BMC
Medical Research Methodology

RESEARCH ARTICLE Open Access

Testing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints

Nian-Sheng Tang[1*], Bin Yu[1] and Man-Lai Tang[2]

Abstract

Background: A two-arm non-inferiority trial without a placebo is usually adopted to demonstrate that an experimental treatment is not worse than a reference treatment by a small pre-specified *non-inferiority margin* due to ethical concerns. *Selection of the non-inferiority margin* and *establishment of assay sensitivity* are two major issues in the design, analysis and interpretation for two-arm non-inferiority trials. Alternatively, a three-arm non-inferiority clinical trial including a placebo is usually conducted to assess the assay sensitivity and internal validity of a trial. Recently, some large-sample approaches have been developed to assess the non-inferiority of a new treatment based on the three-arm trial design. However, these methods behave badly with small sample sizes in the three arms. This manuscript aims to develop some reliable small-sample methods to test three-arm non-inferiority.

Methods: Saddlepoint approximation, exact and approximate unconditional, and bootstrap-resampling methods are developed to calculate p-values of the Wald-type, score and likelihood ratio tests. Simulation studies are conducted to evaluate their performance in terms of type I error rate and power.

Results: Our empirical results show that the saddlepoint approximation method generally behaves better than the asymptotic method based on the Wald-type test statistic. For small sample sizes, approximate unconditional and bootstrap-resampling methods based on the score test statistic perform better in the sense that their corresponding type I error rates are generally closer to the prespecified nominal level than those of other test procedures.

Conclusions: Both approximate unconditional and bootstrap-resampling test procedures based on the score test statistic are generally recommended for three-arm non-inferiority trials with binary outcomes.

Keywords: Approximate unconditional test, Bootstrap-resampling test, Non-inferiority trial, Rate difference, Saddlepoint approximation, Three-arm design

Background

The objective of a non-inferiority trial is to demonstrate the efficacy of an experimental treatment not being inferior to a reference treatment by some pre-specified non-inferiority margin. Many authors considered two-arm non-inferiority trials without a placebo since the comparison between the experimental and reference treatments is direct and the potential ethical problems encountered in traditional placebo-controlled trials are avoided (for example, see Dunnett and Gent [1], Tango [2], and Tang et al. [3]). However, there are two major concerns for two-arm non-inferiority trials [4]. The first issue is the choice of the non-inferiority margin, which is the clinically acceptable amount or a combination of statistical reasoning and clinical judgement. The other issue is the evaluation of assay sensitivity, which refers to the ability of a trial to differentiate an effective treatment from a less effective or ineffective treatment [5]. Without a placebo arm, the assay sensitivity of a trail is not demonstrable from the trial data and ones must rely on some external information (e.g., historical placebo trails) for the reference treatment [4]. Without the trial assay sensitivity, any non-inferiority testing results from the comparison of the experimental and reference treatments will become unconvincing. There are some indications where it is considered ethically acceptable to continue to randomize patients to placebo despite the fact that an effective treatment exists and there is interest in seeing not only

*Correspondence: nstang@ynu.edu.cn
[1]Department of Statistics, Yunnan University, No.2 Cuihu North Road, 650091 Kunming, China
Full list of author information is available at the end of the article

whether the new treatment works at all but also how it measures up to accepted therapy. In this case, a three-arm non-inferiority clinical trail including the experimental treatment, an active reference treatment and a placebo is usually conducted to assess assay sensitivity and internal validation of a trail [6]. Indeed, three-arm trials are recommended in the guidelines of the ICH (The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use) and EMEA/CPMP (European Medicines Agency/Committee for Proprietary Medical Products) as a useful approach to the assessment of assay sensitivity and internal validation (e.g., see [7]).

Statistical inference based on three-arm non-inferiority clinical trials with normally distributed outcomes has received considerable attention in recent years. For example, Koch and Tangen [8] and Pigeot et al. [9] considered the problem of three-arm non-inferiority testing for normally distributed endpoints with a common but unknown variance. Koti [10] presented a new approach for normally distributed endpoints based on the Fieller-Hinkley distribution. Hasler, Vonk and Hothorn [11] proposed the usage of the t-distribution in the presence of heteroscedasticity. Hida and Tango [7] proposed a test procedure for assessing the assay sensitivity with a pre-specified margin defined as a difference between treatments in the presence of homoscedasticity. Ghosh, Nathoo, Gönen and Tiwari [12] developed a Bayesian approach in the presence of heteroscedasticity by incorporating both parametric and semi-parametric models. Gamalo, Muthukumarana, Ghosh and Tiwari [13] extended the existing generalized p-value approach for assessing the non-inferiority of a new treatment in a three-arm trial.

Recently, some statistical methods have also been developed for three-arm non-inferiority testing with binary endpoints. For example, Tang and Tang [14] proposed two asymptotic approaches for testing three-arm non-inferiority via rate difference based on Wald-type and score test statistics. Kieser and Friede (2007) revisited the performance of Tang and Tang's [14] asymptotic test statistics via simulation studies and derived approximate sample size formulae for achieving the desired power. Munk, Mielke, Skipka and Freitag [15] developed likelihood ratio tests. Li and Gao [4] used the closed testing principle to establish the hierarchical testing procedure and proposed a group sequential type design. Liu, Tzeng and Tsou [16] presented a three-step testing procedure and derived an optimal sample size allocation rule in an ethical and reliable manner that minimizes the total sample size.

All aforementioned approaches for testing non-inferiority of a new treatment in a three-arm clinical trial with binary endpoints are based on large sample theory, and their accuracy has long been suspected and criticized when sample sizes are small or the data structure is sparse. To the best of our knowledge, limited work have been done to address these issues. Motivated by Jensen [17], we derive saddlepoint approximations to the cumulative distribution functions of Wald-type, score and likelihood ratio test statistics. Inspired by Tang and Tang [18], we also propose the exact unconditional, approximate unconditional and Bootstrap-resampling p-value calculation procedures for testing three-arm non-inferiority with small sample sizes.

The rest of this article is organized as follows. We first review three test statistics for assessing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints. We also propose saddlepoint approximation, exact and approximate unconditional, and bootstrap-resampling approaches for calculating p-values. Simulation studies are conducted to investigate the performance of all test statistics based on different p-value calculation approaches in terms of type I error rate and power. An example is analyzed to demonstrate our methodologies. Finally, we discuss the performance of our proposed methodologies and present some conclusions.

Methods

Model

Let consider a clinical trial with the test (T), reference (R) and placebo (P) treatments, and assume their primary clinical outcomes X_T, X_R and X_P be independent and binomially distributed as $X_T \sim \text{Bin}(n_T, \pi_T)$, $X_R \sim \text{Bin}(n_R, \pi_R)$ and $X_P \sim \text{Bin}(n_P, \pi_P)$, respectively. Here, X_T, X_R and X_P are the numbers of responses in groups T, R and P, respectively, π_T, π_R and π_P represent their corresponding response probabilities with higher probability indicating a more favorable outcome, and n_T, n_R and n_P denote their corresponding sample sizes. Thus, the joint probability density function of (x_T, x_R, x_P) is given by

$$\begin{aligned} &f(x_T, x_R, x_P|\pi_T, \pi_R, \pi_P) \\ &\quad = \tbinom{n_T}{x_T}\tbinom{n_R}{x_R}\tbinom{n_P}{x_P}\pi_T^{x_T}(1-\pi_T)^{n_T-x_T}\pi_R^{x_R} \\ &\qquad \times (1-\pi_R)^{n_R-x_R}\pi_P^{x_P}(1-\pi_P)^{n_P-x_P}. \end{aligned} \tag{2.1}$$

It can be easily shown from Equation (2.1) that the maximum likelihood estimates (MLEs) of π_T, π_R and π_P are given by $\hat{\pi}_T = x_T/n_T$, $\hat{\pi}_R = x_R/n_R$ and $\hat{\pi}_P = x_P/n_P$, respectively.

Test statistics

Following Hida and Tango [7], to test the non-inferiority of the experimental treatment to the reference with the assay sensitivity in a three-arm trial, we have to simultaneously demonstrate (i) the superiority of the experimental treatment to the placebo, (ii) the non-inferiority of the experimental treatment to the reference with a non-inferiority margin $\Delta > 0$, and (iii) the superiority of the

reference treatment to the placebo by more than Δ. That is, π_T, π_R and π_P must satisfy the following inequalities: $\pi_P < \pi_R - \Delta < \pi_T$, which can be written as the following two hypotheses:

$$H_0 : \pi_T \le \pi_R - \Delta \quad \text{versus} \quad H_1 : \pi_T > \pi_R - \Delta,$$

$$K_0 : \pi_R \le \pi_P + \Delta \quad \text{versus} \quad K_1 : \pi_R > \pi_P + \Delta.$$

Similar to Pigeot et al. [9], we take the margin Δ as a fraction f of the effect size of the reference treatment, i.e., $\Delta = f(\pi_R - \pi_P)$. Generally, one can select $f = 1/2$ and $1/3$ [14]. Thus, the second hypothesis can be expressed as $K_0 : \pi_R \le \pi_P$ versus $K_1 : \pi_R > \pi_P$. If K_0 is rejected, letting $f = 1-\theta$ yields the following non-inferiority hypothesis:

$$H_0 : \frac{\pi_T - \pi_P}{\pi_R - \pi_P} \le \theta \quad \text{versus} \quad H_1 : \frac{\pi_T - \pi_P}{\pi_R - \pi_P} > \theta, \tag{2.2}$$

where $\theta \in (0, 1)$ is a fixed retention fraction [8]. Rejecting H_0 implies that the test treatment preserves at least $100\theta\%$ of the efficacy of the reference treatment compared to placebo [19]. Similar to Tang and Tang [14], we only consider hypothesis H_0 and assume that K_0 is rejected at some pre-given significant level. Thus, the non-inferiority hypothesis (2.2) can be rewritten as

$$\begin{aligned} &H_0 : \pi_T - \theta\pi_R - (1-\theta)\pi_P \le 0 \quad \text{versus} \\ &H_1 : \pi_T - \theta\pi_R - (1-\theta)\pi_P > 0. \end{aligned} \tag{2.3}$$

Let $\psi = \pi_T - \theta\pi_R - (1-\theta)\pi_P$. The non-inferiority hypothesis (2.3) can be expressed as

$$H_0 : \psi \le 0 \quad \text{versus} \quad H_1 : \psi > 0. \tag{2.4}$$

The restricted maximum likelihood estimates (RMLEs) (denoted by $\tilde{\pi}_T, \tilde{\pi}_R, \tilde{\pi}_P$) of π_T, π_R and π_P can be computed as follows. If the MLEs $\hat{\pi}_T, \hat{\pi}_R, \hat{\pi}_P$ of π_T, π_R, π_P satisfy the conditions: $\hat{\pi}_T - \theta\hat{\pi}_R - (1-\theta)\hat{\pi}_P \le 0$ and $\hat{\pi}_R - \hat{\pi}_P > 0$, we take $\tilde{\pi}_T = \hat{\pi}_T$, $\tilde{\pi}_R = \hat{\pi}_R$ and $\tilde{\pi}_P = \hat{\pi}_P$; otherwise, the RMLEs can be calculated by setting $\pi_T = \theta\pi_R + (1-\theta)\pi_P$ in the likelihood function (2.1) and maximizing it with respect to π_R and π_P. For the latter, it follows from Equation (2.1) that the RMLEs of π_R and π_P can be obtained by simultaneously solving the following equations in the parameter space $\Theta = \{(\pi_P, \pi_R) : 0 \le \pi_P < \pi_R \le 1\}$:

$$\begin{aligned} &\frac{x_T - n_T(\theta\pi_R + (1-\theta)\pi_P)}{(\theta\pi_R + (1-\theta)\pi_P)(1 - \theta\pi_R - (1-\theta)\pi_P)} \\ &= \frac{n_R\pi_R - x_R}{\theta\pi_R(1-\pi_R)} = \frac{n_P\pi_P - x_P}{(1-\theta)\pi_P(1-\pi_P)}. \end{aligned}$$

It is possible that there is no point $(\pi_P, \pi_R) \in \Theta$ such that it satisfies the above equations, which implies that the likelihood function given in Equation (2.1) attains its maximum on the boundary of the parameter space Θ.

Following Tang and Tang [14], ψ can be estimated by $\hat{\psi} = \hat{\pi}_T - \theta\hat{\pi}_R - (1-\theta)\hat{\pi}_P$, and its variance is given by $\text{var}(\hat{\psi}) = \pi_T(1-\pi_T)/n_T + \theta^2\pi_R(1-\pi_R)/n_R + (1-\theta)^2\pi_P(1-\pi_P)/n_P$, which can be estimated by $\sigma^2(\breve{\pi}) \triangleq \widehat{\text{var}}(\hat{\psi}) = \breve{\pi}_T(1-\breve{\pi}_T)/n_T + \theta^2\breve{\pi}_R(1-\breve{\pi}_R)/n_R + (1-\theta)^2\breve{\pi}_P(1-\breve{\pi}_P)/n_P$, where $\breve{\pi} = (\breve{\pi}_T, \breve{\pi}_R, \breve{\pi}_P)$ is some appropriate estimate of $\pi = (\pi_T, \pi_R, \pi_P)$, for example, taking $\breve{\pi}$ to be $\hat{\pi} = (\hat{\pi}_T, \hat{\pi}_R, \hat{\pi}_P)$ or $\tilde{\pi} = (\tilde{\pi}_T, \tilde{\pi}_R, \tilde{\pi}_P)$ which is the RMLE of π. Thus, the statistics for testing hypothesis (2.4) are given by

$$T_W = \hat{\psi}/\sigma(\hat{\pi}) \text{ and } T_R = \hat{\psi}/\sigma(\tilde{\pi}),$$

which are asymptotically distributed as the standard normal distribution under H_0 as n_T, n_R and n_P are sufficiently large. Hence, non-inferiority can be claimed if $T_W > z_{1-\alpha}$ (or $T_R > z_{1-\alpha}$), where $z_{1-\alpha}$ is the $(1-\alpha)$-quantile of the standard normal distribution. When $\pi_P = 0$, T_W is the Wald-type statistic proposed in Blackwelder [20] and T_R is the test statistic given by Farrington and Manning [21] for two-arm noninferiority trials.

The signed root of the likelihood ratio statistic for testing hypothesis (2.4) is given by

$$T_L = \text{sgn}(\hat{\psi})\sqrt{2\{\ell(\hat{\pi}) - \ell(\tilde{\pi})\}},$$

which is asymptotically distributed as the standard normal distribution under H_0 as n_T, n_R and n_P are sufficiently large, where $\ell(\pi) = x_T\log(\pi_T) + (n_T - x_T)\log(1-\pi_T) + x_R\log(\pi_R) + (n_R - x_R)\log(1-\pi_R) + x_P\log(\pi_P) + (n_P - x_P)\log(1-\pi_P) + \mathcal{C}$ with $\mathcal{C} = \log\{n_T!\, n_R!\, n_P!\} - \log\{x_T!\, x_R!\, x_P!\, (n_T - x_T)!\, (n_R - x_R)!\, (n_P - x_P)!\}$. Thus, non-inferiority can be claimed if $T_L > z_{1-\alpha}$.

p-value calculation methods

The non-inferiority hypothesis (2.2) can be claimed via the *p*-value method with the rule: H_0 is rejected if the *p*-value is less than or equal to the prespecified significance level α. In what follows, we introduce five approaches for calculating *p*-values based on t_j^0, which is the observed value of test statistic T_j ($j = W, R, L$) for the observed value $\left(x_T^o, x_R^o, x_P^o\right)$ of (X_T, X_R, X_P).

(1) Asymptotic method (AM)

It follows from the above arguments that all statistics T_j's ($j = W, R, L$) asymptotically follow the standard normal distribution under the null hypothesis $H_0 : \psi \le 0$. Thus, the asymptotic *p*-value for testing hypothesis (2.2) via statistic T_j ($j = W, R, L$) based on $\left(x_T^o, x_R^o, x_P^o\right)$ can be calculated by $p_j^{AM}\left(x_T^o, x_R^o, x_P^o\right) = P\left(T_j \ge t_j^o | H_0\right) = 1 - \Phi(t_j^o)$, where $\Phi(\cdot)$ is the standard normal distribution function.

The above asymptotic approach for calculating *p*-value of testing hypothesis (2.2) via statistic T_j ($j = W, RW, L$) is established under the large sample theory. Its accuracy has long been suspected and criticized, especially when n_T,

n_R and/or n_P are small since the skewness of the underlying binomial distributions is not taken into consideration. Some higher order corrections such as the saddlepoint approximation [17] have been proposed to improve the accuracy of the normal approximation. In what follows, we will derive saddlepoint approximations to distributions of the three test statistics.

(2) Saddlepoint approximation method (SAM)

Since X_T, X_R and X_P are independent and $X_i \sim \text{Bin}(n_i, \pi_i)$ $(i = T, R, P)$, the moment generating function of $\hat{\psi}$ is given by

$$\varphi(t) = \left(1-\pi_T+\pi_T e^{t/n_T}\right)^{n_T}\left(1-\pi_R+\pi_R e^{-\theta t/n_R}\right)^{n_R} \times \left(1-\pi_P+\pi_P e^{(\theta-1)t/n_P}\right)^{n_P},$$

with the cumulant generating function being

$$K(t) = n_T \log\left(1-\pi_T+\pi_T e^{t/n_T}\right) + n_R \log\left(1-\pi_R+\pi_R e^{-\theta t/n_R}\right) + n_P \log\left(1-\pi_P+\pi_P e^{(\theta-1)t/n_P}\right),$$

where $-1 \le t \le 1$. Thus, the first two derivatives of the cumulant generating function $K(t)$ are given by

$$\dot{K}(t) = \frac{\pi_T e^{t/n_T}}{1-\pi_T+\pi_T e^{t/n_T}} + \frac{-\theta\pi_R e^{-\theta t/n_R}}{1-\pi_R+\pi_R e^{-\theta t/n_R}} + \frac{(\theta-1)\pi_P e^{(\theta-1)t/n_P}}{1-\pi_P+\pi_P e^{(\theta-1)t/n_P}}, \text{ and}$$

$$\ddot{K}(t) = \frac{(1-\pi_T)\pi_T e^{t/n_T}}{n_T(1-\pi_T+\pi_T e^{t/n_T})^2} + \frac{\theta^2(1-\pi_R)\pi_R e^{-t\theta/n_R}}{n_R(1-\pi_R+\pi_R e^{-t\theta/n_R})^2} + \frac{(\theta-1)^2(1-\pi_P)\pi_P e^{(\theta-1)t/n_P}}{n_P(1-\pi_P+\pi_P e^{(\theta-1)t/n_P})^2},$$

respectively. To obtain the saddlepoint approximation to $P(\hat{\psi} \ge b)$, we need to solve the following saddlepoint equation: $\dot{K}(t) = b$ whose unique solution is denoted as $\hat{t}$. Following Jing and Robinson [22], the saddlepoint approximation to the cumulative distribution function of statistic $\hat{\psi}$ is given by

$$P(\hat{\psi} \ge b) \approx 1 - \Phi(\omega) + \phi(\omega)(1/\upsilon - 1/\omega),$$

where $\omega = \text{sgn}(\hat{t})\sqrt{2\{\hat{t}b - K(\hat{t})\}}$ and $\upsilon = \hat{t}\sqrt{\ddot{K}(\hat{t})}$. Thus, the saddlepoint approximation to $P\left(T_j \ge t_j^o | H_0\right)$ $(j = W, R, L)$ is given by

$$p_j^{SA}\left(x_T^o, x_R^o, x_P^o\right) = P\left(T_j \ge t_j^o | H_0\right) \approx 1 - \Phi\left(\omega_j^o\right) + \phi\left(\omega_j^o\right)\left(1/\nu_j^o - 1/\omega_j^o\right),$$

where $\omega_j^o = \text{sgn}(\hat{A}_j)\sqrt{2\left\{\hat{A}_j t_j^o - K(\hat{A}_j/B_j)\right\}}$ and $\upsilon_j^o = \hat{A}_j B_j^{-1}\sqrt{\ddot{K}(\hat{A}_j/B_j)}$, $\hat{A}_j$ is the unique solution to equation: $\dot{K}(\hat{A}_j/B_j) = t_j^o B_j$ for $j = W, R$ with $B_W = \sigma(\hat{\pi})$ and $B_R = \sigma(\tilde{\pi})$, $\omega_L^o = \text{sgn}(\hat{\psi})\sqrt{2\{\ell(\hat{\pi}) - \ell(\tilde{\pi})\}}$ and $\upsilon_L^o = \hat{\psi}\sqrt{n_T\mathcal{H}_1/\mathcal{H}_2}$ with $\mathcal{H}_1 = n_T n_R n_P(\theta\hat{\pi}_R + (1-\theta)\hat{\pi}_P)(1 - \theta\hat{\pi}_R - (1-\theta)\hat{\pi}_P)\hat{\pi}_R(1-\hat{\pi}_R)\hat{\pi}_P(1-\hat{\pi}_P)$, and $\mathcal{H}_2 = n_R n_P \tilde{\pi}_R(1-\tilde{\pi}_R)\tilde{\pi}_P(1-\tilde{\pi}_P)$.

(3) Exact unconditional method (EUM)

When sample sizes (i.e., n_T, n_R, n_P) are small, asymptotic methods may yield inflated type I error rates and their exact versions may provide reliable alternative. Under $H_0 : \psi \le 0$ with $\pi_P < \pi_R$, parameters π_R and π_P must belong to the following constrained parameter space $\Omega = \{(\pi_P, \pi_R) : 0 \le \pi_P < \pi_R \le 1$ if $-\theta\pi_R < \psi < 0$, $(-\psi - \theta\pi_R)/(1-\theta) \le \pi_P < \pi_R < 1$ if $-\pi_R < \psi \le -\theta\pi_R$, and empty set otherwise$\}$. Under the null hypothesis, the probability density function (2.1) can be reexpressed by $\pi_T = \psi + \theta\pi_R + (1-\theta)\pi_P$ with π_R, π_P and ψ being nuisance parameters. These nuisance parameters can be eliminated by maximizing the null likelihood over the complete domain Ω. Similar to Tang and Tang [18], the exact unconditional p-value for testing $H_0 : \psi \le 0$ via statistic T_j $(j = W, R, L)$ based on $\left(x_T^o, x_R^o, x_P^o\right)$ is defined as

$$p_j^{\text{EU}}\left(x_T^o, x_R^o, x_P^o\right) = \sup_{\psi \le 0}\left\{\sup_{(\pi_R,\pi_P)\in\Omega} P\left(T_j \ge t_j^o | \psi, \pi_R, \pi_P\right)\right\},$$

where

$$P\left(T_j \ge t_j^o | \psi, \pi_R, \pi_P\right) = \sum_{x_T=0}^{n_T}\sum_{x_R=0}^{n_R}\sum_{x_P=0}^{n_P} \times \frac{n_T!}{x_T!(n_T-x_T)!}\frac{n_R!}{x_R!(n_R-x_R)!}\frac{n_P!}{x_P!(n_P-x_P)!} \times (\psi+\theta\pi_R+(1-\theta)\pi_P)^{x_T} \times (1-\psi-\theta\pi_R-(1-\theta)\pi_P)^{n_T-x_T} \times \pi_R^{x_R}(1-\pi_R)^{n_R-x_R}\pi_P^{x_P}(1-\pi_P)^{n_P-x_P} I \times \left\{T_j(x_T, x_R, x_P) \ge t_j^o\right\},$$

and $I\left\{T_j(x_T, x_R, x_P) \ge t_j^o\right\}$ is 1 if $T_j(x_T, x_R, x_P) \ge t_j^o$ and 0 otherwise.

(4) Approximate unconditional method (AUM)

According to Tang and Tang [18] and Tang, Tang and Rosner [23], the exact unconditional test is always conservative, i.e., its corresponding type I error rate is always less than or equal to the prespecified significance level. Following Tang and Tang [18], these nuisance parameters can be eliminated by evaluating their values at their corresponding RMLEs under $\psi = 0$. The approximate unconditional p-value for testing $H_0 : \psi \le 0$ via statistic T_j $(j = W, R, L)$ based on $\left(x_T^o, x_R^o, x_P^o\right)$ can be defined as $p_j^{\text{AU}}\left(x_T^o, x_R^o, x_P^o\right) = P\left(T_j \ge t_j^o | \psi = 0, \pi_R = \tilde{\pi}_R, \pi_P = \tilde{\pi}_P\right)$.

(5) Bootstrap-resampling method (BTM)

Hypothesis testing based on the bootstrap-resampling method is usually recommended when sample sizes (i.e., n_T, n_R and n_P) are small [24] or data structure is sparse (e.g., x_T or x_R or x_P is close to zero or n_T, n_R and n_P,

respectively). Given the observation $\left(x_T^o, x_R^o, x_P^o\right)$, we compute the RMLEs $\tilde{\pi}_T, \tilde{\pi}_R$ and $\tilde{\pi}_P$ of parameters π_T, π_R and π_P, and calculate the observed value t_j^0 of statistic T_j ($j = W, R, L$). Based on the RMLEs $\tilde{\pi}_T, \tilde{\pi}_R$ and $\tilde{\pi}_P$, we generate B bootstrap samples $\left\{\left(x_T^b, x_R^b, x_P^b\right) : b = 1, \ldots, B\right\}$ from the following distribution: $x_k^b \sim \text{Bin}(n_k, \tilde{\pi}_k)$ for $k = T, R$ and P. For each of the B bootstrap samples, we compute the observed value t_j^b of statistic T_j ($j = W, R, L$). Hence, an approximate p-value for testing $H_0 : \psi \leq 0$ via statistic T_j based on $\left(x_T^o, x_R^o, x_P^o\right)$ is given by $\hat{p}_j^{BT}\left(x_T^o, x_R^o, x_P^o\right) = \frac{1}{B}\sum_{b=1}^{B} I\left(t_j^b \geq t_j^0\right)$.

For any given observation $\left(x_T^o, x_R^o, x_P^o\right)$, test statistic T_j ($j = W, R, L$) and p-value calculation method, we reject the null hypothesis H_0 at the significance level α if $p_j^k\left(x_T^o, x_R^o, x_P^o\right) \leq \alpha$ for k =AM, SA, EU, AU and BT.

Simulation study

Simulation studies are conducted to investigate the performance of various test statistics together with the five p-value calculation methods in small-sample designs (e.g., $n = 30$ and 60, where $n = n_P + n_R + n_T$ with the allocation ratios λ_P:λ_R:λ_T=1:n_R/n_P:n_T/n_P taking to be 1:1:1, 1:2:2 and 1:2:3) in terms of type I error rate and power. For each (n_P, n_R, n_T), we consider the following probability settings [19]: $\pi_P = 0.05, 0.10, 0.15, \ldots, 0.50$, $\pi_R = \pi_P + 0.05, \pi_P + 0.10, \ldots, 0.95$, and $\pi_T = \theta\pi_R + (1-\theta)\pi_P$, which corresponds to a total of 11,340 configurations of (π_P, π_R, π_T), and the following two non-inferiority margins: $\theta = 0.6$ and 0.8. The nominal level is taken to be $\alpha = 0.05$. For the given values of n and allocation ratio λ_P:λ_R:λ_T, n_k is given by $n_\ell = n\lambda_k/(\lambda_P + \lambda_R + \lambda_T)$ for $\ell = P, R$ and T. Thus, given n, allocation ratio and (π_P, π_R, π_T), the type I error rate for testing hypothesis $H_0 : \psi \leq 0$ versus $H_1 : \psi > 0$ via test statistic T_j ($j = W, R, L$) at the significance level α is calculated by

$$\alpha_j^k = \sum_{x_T^o=0}^{n_T}\sum_{x_R^o=0}^{n_R}\sum_{x_P^o=0}^{n_P} f\left(x_T^o, x_R^o, x_P^o | \pi_T, \pi_R, \pi_P, H_0\right) \times I\left\{p_j^k\left(x_T^o, x_R^o, x_P^o\right) \leq \alpha\right\}$$

for $k = AM, SAM, EUM, AUM$ and BTM, whilst the corresponding power can be evaluated by replacing H_0 in $f\left(x_T^o, x_R^o, x_P^o | \pi_T, \pi_R, \pi_P, H_0\right)$ by H_1.

Results

Simulation study

To compare the performance of AM, SAM, EUM, AUM and BTM together with test statistics T_W, T_R and T_L under the balanced and unbalanced designs, Figure 1 presents boxplots of their corresponding type I error rates for $n = 30$ and 60, and λ_P:λ_R:λ_T=1:1:1, 1:2:2 and 1:2:3, where AMk, SAk, EUk, AUk and BTk represent AM, SAM, EUM, AUM and BTM for test statistic T_k with k=W, R and L, respectively. Here, each boxplot in Figure 1 contains 2 (i.e., the number of non-inferiority margins)$\times$11,340 (i.e., the number of configurations for (π_P, π_R, π_T))=22,680 data points. From Figure 1, we have the following findings. First, the medians of the type I error rates based on AUM and BTM are closer to the prespecified nominal level $\alpha = 0.05$ than those based on the other three p-value calculation methods for all three test statistics under consideration. Second, for AUM and BTM, the medians of the type I error rates for test statistics T_W and T_R, which are 0.0495 and 0.0501 for AUM and 0.0494 and 0.0494 for BTM respectively, are closer to $\alpha = 0.05$ than those for test statistic T_L, which are 0.0442 for AUM and 0.0442 for BTM. Third, for AM, SAM and EUM, their corresponding medians of type I error rates are 0.0649, 0.0455 and 0.0260 for test statistic T_W, 0.0504, 0.0455 and 0.0488 for test statistic T_R, and 0.0663, 0.1285 and 0.0332 for test statistic T_L, respectively, which indicate that (i) the AM is liberal for test statistics T_W and T_L, whilst it is valid for test statistic T_R; (ii) the SAM can improve the accuracy of the normal approximation for test statistics T_W and T_R; and (iii) the EUM is conservative for all test statistics. Fourth, the proportions of configurations whose type I error rates lie in the interval $(0.045, 0.055)$ for AM, SAM, EUM, AUM and BTM are 0.0747, 0.4691, 0.0710, 0.5154 and 0.7994 for T_W, 0.5605, 0.4605, 0.4753, 0.7167 and 0.8370 for T_R, and 0.0784, 0.0800, 0.0691, 0.4056 and 0.4889 for T_L, respectively, which show that (i) AUM and BTM outperform the other three p-value calculation procedures, and (ii) T_R behaves better than the other two test statistics regardless of p-value calculation procedures. Fifth, the median of the type I error rates becomes more close to the prespecified nominal level as the total sample size n increases, whilst at the same time the variability of the type I error rates decreases. Sixth, the variability of the type I error rates for unbalanced designs is not significantly different from that for the balanced designs.

To investigate the sensitivity of various p-value calculation procedures (i.e., AM, SAM, EUM, AUM and BTM) to different test statistics, Figure 2 presents boxplots of their corresponding type I error rates against π_P for test statistics T_W, T_R and T_L. Examination of Figure 2 shows that there is no significant effect of π_P on the type I error rate.

We also calculate powers of the five p-value calculation procedures together with the three test statistics at the nominal level $\alpha = 0.05$ when $\pi_T = \pi_R$ and $\theta = 0.6$ with the following settings: $n = 30$ and 60, $\pi_P = 0.15$ and 0.3, and $\pi_R = 0.5, 0.8$ and 0.95 for the balanced allocation 1:1:1 and unbalanced allocation 1:2:3. Results are reported in Table 1. Examination of Table 1 indicates that (i) T_R is generally more powerful than T_W and T_L for the EUM

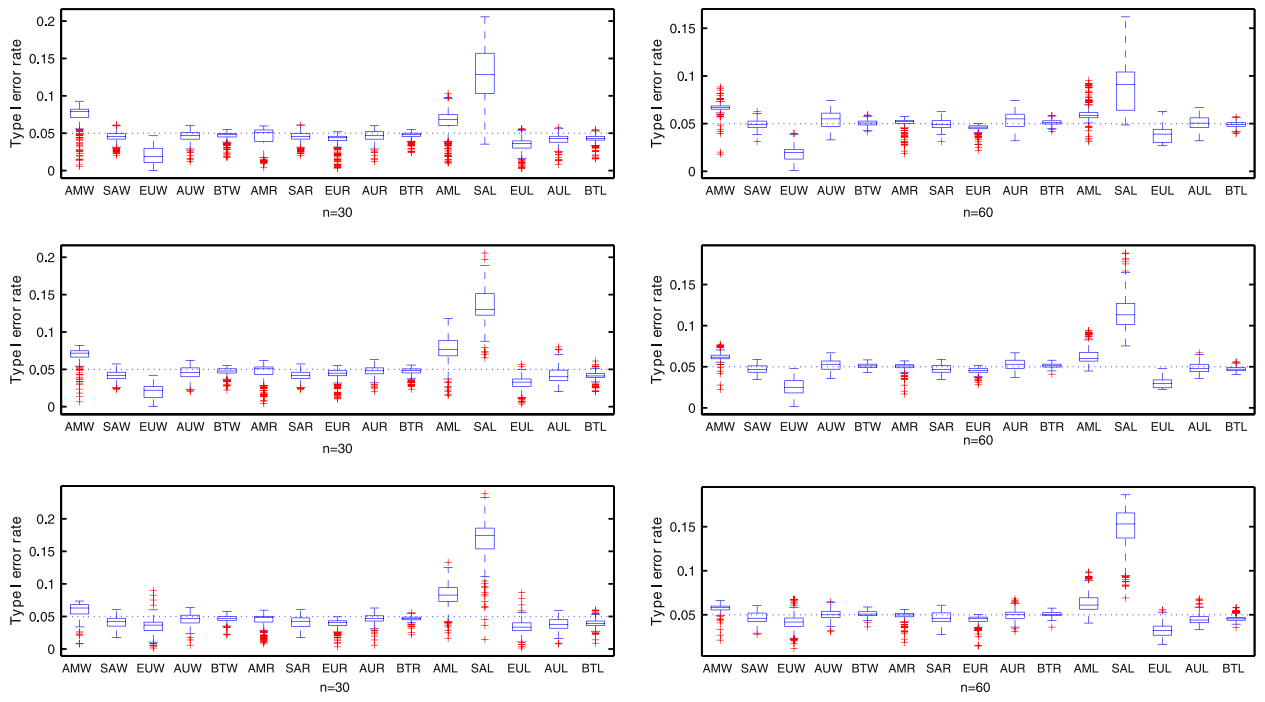

Figure 1 Boxplots of the type I error rates of various test procedures together with three statistics when testing the non-inferiority hypothesis (2.2) at $\alpha = 0.05$. AMk, SAk, EUk, AUk and BTk represent the AM, SA, EU, AU and BT test procedures with test statistic T_k for k = W, R and T, respectively.

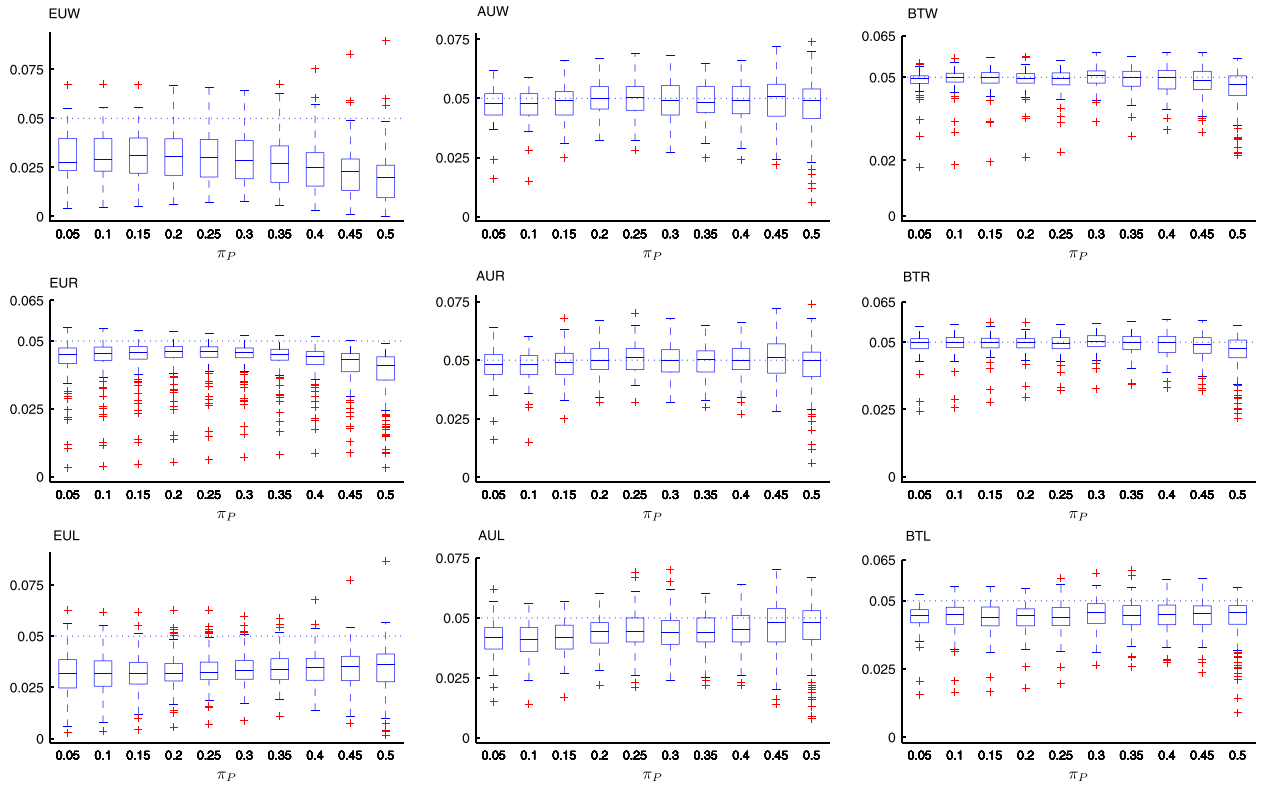

Figure 2 Boxplots of the type I error rates of various test procedures together with three statistics against π_P when testing the non-inferiority hypothesis (2.2) at $\alpha = 0.05$. EUk, AUk and BTk represent the EU, AU and BT test procedures with statistic T_k for k = W,R and T, respectively.

Table 1 Exact powers (%) of various test procedures together with three statistics when $\pi_T = \pi_R$ with $n = 30$ and 60, $\theta = 0.6$ and $\alpha = 0.05$

				AM			SAM			EUM			AUM			BTM		
n	λ_P:λ_R:λ_T	π_P	π_R	T_W	T_R	T_L	T_W	T_R	T_L	T_W	T_R	T_L	T_W	T_R	T_L	T_W	T_R	T_L
30	1:1:1	0.15	0.5	13.4	12.3	13.9	43.6	21.2	22.5	5.0	18.1	15.3	18.2	18.2	14.9	17.9	18.0	16.2
			0.8	44.2	43.9	38.0	42.6	72.4	30.0	28.1	43.5	40.3	42.1	42.1	39.7	43.7	43.9	42.1
			0.95	86.0	85.8	75.2	95.6	97.7	36.8	67.8	79.1	76.0	74.2	74.2	71.3	75.8	75.4	75.0
		0.3	0.5	8.8	7.5	8.3	39.9	23.2	14.8	2.9	10.7	8.5	10.8	10.8	9.3	11.1	11.1	9.1
			0.8	30.1	29.2	22.0	21.1	35.6	26.6	15.9	29.2	24.8	30.0	30.0	26.8	30.2	30.7	29.0
			0.95	66.3	64.1	45.6	80.6	88.3	26.4	42.5	52.7	49.8	59.4	59.4	57.1	60.0	59.7	59.4
	1:2:3	0.15	0.5	14.2	11.9	18.6	33.2	28.2	24.7	13.2	17.9	13.7	21.7	21.7	19.7	20.3	20.0	18.4
			0.8	43.6	41.4	48.3	58.9	81.3	29.0	43.4	45.4	36.3	53.1	52.3	44.8	51.5	51.1	44.4
			0.95	83.0	82.9	83.9	97.1	98.8	36.0	85.6	79.6	82.6	85.9	84.3	79.9	85.4	84.9	80.6
		0.3	0.5	8.6	7.2	10.3	36.2	25.0	18.7	7.9	10.8	7.6	12.4	12.2	10.5	11.9	11.7	9.8
			0.8	29.7	27.4	30.1	26.1	57.9	25.9	28.9	31.9	22.7	35.0	33.6	28.2	34.4	33.7	29.5
			0.95	62.4	61.8	62.8	85.5	95.3	36.9	66.4	63.0	60.9	65.8	64.5	62.3	66.3	65.9	64.0
60	1:1:1	0.15	0.5	19.0	19.0	18.4	33.7	47.5	24.3	10.7	11.3	14.2	29.3	29.4	28.3	28.0	28.1	27.2
			0.8	65.8	67.8	59.6	85.2	92.4	38.9	55.0	56.3	48.7	71.4	71.4	71.4	71.1	71.1	70.7
			0.95	97.9	98.6	96.6	96.6	96.7	50.3	95.9	96.7	89.3	97.7	97.7	97.7	97.7	97.7	97.7
		0.3	0.5	9.7	9.4	9.5	43.2	21.5	16.9	4.7	5.3	4.2	17.0	17.2	15.3	14.1	14.3	13.1
			0.8	46.5	47.1	39.8	54.6	79.8	33.2	35.4	36.9	37.1	50.9	50.9	50.8	49.7	50.3	50.0
			0.95	91.3	93.3	85.7	95.3	96.0	47.0	85.9	87.8	71.3	88.0	88.0	88.0	89.6	89.3	89.5
	1:2:3	0.15	0.5	20.5	20.2	22.2	29.6	53.8	27.9	24.1	22.1	24.2	31.0	30.8	28.3	31.7	31.1	28.2
			0.8	72.5	72.5	69.1	92.3	96.4	40.9	73.9	73.3	79.2	77.0	76.9	75.9	78.3	78.1	76.7
			0.95	98.6	98.6	98.0	99.9	99.9	50.6	98.6	98.9	92.4	99.1	99.0	99.0	98.5	98.5	98.4
		0.3	0.5	10.3	10.0	10.3	42.9	25.0	20.1	12.3	10.3	10.1	15.8	15.7	13.5	15.8	15.4	13.4
			0.8	49.3	49.2	45.1	64.6	84.1	36.2	52.0	48.4	38.0	52.0	52.0	51.0	55.5	55.4	54.1
			0.95	90.4	90.3	88.7	99.1	99.4	51.3	90.9	92.5	82.4	92.1	92.0	92.0	91.8	91.7	91.7

Table 2 Various *p*-values for the pharmacological data set at the nominal level $\alpha = 5\%$

	$\theta = 0.6$			$\theta = 0.8$		
Test method	T_W	T_R	T_L	T_W	T_R	T_L
AM	0.173	0.162	0.164	0.234	0.229	0.230
SAM	0.494	0.494	0.140	0.497	0.497	0.162
EUM	0.185	0.181	0.192	0.233	0.202	0.210
AUM	0.166	0.165	0.186	0.232	0.230	0.249
BTM	0.504	0.502	0.519	0.516	0.514	0.530

except for $\pi_R = 0.95$ with the unbalanced designs, (ii) T_W and T_R have similar powers for AM, AUM and BTM under our considered settings, (iii) a slight power difference is observed between T_R and T_L for AUM and BTM, (iv) there is slight power difference between balanced and unbalanced designs, and (v) power increases as n increases regardless of p-value calculation procedures or test statistics. Hence, we would recommend both AUM and BTM with T_R for hypothesis testing.

Real data example

An example from a pharmacological study of patients with functional dyspepsia (FD) and a placebo-controlled trail of subjects with acute migraine is used to illustrate our proposed methodologies. This example has been analyzed by Holtmann et al. [25] and Tang and Tang [14]. In this example, cisapride and simethicone can be regarded as the existing reference and new experimental treatments, respectively. In that study, among $n = 178$ patients of FD, $n_P = 61$, $n_R = 59$ and $n_T = 58$ were randomized and treated in a doubly dummy technique with placebo, cisapride and simethicone, respectively; adverse events (e.g., diarrhea and pain) were happened in $x_P = 7$, $x_R = 10$ and $x_T = 12$ patients treated with placebo, cisapride and simethicone, respectively. It is of interest to test if simethicone is not inferior to cisapride in terms of rate of reporting adverse event in the presence of placebo. Given $\theta = 0.6$ and 0.8, the corresponding p-values for testing $H_0 : (\pi_T - \pi_P)/(\pi_R - \pi_P) \leq \theta$ versus $H_1 : (\pi_T - \pi_P)/(\pi_R - \pi_P) > \theta$ based on the five p-value calculation procedures and three test statistics are reported in Table 2. By Table 2, there is no evidence to show that simethicone is noninferior to cisapride in the presence of placebo at the nominal level $\alpha = 0.05$, which is consistent with that given in Tang and Tang [14].

Discussion

Simulation results demonstrate that our proposed score test statistic outperforms other test statistics in terms of type I error rate and power under our considered settings. The approximate unconditional and bootstrap-resampling methods perform better than other p-value calculation procedures in the sense that their corresponding type I error rates are closer to the prespecified nominal level and their corresponding powers are larger than those of other p-value calculation procedures. The exact unconditional method is conservative and time-consuming when sample sizes are large (e.g., see the 6th column in Table 3). The asymptotic tests are liberal since their type I error rates are greater than the prespecified nominal level $\alpha = 0.05$ in most cases. Comparing the approximate and exact unconditional methods, the approximate unconditional method provides a good alternative to the exact unconditional method in terms of computing time (e.g., see the 6th and 7th columns in Table 3) and type I error rate when sample sizes are large. In contrast, the computing burden of the bootstrap-resampling method is heavier than that of the approximate unconditional method (e.g., see the last two columns in Table 3).

In this article, we concentrate on a three-arm non-inferiority trial with binary endpoints in which the marginal is defined as a fraction of the unknown difference in response probabilities between reference and placebo. The corresponding hypothesis (i.e., $H_0 : \frac{\pi_T - \pi_P}{\pi_R - \pi_P} \leq \theta$ or $H_0 : \pi_T - \theta\pi_R - (1-\theta)\pi_P \leq 0$) is considered since it is simple and only one single hypothesis is involved (e.g., see [6,9,14]). However, three-arm non-inferiority hypotheses with the marginal defined as the prespecified difference between treatments have received a considerable attention in recent years (e.g., see [5,7]). They can be generally classified as the union type hypotheses (i.e., H_{U0}:

Table 3 Computing time (minutes) of the Type I error rates for 11340 configurations of (π_P, π_R, π_T) together with three test statistics under five test methods

λ_P:λ_R:λ_T	θ	n	AM	SAM	EUM	AUM	BTM
1:2:3	0.6	30	3.3	269	2920	55.75	11700
		60	3.8	356	130950	357.3	20700

$\pi_R \geq h_P(\pi_P)$ or $\pi_R \geq h_T(\pi_T)$) or the intersection type hypotheses (i.e., H_{IU0}: $\pi_R \geq h_P(\pi_P)$ and $\pi_R \geq h_T(\pi_T)$), where $h_P(.)$ and $h_T(.)$ are any functions [15]. For specific choices of $h_P(.)$ and $h_R(.)$, this includes, for examples, hypotheses on the differences, the relative risks or the odds ratio of the proportions. While the union type hypotheses are suitable for showing both the superiority of the standard treatment as compared to placebo and the inferiority of the test treatment as compared to the standard treatment, the intersection type hypotheses are suitable for showing the test treatment is as effective as the standard or placebo treatments. We are working on statistical inference on a three-arm non-inferiority trial with the margin being a prespecifided difference between treatments when the primary endpoints are binary.

Conclusions

According to the aforementioned observations, we can draw the following conclusions. In terms of type I error rates and powers, the approximate unconditional and bootstrap-resampling methods with score test statistic are recommended for hypothesis testing purpose when sample sizes are small in a three-arm non-inferiority trial. In terms of time-consuming and type I error rates and powers, the approximate unconditional method with score test statistic behaves the best among our considered p-value calculation procedures and test statistics.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
NST conceived of research questions, developed methods and revised the manuscript; BY carried out statistical analysis and drafted the manuscript; MLT interpreted results and revised the manuscript. All authors commented on successive drafts, and read and approved the final manuscript.

Acknowledgements
This work was supported by the grants from the National Science Foundation of China (11225103), and Research Fund for the Doctoral Program of Higher Education of China (20115301110004). The work of the third author was partially supported by the General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/FDS14/P01/14).

Author details
[1]Department of Statistics, Yunnan University, No.2 Cuihu North Road, 650091 Kunming, China. [2]Department of Mathematics and Statistics, Hang Seng Management College, Hang Seng Link, Siu Lek Yuen, Shatin NT, Hong Kong, China.

Received: 4 August 2014 Accepted: 12 December 2014
Published: 18 December 2014

References

1. Dunnett CW, Gent M: **Significance testing to establish equivalence between treatments with special reference to data in the form of 2 × 2 tables.** *Biometrics* 1977, **33:**593–602.
2. Tango T: **Equivalence test and confidence interval for the difference in proportions for the paired-sample design.** *Stat Med* 1998, **17:**891–908.
3. Tang NS, Tang ML, Chan ISF: **On tests of equivalence via non-unity relative risk for matched-pair design.** *Stat Med* 2003, **22:**1217–1233.
4. Li G, Gao S: **A group sequential type design for three-arm non-inferiority trials with binary endpoints.** *Biom J* 2010, **52:**504–518.
5. Hida E, Tango T: **Three-arm noninferiority trials with a prespecified margin for inference of the difference in the proportions of binary endpoints.** *J Biopharm Stat* 2013, **23:**774–789.
6. Koch GG, Röhmel J: **Hypothesis testing in the gold standard design for proving the efficacy of an experimental treatment relative to placebo and a reference.** *J Biopharm Stat* 2004, **14:**315–325.
7. Hida E, Tango T: **On the three-arm non-inferiority trial including a placebo with a prespecified margin.** *Stat Med* 2011, **30:**224–231.
8. Koch GG, Tangen CM: **Nonparametric analysis of covariance and its role in non-inferiority clinical trials.** *Drug Inf J* 1999, **33:**1145–1159.
9. Pigeot I, Schafer J, Rohmel J, Hauschke D: **Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo.** *Stat Med* 2003, **22:**883–899.
10. Koti KM: **Use of the fieller-hinkley distribution of the ratio of random variables in testing for noninferiority.** *J Biopharm Stat* 2007, **17:**215–228.
11. Hasler M, Vonk R, Hothorn LA: **Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity.** *Stat Med* 2008, **27:**490–503.
12. Ghosh P, Nathoo F, Gönen M, Tiwari RC: **Assessing noninferiority in a three-arm trial using the bayesian approach.** *Stat Med* 2011, **30:**1795–1808.
13. Gamalo MA, Muthukumarana S, Ghosh P, Tiwari RC: **A generalized p-value approach for assessing noninferiority in a three-arm trial.** *Stat Methods Med Res* 2013, **22:**261–277.
14. Tang ML, Tang NS: **Tests of non-inferiority via rate difference for three-arm clinical trials with placebo.** *J Biopharm Stat* 2004, **14:**337–347.
15. Munk A, Mielke M, Skipka G, Freitag G: **Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics.** *Canadiaan J Stat* 2007, **35:**413–431.
16. Liu JT, Tzeng CS, Tsou HH: **Establishing non-inferiority of a new treatment in a three-arm trial: apply a step-down hierarchical model in a papulopustular acne study and an oral prophylactic antibiotics study.** *Intl J Stat Med Res* 2014, **3:**11–20.
17. Jensen J: *Saddlepoint Approximations*. Oxford: Oxford Science Publications; 1995.
18. Tang NS, Tang ML: **Exact unconditional inference for risk ratio in a correlated 2 × 2 table with structural zero.** *Biometrics* 2002, **58:**972–980.
19. Kieser M, Friede T: **Planning and analysis of three-arm non-inferiority trials with binary endpoints.** *Stat Med* 2007, **26:**253–273.
20. Blackwelder WC: **Proving the null hypothesis in clinical trials.** *Control Clin Trials* 1982, **3:**345–353.
21. Farrington CP, Manning G: **Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk.** *Stat Med* 1990, **9:**1447–1454.
22. Jing BY, Robinson J: **Saddlepoint approximations for marginal and conditional probabilities of transformed variables.** *Ann Stat* 1994, **22:**1115–1132.
23. Tang ML, Tang NS, Rosner B: **Statistical inference for correlated data in ophthalmologic studies.** *Stat Med* 2006, **25:**2271–2783.
24. Efron B, Tibshirani RJ: *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall; 1993.
25. Holtmann G, Gschossmann J, Mayr P, Talley NJ: **A randomized placebo-controlled trail of simethicone and cisapride for the treatment of patients with functional dyspepsia.** *Aliment Pharmacol Ther* 2002, **16:**1641–1648.

doi:10.1186/1471-2288-14-134
Cite this article as: Tang *et al.*: Testing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints. *BMC Medical Research Methodology* 2014 **14**:134.