, M. (2019). Power gains by using external information in clinical trials are typically not possible when requiring strict type I error control. *Biometrical Journal* **62**, 361–374.

KULLDORFF, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods* **26**, 1481–1496.

LU, N. T., XU, Y., AND YANG, Y. (2018). Incorporating a companion test into the noninferiority design of medical device trials. *Journal of Biopharmaceutical Statistics* **29**, 143–150.

MCINTOSH, J. (2001). Analyzing counts, durations, and recurrences in clinical trials. *Journal of Biopharmaceutical Statistics* **11**, 65–74.

MIELKE, M., MUNK, A., AND SCHACHT, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Statistics in Medicine* **27**, 5093–5110.

MÜTZE, T., MUNK, A., AND FRIEDE, T. (2016). Design and analysis of three-arm trials with negative binomially distributed endpoints. *Statistics in Medicine* **35**, 505–521.

NOSEWORTHY, J. H. (2003). Management of multiple sclerosis: current trials and future options. *Current Opinion in Neurology* **16**, 289–297.

PIGEOT, I., SCHÄFER, J., RÖHMEL, J., AND HAUSCHKE, D. (2003). Assessing noninferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine* **22**, 883–899.

PLUMMER, M., STUKALOV, A., AND DENWOOD, M. (2016). Bayesian graphical models using MCMC. *R News* **364**, 1.

PSIODA, M. A. AND IBRAHIM, J. G. (2018). Bayesian clinical trial design using historical data that inform the treatment effect. *Biostatistics* **20**, 400–415.

SCHUMI, J. AND WITTES, J. T. (2011). Through the looking glass: understanding non-inferiority. *Trials* **12**, 106–118.

SILCOCKS, P., WHITHAM, D., AND WHITEHOUSE, W. P. (2010). P3MC: a double blind parallel group randomised placebo controlled trial of Propranolol and Pizotifen in preventing migraine in children. *Trials*, **11**, 71.

SIMON, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics* **55**, 484–487.

SOEIRO-DE SOUZA, M. G., ANDREAZZA, A. C., CARVALHO, A. F., MACHADO-VIEIRA, R., YOUNG, L. T., AND MORENO, R. A. (2013). Number of manic episodes is associated with elevated DNA oxidation in bipolar I disorder. *International Journal of Neuropsychopharmacology* **16**, 1505–1512.

SORMANI, M., BRUZZI, P., ROVARIS, M., BARKHOF, F., COMI, G., MILLER, D., CUTTER, G., AND FILIPPI, M. (2001). Modelling new enhancing MRI lesion counts in multiple sclerosis. *Multiple Sclerosis Journal* **7**, 298–304.

STUCKE, K. AND KIESER, M. (2013). Sample size calculations for noninferiority trials with Poisson distributed count data. *Biometrical Journal* **55**, 203–216.

TSILIMINGRAS, D., SCHNIPPER, J., DUKE, A., AGENS, J., QUINTERO, S., BELLAMY, G., JANISSE, J., HELMKAMP, L., AND BATES, D. W. (2015). Post-discharge adverse events among urban and rural patients of an urban community hospital: a prospective cohort study. *Journal of General Internal Medicine* **30**, 1164–1171.

TSONG, Y. AND ZHANG, J. (2007). Simultaneous test for superiority and noninferiority hypotheses in active-controlled clinical trials. *Journal of Biopharmaceutical Statistics* **17**, 247–257.

WANGGE, G., PUTZEIST, M., KNOL, M. J., KLUNGEL, O. H., GISPEN-DE WIED, C. C., DE BOER, A., HOES, A. W., LEUFKENS, H. G., AND MANTEL-TEEUWISSE, A. K. (2013). Regulatory scientific advice on non-inferiority drug trials. *PLoS One* **8**, e74818.

WU, Y., LI, Y., HOU, Y., LI, K., AND ZHOU, X. (2018). Study duration for three-arm non-inferiority survival trials designed for accrual by cohorts. *Statistical Methods in Medical Research* **27**, 507–520.

XIE, T. AND AICKIN, M. (1997). A truncated Poisson regression model with applications to occurrence of adenomatous polyps. *Statistics in Medicine* **16**, 1845–1857.

ZASLAVSKY, B. G. (2013). Bayesian hypothesis testing in two-arm trials with dichotomous outcomes. *Biometrics* **69**, 157–163.

[*Received July 27, 2019; revised December 9, 2019; accepted for publication February 16, 2020*]