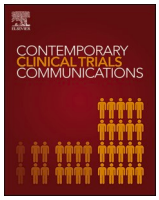
Contemporary Clinical Trials Communications 42 (2024) 101383

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc



Optimization the design of fixed and group sequential three-arm non-inferiority trials with dichotomous endpoints of risk difference and odds ratio

Wenwen Wang [1], Yaru Huang [1], Jielai Xia, Ling Wang, Chen Li [*]

Department of Health Statistics, School of Preventive Medicine, Ministry of Education Key Lab of Hazard Assessment and Control in Special Operational Environment, Fourth Military Medical University, Xi'an, 710042, Shaanxi Province, China

ARTICLE INFO

Keywords:
Risk difference
Three-arm non-inferiority trials
Group sequential clinical trial
Odds ratio
Dichotomous

ABSTRACT

Although the risk difference (RD) is the most common and well explored functional form for testing efficacy with dichotomous endpoint, odds ratio (OR) is also suggested and well applied measure for non-inferiority (NI) trials. Since the construction and interpretation of these function forms are quite different, this study aims to provide detailed discussions and comprehensive comparisons on the design and testing approach for RD and OR scales for the fixed and group sequential three-arm NI trials under various of situations. The sample size determinations and testing approaches for assessing NI of a new treatment in three-arm clinical trials for RD and OR scales were reviewed comprehensively. Simulation studies are conducted for hundreds of scenarios with parameter configurations of the response rates, randomized allocations, NI margins and interim analysis. The operating characteristic (OC) of RD and OR scales based on the MLE and RMLE methods were thoroughly investigated. A trial example was designed and analyzed to demonstrate the methodologies. It is found that sample size determination on OR scale gives smaller sample size and robust procedure compared to RD scale in the majority of situations. When evaluating the behaviors of the attained power, the RMLE methods based on OR scale outperforms the MLE method and tend to have more power to reject the null hypothesis especially under the small sample size situations. Compared to the fixed design, the group sequential design has better OC, which provides a comparable power while needing smaller total average sample sizes for all cases. In addition, we suggest a lower significance level with a higher power for the sample size determination in the superiority test stage in the group sequential design, which can significantly reduce the total sample sizes while the number of subjects in the placebo group does not increase much. It can offer some recommendations for the investigators to choose the optimal endpoints and parameter configurations to design a three-arm NI trial under certain situations.

1. Introduction

When an effective treatment for a disease is already established, newly developed treatment might offer an attractive alternative such as easier intervention or fewer side effects, even if it is not superior in terms of efficacy. The non-inferiority (NI) trial aims to investigate whether the efficacy of a new treatment, relative to the established active standard reference, does not fall below a clinically relevant value by some pre-specified margin [1]. Although two-arm NI trial with a head-to-head comparison of a test and a standard treatment is an attractive option, it is always questioned for lack of assay sensitivity, which refers to the ability of a trial to differentiate an effective treatment from ineffective treatment [2,3]. For this design without a placebo arm, the NI evidence depends on an indirect inference combination of the direct comparison of the test treatment with standard treatment control from this NI trial and the assessment of the standard control's effect vs. placebo from the historical data [2]. The interpretability of the NI trial rests upon the reliability of historical evidence selected (at least, partially subjectively for assessment of the standard control's efficacy, assay sensitivity with historical trials and the noninferiority trial, and applicability of the constancy assumption that the standard control's efficacy in the historical trials does not change in this trial [4]. With an additional placebo group, it is possible to discern situations where the reference treatment is not efficacious in the patient population under investigation and

* Corresponding author.
E-mail address: lc.biosta@qq.com (C. Li).
[1] Wenwen Wang and Yaru Huang Contribute equally.

https://doi.org/10.1016/j.conctc.2024.101383
Received 24 July 2024; Received in revised form 10 October 2024; Accepted 13 October 2024
Available online 18 October 2024

where [3]. This issue is also controversial when defining the NI margin [5].In that case, it is recommended that a placebo be included in addition to the standard reference treatment included for internal validation if ethically acceptable and practically feasible [6]. The three-arm NI trials with both active standard treatment and placebo as control groups, known as "gold-standard" design [3], are therefore recommended to demonstrate the efficacy as well as safety of the test treatment in the guidelines of the ICH E10 and EMEA/CPMP (European Medicines Agency/Committee for Proprietary Medical Products) [7,8].

For the design of this gold-standard trial, Pigeot [9] first proposed the fraction margin approach for normally distribution endpoints, where NI margin is adaptively formulated as the pre-specified fraction of the effect size of the reference treatment over placebo in the three-arm trial. Tang and Tang [10] extended two asymptotic approaches for the three-arm NI testing with dichotomous endpoints via risk difference(RD) based on Wald-type and score test statistics. Mielke proposed Wald-type test procedure for the gold standard design with exponentially distributed censored, time-to-event data [11]. Then, multiple researches have updated and optimized the design and test methods for different types of endpoints. For example, Hasler conducted adjusted methods for the normal distribution endpoints in the presence of heteroscedasticity [12]. Kombrink proposed semiparametric methods for the design and analysis of time-to-event endpoint with Weibull distribution [13]. Wu considered the cohorts accrual [14] and non-compliance situation [15] to optimize the total sample sizes. Recently, the negative binomial [16] and Poisson distributed endpoints [17] for the gold-standard design have also been proposed. The NI margin and the optimization of randomization allocations have also been fully developed and discussed [17–20].

As for the dichotomous endpoint, Kieser and Friede [21] revisited the performance of Tang and Tang's asymptotic test statistics [10] via simulation studies and derived approximate sample size formulae for achieving the desired power. Then, Munk [22] developed likelihood ratio tests. Though RD is the simplest functional form for dichotomous outcomes, other functionals (e.g., risk ratio (RR), odds ratio (OR), number needed to treat (NNT), risk reduction, etc.) are always used in NI trials. Therefore, Chowdhury [23] proposed frequentist test based on traditional fraction margin approach for RR, OR and NNT and developed a conditional testing approach in NI testing. Besides, when it served as an asymptotic normally distributed test statistics, there are two proposed methods for estimation of the variance on the testing procedure for dichotomous endpoints: the straight-forward procedure employing the maximum likelihood estimates (MLE) and the restricted ML estimates (RMLE) [10,24]. Tang [10] and Kieser [24] have assessed the performance of both test procedures via various simulation scenarios. And also, Bayesian approaches have been applied and compared with the frequentist methods for risk difference [20], RR and OR scales [24]. However, it is still lack of unified consensus for the optimal application of the dichotomous endpoint forms or the choice of variance estimation methods in the three-arm NI trial.

On the other hand, group sequential designs enable earlier detection of highly effective or noneffective treatment, which offer an ethical advantage over single-stage designs. In the pioneer work, Schlömer and Brannath [25] proposed approximately optimal rejection boundaries for the group sequential design for three-arm NI trial with sample size allocation ratios. Ochiai [26] proposed a group sequential design for the effect retention approach and compared its operating characteristics to the fixed margin approach. Liu [27] 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. Li [28] used the closed testing principle to establish the hierarchical testing procedure. Meis [29] presented a variation of the hierarchical testing procedure, which allows for the incorporation of binding futility boundaries at interim analyses to optimize the two-stage group sequential three-arm NI trial. However, to the best of our knowledge, there is still no works for the comparisons of different dichotomous endpoint forms with different variance estimation methods in group sequential designs. This motivates us to conduct the comprehensive comparisons between the RD and OR scales based on the MLE and RMLE for the fixed sample and group sequential three-arm NI trial, which may offer recommendations for the researchers to choose the appropriate endpoint forms and estimation methods under certain situations.

The rest of this article is organized as follows. We first review the test statistics and sample size determinations in three-arm NI trials with dichotomous endpoints in the fixed sample and group sequential designs in section 2. Then numerical studies are conducted to investigate the performance of RD and OR scales based on different estimations approaches in terms of power and sample sizes in section 3. An example is used to illustrate the design and test procedures of a three-arm NI trial in clinical practice in section 4. Finally, we conclude with a discussion in section 5.

2. Methods

This study focused on the theoretical research, which is primarily based on theoretical deduction and simulation studies. It did not involve any experiments or trials conducted on human, therefore, ethics approval was not required for this study.

2.1. Fixed NI design and testing procedure

2.1.1. Hypothesis

Let $x_k \sim Bin(n_k, \pi_k)$, $k = T, R, P$, respectively, where π_T, π_R and π_P represent the true efficacy rates given the sample size of n_T, n_R, n_P in the test treatment, reference treatment and placebo groups. Assuming higher response rate indicating better treatment benefit, a general functional form in three-arm NI testing procedures are as follows:

$$\begin{aligned} &H_{0,h(\theta_k)} : h(\theta_T) - h(\theta_P) \le \Delta\cdot(h(\theta_R) - h(\theta_P)\,) \text{ vs.} \\ &H_{1,h(\theta_k)} : h(\theta_T) - h(\theta_P) > \Delta\cdot(h(\theta_R) - h(\theta_P)\,) \end{aligned} \tag{2.1.1}$$

[30].where $\theta_k \in \Theta \subseteq \mathbb{R}$, $k = T, R, P$. $h(\cdot)$ is a monotone, real-valued function on the parameter space Θ representing the treatment efficacy, which is determined by the distribution of the interested endpoint. $\Delta \in [0, \infty)$ is a constant implying the amount of the active control effect relative to placebo which should be retained. In practice, clinical consideration could drive the choice of $\Delta = 1 + f(f \in [-1, 0])$ and common values for f be −10 %, −20 %, etc. To assess the assay sensitivity, equation (2.1.1) firstly validates the standard reference efficacy with the test of $h(\theta_R) - h(\theta_P)$, then conduct the NI test for the test treatment compared to the standard reference. Although the superiority of the reference treatment to placebo would not be a mandatory aim for the interpretation of the three-arm design, if superiority of the test treatment over placebo cannot be shown indirectly, the trial would have inevitably failed. Following this argumentation, the related test procedure starts testing the null hypothesis in (2.1.1) (3). Therefore, rejecting the null hypothesis indicates the test treatment preserves at least $\Delta \times 100$ percent of the treatment effect of the reference relative to placebo. So, it could claim the test treatment are not inferior to the reference standard treatment, both comparing to the placebo.

For dichotomous endpoints, the most common functional form for $h(\theta_k)$ is the effect rates π_k $(k = T, R, P)$, so the general hypothesis of (2.1.1) can be written based on RD scale as

$$\begin{aligned} &H_{0,\pi_k} : \pi_T - \pi_P \le \Delta\cdot(\pi_R - \pi_P) \text{ vs.} \\ &H_{1,\pi_k} : \pi_T - \pi_P > \Delta\cdot(\pi_R - \pi_P) \end{aligned} \tag{2.1.2}$$

([24,30])

Meanwhile, odds ratio of the effect rate is always used in dichotomous endpoints. Its logarithmic transformed $\log(\pi_k/(1-\pi_k))$ is asymptotically normal distribution. Therefore, the hypothesis (2.1.1) can be expressed for the OR scale as

$$H_{0,\log\left(\frac{\pi_k}{1-\pi_k}\right)}: \log\left(\frac{\pi_T}{1-\pi_T}\right) - \log\left(\frac{\pi_R}{1-\pi_R}\right) \le \Delta\left(\log\left(\frac{\pi_R}{1-\pi_R}\right) - \log\left(\frac{\pi_P}{1-\pi_P}\right)\right)$$

vs.

$$H_{1,\log\left(\frac{\pi_k}{1-\pi_k}\right)}: \log\left(\frac{\pi_T}{1-\pi_T}\right) - \log\left(\frac{\pi_R}{1-\pi_R}\right) > \Delta\left(\log\left(\frac{\pi_R}{1-\pi_R}\right) - \log\left(\frac{\pi_P}{1-\pi_P}\right)\right) \quad (2.1.3)$$

And the margin Δ can be transformed on OR scale as $\Delta_{OR} = (\pi_P/(1-\pi_P))/(\pi_R/(1-\pi_R))^{1-\Delta}$ [31].

2.1.2. Test statistic

For the hypothesis (2.1.1) in three arm NI test, let

$$\psi = h(\theta_T) - \Delta h(\theta_R) - (1-\Delta)h(\theta_P) \quad (2.1.4)$$

The MLE of $h(\theta_k) k = T, R, P$ can be obtained by plugging in the MLE $\widehat{\theta}_k$ of θ_k, which is asymptotically normally distributed with variance

$$\sigma^2(\psi) = \sigma_T^2 + \frac{\Delta^2\sigma_R^2}{w_R} + \frac{(1-\Delta)^2\sigma_P^2}{w_P} \quad (2.1.5)$$

where $w_k = \frac{n_k}{n_T}$ denotes the sample size ratio between the $k(= R, P)$ group and the treatment group. Thus, we obtain the following test-statistic

$$T = \sqrt{n_T + n_R + n_P}\cdot\frac{\widehat{\psi}}{\sigma(\widehat{\psi})} = \sqrt{n_T + n_R + n_P}\cdot\frac{h(\widehat{\theta}_T) - \Delta h(\widehat{\theta}_R) - (1-\Delta)h(\widehat{\theta}_P)}{\sigma(\widehat{\psi})} \quad (2.1.6)$$

which is asymptotically standard normally distributed at the boundary of $H_{0,h(\theta_k)}$, i.e., that is $\psi = 0$. Therefore, $H_{0,h(\theta_k)}$ can be rejected if $T > z_{1-\alpha}$, where $z_{1-\alpha}$ is the $1-\alpha$-quantile of the standard normal distribution and α a specified significance level [30].

2.1.3. MLE and RMLE for the variance estimation

For the estimation of the variance $\sigma^2(\psi)$, there are two methods proposed in literature [30]. Typically, it is sufficient to obtain the asymptotic variance of the test statistic in (2.1.6) by the MLE as $\sigma^2(\widehat{\psi}) = \widehat{\sigma}^2_{ML}$, which can be obtained by plugging in the MLE's $\widehat{\theta}_k$ of θ_k for equation (2.1.5) when the variance of the test statistic is a consistent estimator and independent of the parameter $h(\theta_k)$. However, for the binary endpoint, the variance $\sigma^2(\psi)$ depends on the parameter $h(\theta_k)$, $\sigma^2(\psi)$ can be obtained by the restricted estimation under the null hypothesis [30], denoted as $\widehat{\sigma}^2_{RML}$. In this case, $\widehat{\sigma}^2_{RML}$ is a consistent estimator only if the true parameters are located in the hypothesis with $\psi \le 0$. If the parameters are located in the alternative $\psi > 0$, $\widehat{\sigma}^2_{RML}$ should be determined by restricting the likelihood function to the boundary of $H_{0,h(\theta_k)}$ via constraining the equation $\psi = h(\widehat{\theta}_T) - \Delta h(\widehat{\theta}_R) - (1-\Delta)h(\widehat{\theta}_P) = 0$. That is, substitute $\theta_T = h^{-1}(\Delta h(\theta_R) + (1-\Delta)h(\theta_P))$ and maximize this with respect to θ_R and θ_P numerically for the variance estimation in (2.1.5).

Based on the two methods for the variance estimation, there can be four estimation methods of test statistics for the binary endpoint. For the RD scale, $h(\theta_k) = \pi_k$ and $\sigma_k^2 = \pi_k(1-\pi_k)$, $k(= T, R, P)$, the test statistic can be obtained by the MLE given by $\widehat{\psi} = \widehat{\pi}_T - \Delta\widehat{\pi}_R - (1-\Delta)\widehat{\pi}_P$ with $\widehat{\pi}_k = \frac{x_k}{n_k}$ as

$$T_{RDML} = \frac{\sqrt{n}\cdot\widehat{\psi}}{\widehat{\sigma}_{RDML}} = \frac{\widehat{\pi}_T - \Delta\widehat{\pi}_R - (1-\Delta)\widehat{\pi}_P}{\widehat{\pi}_T(1-\widehat{\pi}_T)/n_T + \Delta^2\widehat{\pi}_R(1-\widehat{\pi}_R)/n_R + (1-\Delta)^2\widehat{\pi}_P(1-\widehat{\pi}_P)/n_P} \quad (2.1.7)$$

where $N = n_T + n_R + n_P$. The RMLE method constrained the equation $\pi_T - \Delta\pi_R - (1-\Delta)\pi_P = 0$ for $\pi_k(k = T, R, P)$ in the variance $\sigma^2(\psi)$ estimation and the corresponding test statistic.

For the OR scale, $h(\theta_k) = \log(\pi_k/(1-\pi_k))$, $\sigma_k^2 = (\pi_k(1-\pi_k))^{-1}$ [30,32], the MLE test statistics will be

$$T_{ORML} = \frac{\sqrt{n}\cdot\widehat{\psi}}{\widehat{\sigma}_{ORML}} = \frac{log\left(\frac{\widehat{\pi}_T}{1-\widehat{\pi}_T}\right) - \Delta log\left(\frac{\widehat{\pi}_R}{1-\widehat{\pi}_R}\right) - (1-\Delta)log\left(\frac{\widehat{\pi}_P}{1-\widehat{\pi}_P}\right)}{\frac{1}{(\widehat{\pi}_T(1-\widehat{\pi}_T))n_T} + \frac{\Delta^2}{(\widehat{\pi}_R(1-\widehat{\pi}_R))n_R} + \frac{(1-\Delta)^2}{(\widehat{\pi}_P(1-\widehat{\pi}_P))n_P}} \quad (2.1.8)$$

And T_{ORRML} can be obtained as

$$T_{ORML} = \frac{\sqrt{N}\cdot\widehat{\psi}}{\widehat{\sigma}_{ORRML}} \quad (2.1.9)$$

by restricting $\log\left(\frac{\pi_T}{1-\pi_T}\right) - \Delta\log\left(\frac{\pi_R}{1-\pi_R}\right) - (1-\Delta)\log\left(\frac{\pi_P}{1-\pi_P}\right) = 0$ for $\sigma^2(\psi)$ estimation.

2.1.4. Sample size determination and the optimal allocation

Use the notations $\psi_0 = h(\theta_{T,0}) - \Delta h(\theta_{R,0}) - (1-\Delta)h(\theta_{P,0})$ for the alternative to be detected with the true unknown parameter $h(\theta_{k,0})$, $k = T, R, P$ and the variance σ_0^2. For the unrestricted MLE methods, the power approximation to reject the hypothesis $H_{0,h(\theta_k)}$ will be

$$P_{\psi_1}(T > 1-\alpha) \approx 1 - \Phi\left(z_{1-\alpha} - \sqrt{n}\ \frac{\psi_0}{\sigma_0}\right) \quad (2.1.10)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. When the variance $\sigma^2(\psi)$ is estimated unrestrictedly ($\sigma^2(\widehat{\psi}) = \widehat{\sigma}^2_{ML}$) we end up with the simplified power formula. The minimal required total sample size to obtain a power of $1-\beta$ for a given alternative $\psi_0 > 0$ is determined by

$$n_{ML} \approx (z_{1-\alpha} + z_{1-\beta})^2\cdot\left(\frac{\sigma_0}{\psi_0}\right)^2 \quad (2.1.11)$$

For the RMLE method, σ^2_{RML} denotes the same expression as for σ_0^2 but with $h(\theta_{k,0})(k = T, R, P)$ replaced by the values $h(\theta_{k,RML})(k = T, R, P)$ fulfilling the restriction $h(\theta_{T,RML}) - \Delta h(\theta_{R,RML}) - (1-\Delta)h(\theta_{P,RML}) = 0$. The power function will be obtained by substituting σ_0^2 with σ^2_{RML}

$$P_{\psi_1}(T > 1-\alpha) = P_{\psi_1}\left(T\cdot\frac{\sigma_{RML}}{\sigma_0} - \sqrt{n}\ \frac{\psi_0}{\sigma_0} > z_{1-\alpha}\cdot\frac{\sigma_{RML}}{\sigma_0} - \sqrt{n}\ \frac{\psi_0}{\sigma_0}\right) \approx 1 - \Phi\left(z_{1-\alpha}\frac{\sigma_{RML}}{\sigma_0} - \sqrt{n}\ \frac{\psi_0}{\sigma_0}\right) \quad (2.1.12)$$

The sample size formula is derived from (2.13) as

$$n_{RML} \approx \left(z_{1-\alpha}\cdot\frac{\sigma_{RML}}{\sigma_0} + z_{1-\beta}\right)^2\cdot\left(\frac{\sigma_0}{\psi_0}\right)^2 \quad (2.1.13)$$

For the optimal allocation of the samples, it can be considered from two perspectives [1]: minimizing the total sample size at a given power [2]; maximizing the power for a total sample size, which also depends on the allocation through σ_0^2 and σ^2_{RML}. When the variance σ_0^2 is estimated unrestricted in the test procedure ($\sigma^2(\widehat{\psi}) = \widehat{\sigma}^2_{ML}$) we only need to

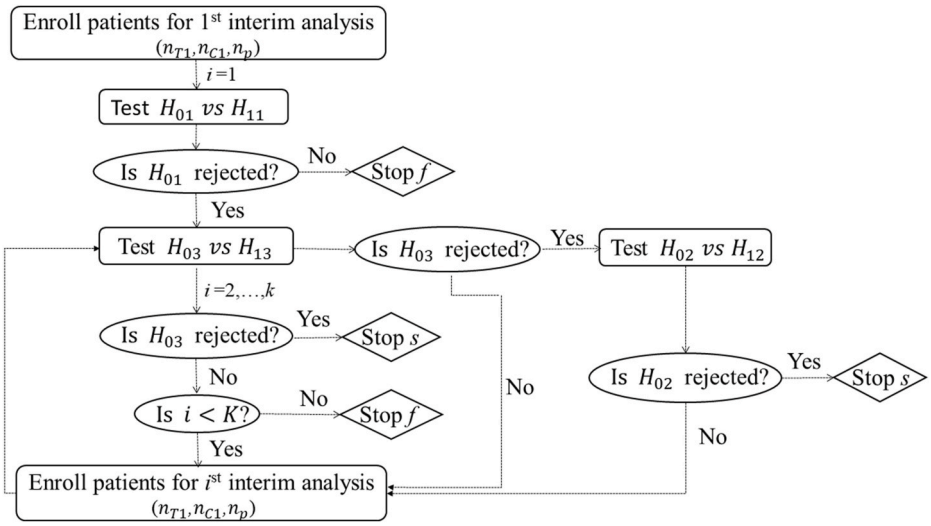


Fig. 1. Flow chart of the test procedures in the three-arm NI group sequential design [28].

minimize σ_0^2 to maximize the power, conferring (2.1.10). Let the randomized ratio $\omega_T : \omega_R : \omega_P = 1 : n_R/n_T : n_P/n_T$, σ_0^2 can be minimum when the allocation ratio $\omega_T : \omega_R : \omega_P = 1 : \Delta\frac{\sigma_{R,0}}{\sigma_{T,0}} : |1-\Delta|\frac{\sigma_{P,0}}{\sigma_{T,0}}$, where $\sigma_{k,0}$ $(k = T, R, P)$ is the standard deviation of each group under the alternative hypothesis. Since the choice of ω_R and ω_P were under the control of the investigator, we obtain the optimal allocation $1 : \Delta : |1-\Delta|$ [9,30,33], that is, the sample size will be minimum under this circumstance. When the variance σ_0^2 is estimated under restriction to $h(\theta_{T,RML}) - \Delta h(\theta_{R,RML}) - (1-\Delta)h(\theta_{P,RML}) = 0$, the asymptotic power in (2.1.12) depends additionally on σ_{RML}/σ_0. Previous studies verify that the asymptotically optimal allocation derived for the unrestricted case is also the (asymptotically) optimal allocation in terms of maximizing the power in (2.1.13) [24,30].

2.2. Group sequential design

2.2.1. Design

As we demonstrated before, there are two major objectives for the three-arm NI trials with a natural hierarchical testing procedure as follow:

$$H_{01} : h(\theta_T) \le h(\theta_P) \text{ vs. } H_{11} : h(\theta_T) > h(\theta_P) \tag{2.2.1}$$

$$H_{02} : h(\theta_R) \le h(\theta_P) \text{ vs. } H_{12} : h(\theta_R) > h(\theta_P) \tag{2.2.2}$$

Firstly, superiority testing is a prerequisite to verify the effectiveness of the test and the reference treatments as (2.2.1) and (2.2.2). Then, the NI objective can be assessed by an effect preservation margin as

$$\text{vs.} \begin{array}{l} H_{03} : h(\theta_T) - h(\theta_R) \le \Delta\cdot(h(\theta_R) - h(\theta_P)) \\ H_{13} : h(\theta_T) - h(\theta_R) > \Delta\cdot(h(\theta_R) - h(\theta_P)) \end{array} \tag{2.2.3}$$

In a fixed design, the sample size is determined upfront and the testing procedure is conducted after enrollments completed. If the null hypothesis of (2.2.1) is not rejected, the trial should be stopped since the test treatment shows no improvement over the placebo. But the data from the reference treatment group might be used just for the assessment of the trial assay sensitivity, in which a waste of resource is unavoidable. This is problematic especially when the response rate of the placebo is highly variable or unknown. From another point of view, since the sample size for the superiority test is generally smaller than that of the NI test, the sample size required for a fixed sample design is usually driven by the NI test and may lead to a higher sample size than actually needed [28]. In this case, a group sequential design may be more preferred than the fixed procedure. With pre-specified stopping boundaries, the trial has good chances of early stopping for futility or efficacy.

The group sequential design can be depicted as two stages (Fig. 1). First, assuming the allocation ratios of $1 : \omega_T : \omega_P$ for the test group, standard treatment group and placebo group, the sample size can be determined based on the superiority test in (2.2.1) with a one-sided type-I error of α_{sup}, a power of $1-\beta_{\text{sup}}$. When the first stage completed, interim analysis can be conducted and stopped early for futility if the superiority of the test treatment over placebo (H_{11} in (2.2.1)) is not demonstrated; otherwise, we continue to enroll patients and perform the second stage of the NI test as (2.2.3).

In the second stage, the enrollment is conducted just for the test and reference groups with an allocation ratio of $1 : \omega_R$. This procedure could apply a conventional group sequential design with K interim analysis with stop boundaries to maintain the overall type I error. In $i^{st}(i = 1, \ldots\ldots, K)$ interim analysis, subjects from the placebo group in the first stage is used to construct a three-arm structure with the samples of the test and reference groups for the three-arm NI test as (2.2.3). If H_{03} in (2.2.3) is rejected, there is still debates that whether H_{02} must also be rejected to make the trial successful [34]. In Li's research [21,28,35], the aim the testing of H_{02} is the trial assay sensitivity. Although the assay sensitivity has been demonstrated following the rejection of H_{01} in (2.2.1) in the first stage, Li suggested to perform the superiority test H_{02} in (2.2.2) at the first time when the null hypothesis H_{03} in (2.2.3) is rejected. If H_{02} is rejected then the rejection of H_{03} can be interpreted as an effect preservation statement, otherwise the NI statement of (2.2.3) can not be claimed because the standard treatment is not demonstrated in the same trial and it needs to continue the enrollment and test H_{03} again for the $i+1^{st}$ phase.

2.2.2. Test statistic and sample size determinations

In the first stage, the superiority hypothesis in (2.2.1) can be tested by the traditional superiority tests. For the second stage, the testing is quite similar to the previous fixed three-arm NI test for each interim analysis. Specifically, let $n_{ik}, i = T, R$ be the sample sizes for the test group and the reference group and $Z_k^{(0)}$ be the test statistic based on the data accumulated at the i^{st} interim analysis. With n_P from placebo group in the first stage, there is:

$$Z_k^{(0)} = \frac{\widehat{\psi}_k}{\sigma(\widehat{\psi}_k)} \tag{2.2.4}$$

where $\widehat{\psi}_k = h(\widehat{\theta}_{Tk}) - \Delta h(\widehat{\theta}_{Rk}) - (1-\Delta)h(\widehat{\theta}_P)$ and $\sigma^2(\widehat{\psi}_k) = \sigma_{Tk}^2 + \frac{\Delta^2\sigma_{Rk}^2}{w_{Rk}} +$

Table 1
Sample size estimation based on different randomized allocation: $\Delta = 0.5$, $\alpha = 2.5\%$, $\beta = 20\%$

(π_T,π_R,π_P)	(ω_R,ω_P)	n_{RDML}	n_{RDRML}	Δ^{OR}	n_{ORML}	n_{ORRML}
(0.3,0.3,0.1)	(1,1)	672	615	0.5092	465	477
	(1,0.5)	605	588	0.5092	508	473
	(0.5,0.5)	567	567	0.5092	449	420
	optimal	552	577	0.5092	435	425
(0.5,0.5,0.1)	(1,1)	198	189	0.3333	102	90
	(1,0.5)	175	183	0.3333	130	97
	(0.5,0.5)	164	168	0.3333	108	81
	optimal	158	179	0.3333	88	80
(0.7,0.7,0.1)	(1,1)	75	87	0.2182	48	36
	(1,0.5)	68	85	0.2182	65	40
	(0.5,0.5)	64	75	0.2182	52	33
	optimal	62	80	0.2182	39	38
(0.5,0.5,0.3)	(1,1)	861	855	0.6547	1728	1734
	(1,0.5)	820	833	0. 6547	1570	1567
	(0.5,0.5)	753	763	0. 6547	1572	1572
	optimal	753	763	0. 6547	1514	1508
(0.7,0.7,0.3)	(1,1)	186	207	0.4286	180	168
	(1,0.5)	180	203	0.4286	185	170
	(0.5,0.5)	164	179	0.4286	161	156
	optimal	164	179	0.4286	160	157
(0.9,0.9,0.3)	(1,1)	45	78	0.2182	60	42
	(1,0.5)	48	75	0.2182	60	45
	(0.5,0.5)	41	63	0.2182	49	41
	optimal	41	64	0.2182	47	254
(0.7,0.7,0.5)	(1,1)	765	813	0.6547	2004	1965
	(1,0.5)	760	790	0.6547	1778	1750
	(0.5,0.5)	692	703	0.6547	1796	1812
	optimal	690	703	0.6547	1697	1693
(0.9,0.9,0.5)	(1,1)	105	156	0.3333	156	120
	(1,0.5)	117	148	0.3333	145	120
	(0.5,0.5)	104	123	0.3333	127	121
	optimal	98	129	0.3333	123	147

$\frac{(1-\Delta)^2\sigma_P^2}{w_P}$. Under the null hypothesis of H_{03}, the test statistics $Z_1^{(0)},\ldots\ldots,Z_K^{(0)}$ follow an asymptotic multivariate normal distribution with $E(Z_k^{(0)})=0$ and $cov\left(Z_i^{(0)},Z_j^{(0)}\right)=\sigma(\widehat{\psi}_j)/\sigma(\widehat{\psi}_i), 1\le i\le j\le K$. Similar to the fixed sample design, we can use the RMLE method to estimate the test statistics $Z_k^{(0)}$ in the i^{st} interim analysis.

Since multiple testing was involved in the sequential testing, the stopping boundaries were used in order to maintain the overall type-I error rate. The trial could be designed for early stopping for either efficacy or futility or both and commonly used methods, like Pocock [36] or O'Brien and Fleming [37], can be used for calculating the critical values $\{b_1,\ldots\ldots,b_k\}$.

In literature, Li [28] have proposed another test procedure. Let $Z_k^{(1)}$ be the test statistics under the alternative hypothesis H_{13} and

$$Z_k^{(1)}=\frac{\widehat{\psi}_k-\psi_1}{\sigma(\widehat{\psi}_k)} \tag{2.2.5}$$

where $\widehat{\psi}^{(1)}=h(\widehat{\theta}_T)^{(1)}-\Delta h(\widehat{\theta}_R)^{(1)}-(1-\Delta)h(\widehat{\theta}_P)^{(1)}$. The distribution of statistics $Z_1^{(1)},\ldots\ldots,Z_K^{(1)}$ is similar to the $Z_1^{(0)},\ldots\ldots,Z_K^{(0)}$, and the power of the NI test at $\psi=\psi^{(1)}$ is

$$prob\left\{\bigcup_{k=1}^{K}\left(Z_j^{(1)}\right)<b_j \text{ for } j=1,\ldots,k-1 \text{ and } Z_k^{(1)}\ge b_k\right)\right\} \tag{2.2.6}$$

For the sample size estimation, there is no explicit formula, but it can be found such that power in (2.2.6) is equal to the target power $1-\beta$ under the assumed response rates at the alternative hypothesis of (2.2.3). The test response rates in the testing procedures can be estimated by the RMLE method (method 1) or the assumed response rates at the alternative hypothesis (method 2). We denote the corresponding sample sizes by N_{00} (method 1) and N_{01} (method 2) respectively [28]. Let δ_{nT} be the sample size increment per one additional interim analysis in the test group for the NI test stages. We can increase δ_{nT} for the numerical search of the sample size until the resulted power achieves the target level.

2.3. Code Availability

All the analyses were performed using the "ThreeArmedTrials" (version 1.0–4) and "mnormt" (version 1.5–6) Packages in the RStudio software (version 4.2.2). The code for the mathematical algorithm and simulation study are available from the corresponding author on reasonable request.

3. Numerical study

3.1. Fixed design

In the following, numerical study is conducted to investigate the operating characteristics (OC) of the three-arm NI trial of RD and OR scales given various design parameters with type I error control of $\alpha = 0.025$ for one-sided and power of $1-\beta = 0.80$.

Firstly, we explore the sample sizes related to different response rates and randomized allocations based on RD and OR scale with MLE or RMLE methods. The response rates were given $\pi_P = 0.1, 0.3, 0.5$, $\pi_R = \pi_P + 0.2$, $\pi_P + 0.4$, $\pi_P + 0.6 (0 \le \pi_R \le 1)$. Since the design operating characteristics (OC) have been fully evaluated for the correlations of π_T, π_R,π_P, especially when $\pi_T > \pi_R$, the power for proving NI is adequate for assay sensitivity regardless of NI margin [10,24,38]. We consider the null situation of $\pi_T = \pi_P$ and alternative as $\pi_T = \pi_R$, in which the better efficacy scenarios ($\pi_T > \pi_R$) not detected in this study. The randomized allocations $\omega_T : \omega_R : \omega_P$ were considered to be balance design as 1:1:1, two types of unbalance designs as 1:1:0.5 and 1:0.5:0.5 and the optimal allocation. The NI margin of $\Delta = (1+f)(\pi_R - \pi_P)$ was considered as 0.5 for the RD endpoint and the log transformation of $\Delta_{OR} = (\pi_P/(1-\pi_P))/(\pi_R/(1-\pi_R))^{1-\Delta}$ for the OR endpoint [31].

Table 1 showed the estimated sample sizes corresponding to a total of 32 parameter configurations. As expected, we can see that the sample size increases quickly with the growing π_P. Taking $\pi_T = \pi_R = 0.7$ for an example, the sample size increases by 3–4 times when π_P increases a unit of 0.2. Meanwhile, if the reference effectiveness $\pi_R - \pi_P$ was doubled, the estimated sample size shrinks by almost 67 % for the RD endpoint, and more reductions for the OR endpoint can be seen. For the randomized allocation, we found that sample sizes obtained from the optimal allocation design were always smaller than those obtained from the other sample size allocation ratios, though the sample sizes with allocation of 1:0.5:0.5 were very close or equal to those in the optimal allocation. That phenomenon is intuitively true since the optimal allocation ratio would be $1:\Delta:|1-\Delta| = 1:0.5:0.5$ employing $\Delta = 0.5$. The necessary sample size is remarkably smaller for the unbalance allocation (1:0.5:0.5) as compared to balance design (1:1:1) and a minor augment is obtained for the unbalance allocation (1:1:0.5). These results are quite similar to those pointed out by Koch [34] and Pigeot [9]. Thus, one may be suggested not to apply a balance design for the three-arm NI trial and the subjects for the placebo should be small for ethical reasons.

For evaluating different estimation methods, it can be found that the RMLE method based on OR scale shows generally better performance than adoption of the MLE method in case of small sample size, while the conclusion is reversed in terms of the RD scale. For example, results with the OR endpoint show a further 44 % sample size reduction compared to the RD scale with the optimal allocation of $(\pi_T,\pi_R,\pi_P) = (0.5,0.5,0.1)$. If we check a smaller reference efficacy rate, such as $(\pi_T,\pi_R,\pi_P) = (0.5,0.5,0.3)$, OR scale results in a considerable increase in sample size. On the one hand, the sample size required more for the small test effectiveness as granted. The OR scale gives smaller sample size when Δ^{OR} smaller or closes to 0.5 in most cases. But when Δ^{OR} increases more than 0.5, the sample size based on OR scale augments rapidly compared to the RD

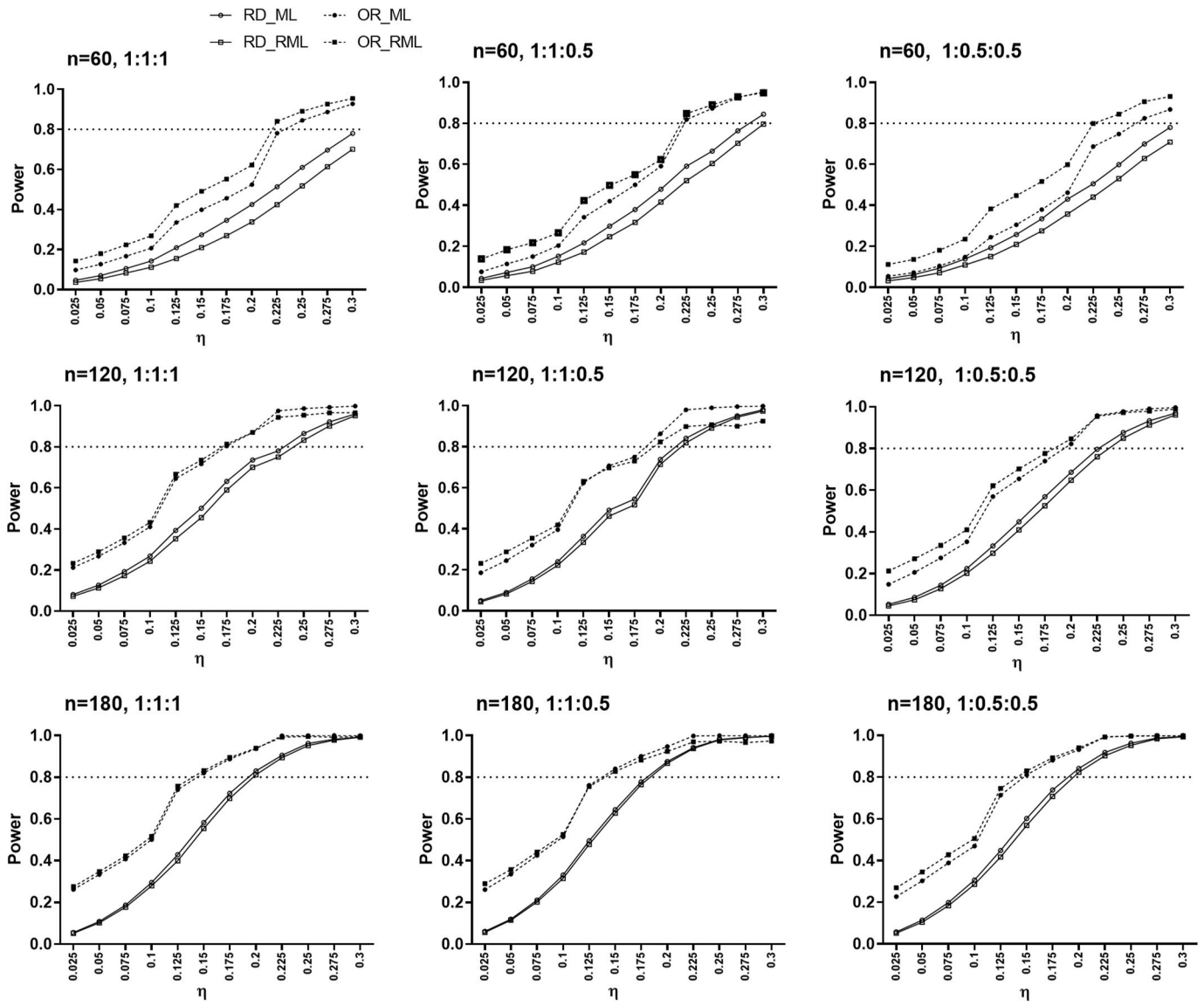


Fig. 2. The simulated power based on four test statistics given $\Delta = 0.5$

scale. We also explored all the considered parameter configurations with the values of $\Delta = 0.5, 0.6$, $0.7, 0.8$, the estimated sample sizes based on higher NI margins tend to be larger with approximately 20 % increase when Δ changes from 0.5 to 0.8, while the variation based on RD endpoint seems more significant than that on OR endpoint. On the whole, we consider that sample size determination on OR scale is a more efficient and robust procedure compared to RD scale.

In addition, we investigated the behaviors of the proposed test procedures in terms of attained power at the nominal level $\alpha = 0.025$ (one-sided). The simulated power of parameter configurations under the small sample size $n = 60, 120, 180$ was detected with the response rates in test group $\pi_T = \Delta\pi_R + (1-\Delta)\pi_P + \eta$ ($\eta = 0.025, 0.05, \ldots, 0.30$), $\pi_R = \pi_P + 0.30$. A total of 480 parameter configurations were simulated with 5000 replications for each scenario.

Fig. 2 displays the simulated average powers based on MLE and RMLE methods under RD and OR scales with line plots by the changing η. The average powers on RD scale are roughly doubled when n increases from 60 to 120, whereas the rising curves trends to be flat when n increases from 120 to 180. In contrast, the power augments from OR scale is more conservative, which increases 10 % on average for every 60 samples increased. It can be seen from Fig. 2 that allocating less subjects to placebo group as compared to the test group and the reference group is advantage in terms of the simulated powers. The average powers are 0.6048 for allocation of 1:1:1, 0.5769 for 1:1:0.5 and 0.5912 for 1:0.5:0.5 in $n = 120$, 0.6624 for 1:1:1,0.6628 for 1:1:0.5 and 0.6797 for 1:0.5:0.5 in $n = 180$, respectively. Almost all parameter configurations based on RD scale could not obtain the desired power for $n = 60$. While, the simulated powers based on OR scale are beyond 80 % when $\eta \geq 0.225$ for $n = 60$ (with a maximum power of 0.8401 for $\eta = 0.225$ in the allocation of 1:0.5:0.5). Although slight deviations of the actual powers may exit, RMLE method outperforms the MLE method, in which the average power based on T_{ORRML} is 56.34 % of configurations. The power curves of test method of T_{RDRML} obtained the lowest powers almost in all parameter constellations (33.65 % for average power), though it is slightly different from that of T_{RDML}. The proportions of configurations whose simulated power lie in the interval (0.80,1.0) for test statistics T_{RDML}, T_{RDRML}, T_{ORML} and T_{ORRML} are 0.0779, 0.0519, 0.2727 and 0.3117 for $n = 60$, 0.1613, 0.1129, 0.4064 and 0.4387 for $n = 120$, and 0.3167, 0.2933, 0.5122 and 0.5329 for $n = 180$, respectively. These results also indicate T_{ORRML} works satisfactorily in the majority of situations. Hence, we would recommend the RMLE methods based on OR scale for the hypothesis testing in view of all the above result.

Table 2
Sample size estimation in three-arm NI group sequential design for K = 3 (the first line) K = 4 (the second line) based on method 1 (N_{00}) and method 2 (N_{01}) at $\alpha_{sup} = 2.5\%$, $\beta_{sup} = 20\%$, $\alpha = 2.5\%$, $\beta = 20\%$

(π_T, π_R, π_P)	Δ	(ω_R, ω_P)	n_P	N_{00}		N_{01}	
				Max	Mean	Max	Mean
(0.7,0.7,0.5)	0.5	(1,1.5)	78	865	736	857	731
				873	708	867	704
		(1,1)	93	979	818	983	821
				987	783	993	787
		(1,0.5)	69	1341	1089	1389	1123
				1353	1033	1407	1069
		(0.5,0.5)	69	1191	970	1239	1004
				1198	920	1248	952
(0.7,0.7,0.3)	0.5	(1:1.5)	20	226	192	214	183
						220	179
		(1,1)	24	256	213	240	202
				258	204	246	197
		(1,0.5)	18	352	285	324	265
				358	272	328	253
		(0.5,0.5)	18	308	250	287	235
				309	237	287	222
(0.9,0.9,0.3)	0.8	(1:1.5)	8	356	305	332	290
				358	291	340	280
		(1,1)	10	398	336	374	321
				402	322	378	308
		(1,0.5)	7	529	440	521	436
				537	422	537	421
		(0.5,0.5)	7	486	403	492	408
				490	384	499	391
(0.9,0.9,0.1)	0.8	(1.5)	4	234	198	186	165
				236	185	188	159
		(1,1)	5	271	226	207	183
				273	216	213	178
		(1,0.5)	4	338	279	246	217
				342	267	252	210
		(0.5,0.5)	4	303	249	228	198
				307	239	231	189

3.2. Group sequential design

We used similar parameter configurations as the previous fixed sample design to explore the OC of the group sequential design. The design elements focused in this section are the group allocations, the number of interim looks and the settings of type I and type II error in the first stage of the superiority test. Therefore, we have conducted the simulation only on the RD scale for better intelligibility and clarity of the magnitude of the treatment effect at each stage of the group sequential trial.

Firstly, we examined required total samples size for the group sequential design. The response rates (π_T, π_R, π_P) was set as (0.7,0.7,0.5), (0.7,0.7,0.3), (0.9,0.9,0.3), (0.9,0.9,0.1). The total overall one-sided type-I error α and power were set as 2.5 % and 80 %, and the superiority test of the test over placebo as 2.5 % and 80 %. We start with $\delta_{nT} = 2$ and increase δ_{nT} by 1 in each step to obtain the statistical power until the resulted power achieves the target level. O'Brien and Fleming methods was used for the stopping boundaries to maintain the overall type-I error rate in the group sequential testing. Table 2 illustrates the sample size determinations among various interim analyses. Generally, increasing interim number of looks K from 3 to 4 results in the sample sizes increase around 10–20 for all scenarios. For the two sample size estimation methods, the difference is tiny though method 2 needs lesser sample sizes than those of the method 1 in general. Besides, the balance design determines less sample size than the unbalance design, while subjects in the placebo group are a little more than the unbalance allocation. This phenomenon is contrast to that in the fixed sample design. Since the number of subjects in the placebo group is determined by the superiority test of (2.2.1) in the first stage, a balance design would require more sample sizes for the placebo group but subsequently less sample size for the NI test. In this way, the total sample sizes were significantly reduced while the subjects in the placebo group do not increase much. We even adjusted the allocation of (ω_R, ω_P) as (1:1.5), the corresponding sample size in the first stage was still less than those of the allocation (1:1). A properly higher allocation of placebo group in the first stage will help reducing the total sample size.

We further explored the parameter configurations for the superiority stage design at a different significance level or a different power level. Here, we focused on a small rate difference (0.7,0.7,0.5) with a lower NI margin 0.5 and a large difference (0.9,0.9,0.3) with a higher margin 0.8. Setting one-sided significant level ($\alpha_{sup} = 2.5\%, 1\%$) and type II error rate ($\beta_{sup} = 20\%, 5\%$) in the superiority stage design, we detected the

Table 3
Sample size estimation in three-arm NI group sequential design when the sample size in the first stage is estimated under the specified α_{sup} (%) and β_{sup} (%) based on K = 3, $\alpha = 2.5\%$, $\beta = 20\%$.

(π_T, π_R, π_P)	Δ	(ω_R, ω_P)	α_{sup}	β_{sup}	n_P	N_{00}		N_{01}	
						Max	Mean	Max	Mean
(0.7,0.7,0.5)	0.5	(1,1)	2.5	20	93	979	818	983	821
			2.5	5	153	811	674	803	667
			1	20	120	860	720	852	714
			1	5	186	798	673	790	668
		(1,0.5)	2.5	20	69	1341	1089	1389	1123
			2.5	5	115	881	716	877	713
			1	20	88	1028	826	1032	829
			1	5	139	829	735	821	734
		(0.5,0.5)	2.5	20	69	1191	970	1239	1004
			2.5	5	115	780	633	780	632
			1	20	88	910	743	919	750
			1	5	139	738	626	738	627
(0.9,0.9,0.3)	0.8	(1,1)	2.5	20	10	398	336	374	321
			2.5	1	20	300	257	280	245
			0.1	20	19	301	258	281	246
			0.1	1	33	283	243	267	233
		(1,0.5)	2.5	20	7	529	439	521	436
			2.5	1	15	325	274	305	261
			0.1	20	13	345	290	321	274
			0.1	1	24	292	244	276	232
		(0.5,0.5)	2.5	20	7	486	403	492	408
			2.5	1	15	278	235	272	231
			0.1	20	13	298	250	292	247
			0.1	1	24	246	207	243	205

Table 4
Sample sizes and simulated powers between the group sequential design and the fixed design when there are uncertainties on the placebo response rate: $\alpha_{sup} = 2.5\%$, K = 3, $\alpha = 0.025$, $\beta = 0.2$.

(π_T, π_R, π_P)	Δ	β_{sup}	ε[a]	(ω_R, ω_P)	Group sequential design			Fixed sample design		
					n_P	$E(N_{01})$	Power (%)	n_P	N_{01}	Power (%)
(0.8,0.8,0.6)	0.5	5 %	1.0	(1,1)	134	559	80.18	221	663	80.54
				(1,0.5)	100	598	81.06	131	655	80.27
				(0.5,0.5)	100	519	79.90	144	576	79.76
			0.9	(1,1)	134	497	94.91	221	663	95.26
				(1,0.5)	100	543	95.21	131	655	95.13
				(0.5,0.5)	100	460	95.26	144	576	95.06
			0.8	(1,1)	134	447	99.35	221	663	99.34
				(1,0.5)	100	512	99.30	131	655	99.24
				(0.5,0.5)	100	422	99.15	144	576	99.26
			0.7	(1,1)	134	419	99.97	221	663	99.96
				(1,0.5)	100	501	99.96	131	655	99.94
				(0.5,0.5)	100	405	99.95	144	576	99.98
(0.8,0.8,0.2)	0.8	1 %	1.0	(1,1)	21	355	80.73	157	471	80.08
				(1,0.5)	15	387	80.24	80	400	80.42
				(0.5,0.5)	15	373	80.86	100	400	80.54
			1.1	(1,1)	21	361	78.02	157	471	78.14
				(1,0.5)	15	392	77.84	80	400	77.82
				(0.5,0.5)	15	378	77.55	100	400	78.53
			1.2	(1,1)	21	365	75.01	157	471	74.52
				(1,0.5)	15	398	73.79	80	400	74.41
				(0.5,0.5)	15	383	74.76	100	400	74.90
			1.3	(1,1)	21	370	70.66	157	471	72.08
				(1,0.5)	15	403	70.81	80	400	71.00
				(0.5,0.5)	15	388	70.41	100	400	71.52

[a] Let π_P and $\widetilde{\pi}_P$ be the assumed and true placebo response rates respectively, and $\varepsilon = \widetilde{\pi}_P/\pi_P$

total sample size determination in the group sequential design. As shown in Table 3, it confirms that the less the treatment difference tested in the first superiority test stage, the more the total sample size requires. In the allocation comparisons, the allocation (1:1:0.5) determines the highest size while balance design determines the lowest. For example, for $(\pi_T, \pi_R, \pi_P) = (0.9, 0.9, 0.3)$ and $\Delta = 0.8$ when $\alpha_{\text{sup}} = 0.025$ and $\beta_{\text{sup}} = 0.20$, the balance design takes three more subjects in placebo group but saves 134 subjects in the NI sequential procedure compared to allocation (1:1:0.5). Thus, increasing the weight of placebo ω_P in (ω_R, ω_P) will properly help reducing the total sample size. We also found that the more placebo subjects at a lower significant level with higher power, the less total sample size determined. For example, for the combinations of response rates (0.9,0.9,0.3), allocation (1,0.5,0.5) and NI margin 0.8, shrinking α_{sup} from 2.5 % to 1 % and β_{sup} from 20 % to 1 % results in a total sample size reduction by half but an augment of 17 subjects in the placebo group. In this way, we may be able to choose a lower significance level with a higher power for the sample size estimation in the superiority test stage.

On the other hand, there exist situations when placebo response rate is either underestimated or overestimated since the placebo response rate is assumed with large uncertainty. We simulated some misestimated situations to make comparison between the group sequential and fixed sample designs to detect the advantage of group sequential design. Let π_P and $\widetilde{\pi}_P$ be the assumed and true placebo response rate respectively and $\varepsilon = \widetilde{\pi}_P/\pi_P$. In the first situation, we make a conservative assumption on the response rates (π_T, π_R, π_P) of (0.8,0.8,0.6) with the NI margin of 0.5, which is assumed to be minimally clinical important. The power of the superiority test stage was set as 95 % with type-I-error as 2.5 %, but the placebo rate is overestimated with $\varepsilon = 0.9, 0.8, 0.7$. In the second situation, we assume a large effect difference for the response rates (π_T, π_R, π_P) of (0.8,0.8,0.2) with the NI margin as 0.8, so the power of the superiority test stage was set as 99 % but with an underestimated placebo effect of $\varepsilon = 1.1, 1.2, 1.3$. Taken $\varepsilon = 1$ as a reference, simulations are conducted for each scenario with 5000 replications to detect the average total sample size and testing power as shown in Table 4.

Compared to the fixed sample design, the group sequential design reduces the total sample sizes by 10%–30 % on average, but maintains comparable powers for the effect preservation test either the placebo effect is under- or over-estimated. Since the sample size for the placebo arm is determined by the superiority test stage, the group sequential design could save more subjects in placebo arm in routine practice. For example, in the treatment combinations of (0.8,0.8,0.2), if the placebo response rate is underestimated with $\varepsilon = 1.1$, the placebo arm requires only 15 subjects for the allocation of (1,0.5,0.5), and the trial can be early stopped via the interim analysis with a total sample size of 378 at most. However, the fixed sample design needs 400 subjects, of which 100 subjects in placebo arm. The trial can be stopped early for either futility or efficacy to save more subjects than that of the fixed sample design.

4. Application

In this section, we use a hypothetical example to illustrate the design and test procedures of a three-arm NI trial for a telemedicine study.

Gout is a form of inflammatory arthritis with elevated levels of uric acid in patients' blood, which can be lowered via lifestyle changes and medicine to provide long-term prevention [39]. Since gout is a chronic metabolic disease, telemedicine may have potential use for guidance of treatment and follow-up to lower the uric acid level and prevent gout attacks. A multicenter study is assumed to be conducted and use a three-arm randomized trial to investigate the efficacy of the telemedicine. The patients with the episodes history of acute gout arthritis by the American college of rheumatology (ACR) diagnostic criteria with hyperuricemia (sUA≥476umol/L) are enrolled and treated previously with colchicine or meloxicam (resistant to colchicine) for 8 weeks to prevent acute gout attack before the randomization. Treatment is to be conducted for 16 weeks after a two-week washout period. According to relevant guidelines, the standard treatments for gout are relatively well-established, so the trial considered the standard arm received a face-to-face treatment in clinics with febuxostat and are scheduled specialist appointment during the follow-up period. The test treatment of telemedicine arm is consistent to the standard arm, but the patients

Table 5
A hypothetical example for the sample size estimation and testing procedures.

Interim	Critical values	Data			Stage1		Stage2			
		T	S	P	T vs. P	Pass	NI test	Pass	S vs. P	Pass
1	4.4161	23/61	28/61	4/61	4.4702	Yes	1.6681	No		
2	2.9150	50/140	64/140	4/61			2.0040	No		
3	2.3307	85/219	110/219	4/61			2.5179	Yes	9.4270	Yes
4	1.9980	–	–	–						

are allocated a smart phone and virtual visited with a transition of care doctor (teledoc). Their medication adherences are documented and all the examinations are measured by the identified institution. To validate the assay sensitivity, an additional placebo group was conducted to received no treatment, but all the patients are informed to improve the lifestyle, such as reducing the consumption of alcohol, fructose-sweetened drinks, seafood and other triggers. The sUA is measured each 4 weeks in the follow-up durations, in which if the last two measurements (the 12th and 16th weeks' values) are lower than 357umol/L, then be considered as effective. The trial aims to detect the effect preservation of the telemedicine compared to the standard treatment.

The trial assumes the response rates as 35 %, 45 %, 5 % for the telemedicine arm, standard arm and placebo arm respectively and a target 50 % of the standard effect to be preserved by the telemedicine treatment. With the 1:1:1 allocation ratio, the sample sizes for the fixed sample design are determined as 672 (224 subjects in each arm) for 80 % power at one-sided significant level of 2.5 % with RD scale. If choose the group sequential design, the required sample size for a maximum of K = 4 sequential NI design is 657 (N_{01}), in which the first stage sample size is 61 in each arm at $\alpha_{sup} = 2.5\%$ and $\beta_{sup} = 1\%$. With the O'Brian & Fleming type boundary $(b_1, b_2, b_3, b_4) = (4.4161, 2.9150, 2.3307, 1.9980)$, the sample size increment is 79 in each arm in the NI test stage while the average of total sample size was estimated as 530. Meanwhile, the sample size was estimated as 477 in the fixed design and 696 designed in group sequential design with the OR scale. Though the fixed sample size with OR scale was smaller than the average sample size of group sequential design with RD scale, the researchers were not confident enough for suspicious compliance of the telemedicine treatment. However, the first stage with RD endpoint would need only 122 patients and can stop early if the superiority of the telemedicine relative to placebo were not verified. Therefore, the trial chose the group sequential design with RD scale for the balance of treatment effect and risk.

When the trial was carried out, the test procedures can be implemented as in Table 5. In the first stage, the telemedicine treatment is shown to be superior to placebo as the test statistic is higher than 1.96. Meanwhile, the placebo arm can be used to construct the three-arm structure in the interim analysis. Since the statistic of three-arm NI test 1.6681 does not cross the success boundary 4.4161, the enrollment should continue and only enroll patients to the telemedicine and standard treatment. Then, the second interim analysis indicated the trial should continue. Until the third interim analysis, the test statistic 2.5179 is higher than the boundary 2.3307 and the standard treatment is superior to placebo. Therefore, the study is stopped early for efficacy and claimed that the telemedicine treatment is not inferior to the standard treatment with actually 499 patients treated.

5. Discussion

In this article, we concentrate on a three-arm NI trial with dichotomous endpoints and made a comparison among RD, OR type functionals with MLE and RMLE methods so that additional guidance can be provided for investigators to design a NI trial for all situation. In practice, RR is also commonly applied in NI trials with binary endpoint. Since Chowdhury [23] had compared RR and OR scale in NI trials, we focus RD and OR as a logarithm scale to observe their perform. It is found that sample size determination on OR scale gives smaller sample size and robust procedure compared to RD scale in the majority of situations. When detecting the behaviors of the attained power, the RMLE methods based on OR scale outperforms the MLE method and would tend to have more power to reject the null hypothesis especially under the small sample size situations.

On the other hand, the three arm NI design has a natural hierarchical structure. A group sequential type design has a great flexibility with the early stopping options which are especially useful when there are great uncertainties in the study assumptions. Simulation results demonstrate that the group sequential design has better OC, which provided a comparable power to that of the fixed design while needing smaller average sample sizes for all cases. In addition, we suggest a lower significance level with a higher power for the sample size determination in the superiority test stage. In this way, the total sample sizes will significantly reduce while the subjects in the placebo group does not increase much. For better understanding and clarity of the magnitude of the treatment effect at each stage in the group sequential trial, we only conducted the simulation on the RD scale. The transformation of OR scale can be made similar to the function (2.1.9) and (2.1.10) in each stage in the group sequential design.

In this article, we restricted ourselves to the frequentist approach only and applied the fraction of the estimated effect size from the reference groups to construct the NI margin. As evident, the information from the historical trials may play a significant role in NI trial design and hierarchical bayesian approach may provide an attractive framework to achieve this in the future.

CRediT authorship contribution statement

Wenwen Wang: Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Yaru Huang:** Writing – review & editing, Writing – original draft, Software, Conceptualization. **Jielai Xia:** Funding acquisition, Conceptualization. **Ling Wang:** Funding acquisition, Conceptualization. **Chen Li:** Methodology, Writing – review & editing, Funding acquisition, Conceptualization.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Funding

This work has been supported by grants from National Natural Science Foundation of China (Nos. 82273728, 82273729 and 82373680) and key research and development program in Shaanxi Province (Nos.2022SF-027, 2023-YBSF-660).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] R.B.S. D'Agostino, J.M. Massaro, L.M. Sullivan, Non-inferiority trials: design concepts and issues - the encounters of academic consultants in statistics, Stat. Med. 22 (2) (2003) 169–186.

[2] H.M.J. Hung, S.J.W. O'Neill, Robert, Issues with statistical risks for testing methods in noninferiority trial without a placebo ARM, J. Biopharm. Stat. (2007).

[3] A. Koch, J. Röhmel, Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference, J. Biopharm. Stat. 14 (2) (2004) 315–325.

[4] H.M. James Hung, S.J. Wang, Y. Tsong, J. Lawrence, R.T. O'Neil, Some fundamental issues with non-inferiority testing in active controlled trials, Stat. Med. 22 (2) (2003) 213–225.

[5] J.A. Lewis, B. Jonsson, G. Kreutz, C. Sampaio, B. van Zwieten-Boot, Placebo-controlled trials and the declaration of helsinki, Lancet 359 (9314) (2002) 1337–1340.

[6] ICH-E10, Choice of Control Group in Clinical Trials, 2000.

[7] E. Hida, T. Tango, On the three-arm non-inferiority trial including a placebo with a prespecified margin, Stat. Med. 30 (3) (2011) 224–231.

[8] Committee for medicinal Products for human use (CHMP) guideline on the choice of the non-inferiority margin, Stat. Med. 25 (10) (2006) 1628–1638.

[9] I. Pigeot, J. Schäfer, J. Röhmel, D. Hauschke, Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo, Stat. Med. 22 (6) (2003) 883–899.

[10] M.L. Tang, N.S. Tang, Tests of noninferiority via rate difference for three-arm clinical trials with placebo, Journal of Biopharmaceutical Stats 14 (2) (2004) 337–347.

[11] M. Mielke, A. Munk, A. Schacht, The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints, Stat. Med. 27 (25) (2008) 5093–5110.

[12] M. Hasler, R. Vonk, L.A. Hothorn, Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity, Stat. Med. 27 (4) (2008) 490–503.

[13] K. Kombrink, A. Munk, T. Friede, Design and semiparametric analysis of non-inferiority trials with active and placebo control for censored time-to-event data, Stat. Med. 32 (18) (2013) 3055–3066.

[14] Y. Wu, Y. Li, Y. Hou, K. Li, X. Zhou, Study duration for three-arm non-inferiority survival trials designed for accrual by cohorts, Stat. Methods Med. Res. 27 (2) (2018) 507–520.

[15] Y. Wu, L. Zhao, Y. Hou, K. Li, X. Zhou, Correcting for non-compliance in randomized non-inferiority trials with active and placebo control using structural models, Stat. Med. 34 (6) (2015) 950–965.

[16] G. Homma, T. Daimon, Sample size calculation for "gold-standard" noninferiority trials with fixed margins and negative binomial endpoints, Stat. Biopharm. Res. 13 (4) (2021) 435–447.

[17] S. Ghosh, E. Paul, S. Chowdhury, R.C. Tiwari, New approaches for testing non-inferiority for three-arm trials with Poisson distributed outcomes, Biostatistics 23 (1) (2022) 136–156.

[18] T. Mütze, F. Konietschke, A. Munk, T. Friede, A studentized permutation test for three-arm trials in the 'gold standard' design, Stat. Med. 36 (6) (2017) 883–898.

[19] H. Lu, H. Jin, W. Zeng, A more efficient three-arm non-inferiority test based on pooled estimators of the homogeneous variance, Stat. Methods Med. Res. 27 (8) (2018) 2437–2446.

[20] E. Paul, R.C. Tiwari, S. Chowdhury, S. Ghosh, A more powerful test for three-arm non-inferiority via risk difference: frequentist and Bayesian approaches, J. Appl. Stat. 50 (4) (2023) 848–870.

[21] G. Li, Comments on 'Planning and analysis of three-arm non-inferiority trials with binary endpoints' by M. Kieser and T. Friede, Stat. Med. 26 (2007) 253–273. Statistics in medicine. 2011;30(3):298-9; author reply 300.

[22] A. Munk, M. Mielke, G. Freitag, G. Skipka, Testing noninferiority in three-armed clinical trials based on likelihood ratio statistics, Canadian Journal of Stats 35 (3) (2007) 413–431.

[23] S. Chowdhury, R.C. Tiwari, S. Ghosh, Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial, Comput. Stat. Data Anal. 132 (2019) 70–83.

[24] M. Kieser, T. Friede, Planning and analysis of three-arm non-inferiority trials with binary endpoints, Stat. Med. 26 (2) (2007) 253–273.

[25] P. Schlömer, W. Brannath, Group sequential designs for three-arm 'gold standard' non-inferiority trials with fixed margin, Stat. Med. 32 (28) (2013) 4875–4889.

[26] T. Ochiai, T. Hamasaki, S.R. Evans, K. Asakura, Y. Ohno, Group-sequential three-arm noninferiority clinical trial designs, J. Biopharm. Stat. 27 (1) (2017) 1–24.

[27] J.-T. Liu, C.S. Tzeng, H.-H. Tsou, 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, Int. J. Stat. Med. Res. 3 (2014) 11–20.

[28] G. Li, S. Gao, A group sequential type design for three-arm non-inferiority trials with binary endpoints, Biom. J. 52 (4) (2010) 504–518.

[29] J. Meis, M. Pilz, C. Herrmann, B. Bokelmann, G. Rauch, M. Kieser, Optimization of the two-stage group sequential three-arm gold-standard design for non-inferiority trials, Stat. Med. 42 (4) (2023) 536–558.

[30] Mielke M, Munk A. The Assessment and Planning of Non-inferiority Trials for Retention of Effect Hypotheses - towards a General Approach. Arxiv preprint. 2009;arXiv:0912.4169.

[31] T.H. Ng, Noninferiority hypotheses and choice of noninferiority margin, Stat. Med. 27 (26) (2008) 5392–5406.

[32] S. Chowdhury, R.C. Tiwari, S. Ghosh, Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial, Comput. Stat. Data Anal. 132 (2019).

[33] T.A. Schwartz, J.S. Denne, A two-stage sample size recalculation procedure for placebo- and active-controlled non-inferiority trials, Stat. Med. 25 (19) (2006) 3396–3406.

[34] A. Koch, J. Röhmel, Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference, Journal of Biopharmaceutical Stats 14 (2) (2004) 315–325.

[35] G. Li, S. Gao, A group sequential type design for three-arm non-inferiority trials with binary endpoints, Biometrical journal Biometrische Zeitschrift 52 (4) (2010) 504–518.

[36] S.J. Pocock, Group sequential methods in the design and analysis of clinical trials, Biometrika 64 (2) (1977) 191–199.

[37] P.C. O'Brien, T.R. Fleming, A multiple testing procedure for clinical trials, Biometrics 35 (3) (1979) 549–556.

[38] N.S. Tang, B. Yu, M.L. Tang, Testing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints, BMC Med. Res. Methodol. 14 (1) (2014) 134.

[39] P. Richette, T. Bardin, Gout. Lancet 375 (9711) (2009) 318–328.