Communications for Statistical Applications and Methods
2024, Vol. 31, No. 6, 613–625

https://doi.org/10.29220/CSAM.2024.31.6.613
Print ISSN 2287-7843 / Online ISSN 2383-4757

A comparison of testing methods in non-inferiority clinical trials

Jieun Park[a], Jae Won Lee[1,a]

[a]Department of Statistics, Korea University, Korea

Abstract

A general non-inferiority (NI) clinical trial is typically conducted using parametric testing methods with large samples. However, patient recruitment challenges often hinder rare disease trials, leading to enrollment failures. In this study, we introduce current parametric and nonparametric NI trial testing methods and propose modifications to enhance the performance of the nonparametric approach. Through a comprehensive simulation study with various sample sizes, data distributions, and sample ratios, we compare empirical levels and statistical powers as criteria for evaluating performance. Our findings indicate that the modified nonparametric methods outperformed the existing methods, particularly under conditions of small sample sizes and non-normal distributions, offering valuable insights for improving the reliability and sensitivity of NI trials in the context of rare diseases.

Keywords: non-inferiority clinical trial, nonparametric method, relative effect, rank-based, rare disease. pseudo rank

1. Introduction

Most non-inferiority (NI) clinical trial is conducted with large-scale enrollment. With this scale, the parametric testing method is used to confirm the non-inferiority in general. Hida and Tango (2011) introduced a parametric t-test to assess the NI trial. In this approach, the mean of each treatment group is used as the measure of the hypotheses, and the test statistic follows a Student's t-distribution and degrees of freedom vary according to variance's homogeneity. However, in the case of rare diseases, it is difficult to recruit patients who meet the eligibility criteria. In this case, a clinical trial is terminated due to insufficient enrollment, and research on the rare disease could not persist. Therefore, there have been some research on NI trials with nonparametric methods, which are less restrictive than parametric methods. The nonparametric method is not affected by the normality of data and the homogeneity of variance.

In situations where nonparametric methods are appropriate, enrollment is small scale or normality of data is not assured, there have been several methods. In this paper, we consider the three-arm design non-inferiority clinical trial. The three-arm design indicates that the clinical trial with experimental, reference drug, and placebo. The importance of three-arm design is highlighted in various research since we can show the assay sensitivity in the presence of a placebo arm. In the ICH E10 guideline (2000), assay sensitivity is defined as the property of a clinical trial defined as the ability to distinguish an effective treatment from a less effective or inefficient treatment. Thus, in many non-inferiority

[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-Ro, Sungbuk-Gu, Seoul 02841, Korea.
E-mail: jael@korea.ac.kr

Published 30 November 2024 / journal homepage: http://csam.or.kr

clinical trials, demonstrating assay sensitivity takes precedence over establishing non-inferiority. Park and Kim (2014) proposed a ratio-shape formulation for the non-inferiority hypotheses. They utilized a Hodges-Lehmann estimator to test this non-inferiority hypothesis. However, there is an alternative approach to assessing nonparametric NI trials known as the 'relative effect' which uses the relative effect as a measure of the hypothesis instead of the mean effect. Munzel (2009) introduced a ratio-shaped hypothesis with relative effect.

In this paper, we introduce a modification of the method by Park and Kim (2014) applicable in a three-arm NI trial, along with a modification of Munzel (2009) method utilizing unweighted relative effect—a measure calculated with pseudo rank, offering advantages over traditional weighted rank. While various NI trial testing methods have been proposed, comprehensive comparisons of their performance remain scarce. Our primary focus lies in introducing and applying these methods. To rigorously assess their performance and effectiveness, we conduct a comprehensive comparison with existing testing methods through a carefully designed simulation study. By evaluating empirical levels and statistical powers, we aim to provide valuable insights into the practical utility of these methods in non-inferiority clinical trials.

This paper is structured as follows. Section 2 introduces existing NI trial testing methods, encompassing both parametric and nonparametric approaches. In Section 3, we present the modification of the method by Park and Kim for three-arm clinical trials and propose the utilization of unweighted relative effect in Munzel's method. Section 4 presents the simulation results, where we evaluate the performance of the testing methods. Finally, Section 5 concludes our results and initiates a discussion on the implications of our findings.

2. Existing testing methods

2.1. Parametric method

Hida and Tango (2010) suggested a parametric testing method when a NI trial includes a single experimental, reference treatment and placebo. Assume that the primary endpoint under the three-arm is $X_{ij}, i = E, R, P, j = 1, \ldots, n_i$, respectively. X_{E_j}, X_{R_j} and X_{P_j} are mutually independent and normally distributed with unknown but common variance σ^2, that is, $X_{E_j}\ N(\mu_E, \sigma^2)$, $j = 1, \ldots, n_E$, $X_{R_j}\ N(\mu_R, \sigma^2)$, $j = 1, \ldots, n_R$ and $X_{P_j}\ N(\mu_P, \sigma^2)$, $j = 1, \ldots, n_P$. The total sample size is $N = n_E + n_R + n_P$. Each sample size of treatment group is not necessarily identical. The non-inferiority trial and assay sensitivity null hypotheses are $H_{0E} : \mu_E - \mu_R \leq -M_2$ and $A_0 : \mu_R - \mu_P \leq M_1$. H_{0E} is the hypothesis of NI trial and A_0 is the hypothesis of assay sensitivity. The corresponding alternative hypotheses are $H_{1E} : \mu_E - \mu_R > -M_2$ and $A_1 : \mu_R - \mu_P > M_1$ respectively. M_1 is the entire effect size of reference drug and M_2 is the largest clinically acceptable difference, i.e., non-inferiority margin. It is required that $M_1 \geq M_2 = r \times M_1$, where $0 < r \leq 1$. Choosing the proper value of M_1 and M_2 is found in FDA guidance for non-inferiority trials (2016).

To test the null hypotheses H_{0E} and A_0, Student t-test is used. T_E is the test statistic for NI trial

$$T_E = \frac{\bar{X}_E - \bar{X}_R + M_2}{\hat{\sigma}_{ER}\sqrt{(1/n_E) + (1/n_R)}},$$

and T_A is the test statistics for proving assay sensitivity.

$$T_A = \frac{\bar{X_R} - \bar{X}_P - M_1}{\hat{\sigma}_{RP}\sqrt{(1/n_R) + (1/n_P)}},$$

where $\bar{X}_i = (1/n_i)\sum_{j=1}^{n_j} X_{ij}, i = E, R, P, j = 1, \ldots, n_i$. And the homogeneous variance $\hat{\sigma}_{ER}$ and $\hat{\sigma}_{RP}$ are defined $\hat{\sigma}^2_{ER} = ((n_E - 1){s_E}^2 + (n_R - 1){s_R}^2)/n_E + n_R - 2$ and $\hat{\sigma}^2_{RP} = ((n_R - 1){s_R}^2 + (n_p - 1){s_P}^2)/n_R + n_P - 2$ where ${s_E}^2$, ${s_R}^2$ and ${s_P}^2$ denote a sample variance of experimental, reference and placebo treatments, respectively. The test statistic T_E follows a t-distribution with degrees of freedom $(n_E + n_R - 2)$ and T_A follows a t-distribution with degrees of freedom $(n_R + n_P - 2)$. And corresponding $100 \times (1 - \alpha)\%$ confidence intervals are

$$\left[\left(\bar{X}_E - \bar{X}_R\right) - t_{\frac{\alpha}{2}(n_E+n_R-2)} \times \hat{\sigma}_{ER}\sqrt{\frac{1}{n_E} + \frac{1}{n_R}}, \left(\bar{X}_E - \bar{X}_R\right) - t_{\frac{\alpha}{2}(n_E+n_R-2)} \times \hat{\sigma}_{ER}\sqrt{\frac{1}{n_E} + \frac{1}{n_R}}\right],$$

and

$$\left[\left(\bar{X}_R - \bar{X}_P\right) - t_{\frac{\alpha}{2}(n_R+n_P-2)} \times \hat{\sigma}_{RP}\sqrt{\frac{1}{n_R} + \frac{1}{n_p}}, \left(\bar{X}_R - \bar{X}_P\right) - t_{\frac{\alpha}{2}(n_R-n_P-2)} \times \hat{\sigma}_{RP}\sqrt{\frac{1}{n_R} + \frac{1}{n_P}}\right]$$

respectively.

In case of heterogenous variance, the identical confidence interval is applied and the only difference is degrees of freedom. Detailed formulas are found in Huang *et al.* (2015).

2.2. Nonparametric methods

As mentioned in the introduction, we can divide nonparametric NI trial testing methods into two categories. The first category involves NI hypotheses testing with mean effect, while the other category focuses on NI hypothesis testing with relative effect. We first introduce the method by Park and Kim (2014). However, in our comparison scenario assuming the clinical trial with a three-arm design, Park-Kim method is not included because Park-Kim method is applied in case of two-arm clinical trial which contains only a single experimental drug and reference drug, not a placebo.

2.2.1. Park-Kim method

Park and Kim (2014) introduced a nonparametric NI trial based on the Wilcoxon rank-sum test and Hodges-Lehmann estimator of reference drug. Let $X_{E_i}, i = 1, \ldots, n_E$ and $X_{R_j}, j = 1, \ldots, n_R$ are the primary endpoints from the experimental group and reference group respectively. Then the null hypothesis of non-inferiority trial is $H_{0E} : (\mu_E - \mu_R)/\mu_R \leq \lambda$ and the corresponding alternative hypothesis is $H_{1E} : (\mu_E - \mu_R)/\mu_R > \lambda$ where λ is $M_2 - 1$ and M_2 is a non-inferiority margin.

For all X_{E_i} and $X_{R_j} (i = 1, \ldots, n_E, j = 1, \ldots, n_R)$, define Q_{ij} as $Q_{ij} = X_{R_j} - X_{E_i}$ and its order statistics as $Q_{(1)}, Q_{(2)}, \ldots, Q_{(n_E n_R)}$. Then the rank-sum statistics are defined as median of Q_{ij}. Therefore, upper and lower limit of $100 \cdot (1 - \alpha)\%$ confidence interval of Wilcoxon rank sum test is $LL_w = Q_{(C_\alpha)}, UL_w = Q_{(n_E n_R + 1 - C_\alpha)}$, respectively, where $C_\alpha = ((n_E(2n_R + n_E + 1))/2) + 1 - w_\alpha$ and w_α is upper $100 \times \alpha^{th}$ quantile of Wilcoxon rank sum statistics $W_{R_j R_j\prime} = (X_{R_j} + X'_{R_j})/2$ when $R_j \leq R_j', (R_j, R_j' = 1, \ldots, n_R)$. Consequently, the lower and upper limit of nonparametric $100 \times (1 - \alpha)\%$ confidence interval of $(\mu_E - \mu_R)/\mu_R$ is

$$LL_N = \frac{Q_{(c_\alpha)}}{\text{med}(W_{R_j R_j\prime})'} \qquad R_j \leq R_j', \left(R_j, R_j' = 1, \ldots, n_R\right)$$

$$UL_N = \frac{Q_{(n_E n_R + 1 - C_\alpha)}}{\text{med}(W_{R_j R_j\prime})'} \qquad R_j \leq R_j', (R_j, R_j' = 1, \ldots, n_R)$$

respectively. In Section 3, we reformulate this method to be used in three-arm design.

2.2.2. Park-Kim method

Before we explain the Munzel method, the relative effect needs to be defined first. Let

$$X_{ij} \sim F_i, \quad i = 1, \ldots, k, \;\; j = 1, \ldots, n_i,$$

where i denotes treatment group and j denotes the individual within the ith treatment. $F_i(x) = P(X_{ij} < x) + (1/2)P(X_{ij} = x) = (1/2)[F_i^- + F_i^+]$ denotes the average of the left and right continuous version of the distribution function, where $F_i^- = P(X < x)$ is the left-continuous version and $F_i^+ = P(X \leq x)$ is the right-continuous version of the cumulative distribution function (cdf) of X. This statistical model does not include any parameter, and it could be used to describe the influence of the treatment to the observation. Thus, the marginal distribution function can be used to describe the relative effect

$$p_i = \int H dF_i = P\left(X_{ij} < x\right) + \frac{1}{2} P\left(X_{ij} = x\right), \quad i = 1, \ldots, k, \;\; j = 1, \ldots, n_i,$$

where H denotes a mean distribution of F. Additional explanation of relative effect is found in Brunner *et al.* (2018, 2021).

Munzel (2009) suggested a non-inferiority testing and assay sensitivity null hypothesis $H_{0E} : (p_E - p_R)/p_R - p_P \leq -\delta$ and $A_0 : p_R - p_P \leq Q_1$ and corresponding alternative hypotheses are $H_{1E} : (p_E - p_R)/p_R - p_P > -\delta$ and $A_1 : p_R - p_P > Q_1$. Where p_E, p_R and p_P are relative effect of experimental, reference and placebo group respectively, and Q_1 is entire relative effect size of R (same as M_1 in parametric setting) and $Q_2 = \delta(p_R - p_P)$ is the margin of the test. Thus, δ plays role in parametric testing as 'γ'.

Applying Fieller's theorem, the two-sided $(1 - \alpha) \times 100\%$ confidence interval for ratio $(p_E - p_R)/p_R - p_P$ is,

$$\frac{1}{(1-g)} \left[\frac{\hat{P}_E - \hat{P}_R}{\hat{P}_R - \hat{P}_P} - \frac{g \cdot \text{cov}(\hat{p}_E - \hat{p}_R, \hat{p}_R - \hat{p}_P)}{\text{var}(\hat{p}_R - \hat{p}_P)} \pm \frac{z_{1-\alpha/2}}{\hat{p}_R - \hat{p}_P} \sqrt{C} \right],$$

where

$$g = \frac{z^2_{1-\frac{\alpha}{2}} \cdot \text{var}\,(p_R - p_P)}{(\hat{p}_R - \hat{p}_P)^2},$$

and

$$\begin{aligned} C = {} & \text{var}\,(\hat{p}_R - \hat{p}_P) - 2 \cdot \frac{\hat{p}_E - \hat{p}_R}{\hat{P}_R - \hat{p}_P} \cdot \text{cov}(\hat{p}_E - \hat{p}_R, \hat{p}_R - \hat{p}_P) + \left(\frac{\hat{p}_E - \hat{p}_R}{\hat{p}_R - \hat{p}_P} \right)^2 \cdot \text{var}(\hat{p}_R - \hat{p}_P) \\ & - g \cdot \left(\text{var}(\hat{p}_E - \hat{p}_R) - \frac{(\text{cov}(\hat{p}_E - \hat{P}_R, \hat{p}_R - \hat{p}_P))^2}{\text{var}(\hat{p}_R - \hat{p}_P)} \right). \end{aligned}$$

Also, define R_{ik} as vector of overall rank of X_{ik} among all N observations and $R_{ik}^{(i)}$ as internal rank of X_{ik} among all n_i observations in the i^{th} treatment group and $R_{ik}^{(-j)}$ as partial rank of X_{ik} among all

$N - n_j$ observations except those in the j^{th} treatment group, then

$$\text{var}\left(\sqrt{N}p_i\right) = \frac{1}{N}\left[\frac{1}{n_i(n_i-1)}R_i^t \cdot R_i + \frac{1}{n_i^2}\sum_{r=1}^{3}\frac{n_r}{n_r-1}R_{ri}^t \cdot R_{ri}\right], \quad \text{and}$$

$$\text{cov}\left(\sqrt{N}\hat{p}_i,\ \sqrt{p}_i\right) = \frac{1}{N}\left[\frac{1}{n_i n_j}\sum_{r=1}^{3}\frac{n_r}{n_r-1}R_{ri}^t \cdot R_{rj} - \frac{1}{n_j(n_i-1)}R_i^t \cdot R_{ij} - \frac{1}{n_i(n_j-1)}R_j^t \cdot R_{ji}\right],$$

where $R_i = \{R_{ik} - R_{ik}^{(i)} - \bar{R}_{i.} + \bar{R}_{i.}^{(i)}\}_{k=1,\ldots,n_i}$ and $R_{ij} = \{R_{ik} - R_{ik}^{(-j)} - \bar{R}_{i.} + \bar{R}_{i.}^{(-j)}\}_{k=1,\ldots,n_i}$ for $j \neq i$ and $i = E, R, P$, and $r = 1, 2, 3$ represent experimental, reference and placebo respectively.

3. Proposed methods

3.1. Modified Park-Kim method

In this section, we modify the Park-Kim method by reformulating it in three-arm design case. We add placebo arm in the hypothesis, and the NI hypothesis $H_0 : (\mu_E - \mu_R)/\mu_R - \mu_P \leq -\gamma$ vs. $(\mu_E - \mu_R)/\mu_R - \mu_P > -\gamma$, where r is pre-specified margin. We have retained the numerator and made a modification to the denominator estimator, changing it to two sample of Hodges-Lehmann estimator.

Suppose $\hat{\Delta} = \text{median}\{(X_{R_j} - W_{P_s}), j = 1, \ldots, n_R, s = 1, \ldots, n_P\}$, where X_{R_j} is the j^{th} response of reference treatment and W_{P_s} is the sth response of placebo treatment. And then define the order statistic of $(X_{R_j} - W_{P_s})$ as $H_{(1)}, H_{(2)}, \ldots, H_{(n_R n_P)}$. When $n_R n_P$ is odd, $n_R n_P = 2b+1$ then $b = (n_R n_P - 1)/2$, the Hodges-Lemann estimator would be $\hat{\Delta} = H_{(b+1)}$, and when $n_R n_P$ is even, $n_R n_P = 2b$ then $b = (n_R n_P)/2$, the Hodges-Lemann estimator would be $\hat{\Delta} = (H_{(b)} + H_{(b+1)})/2$. Therefore, the lower and upper limit of modified nonparametric $100 \times (1-\alpha)\%$ confidence interval of $(\mu_E - \mu_R)/\mu_R - \mu_P$ is

$$LL_N = \frac{Q_{(C_\alpha)}}{\hat{\Delta}}, \quad R_j \leq R_j', \left(R_j, R_j' = 1, \ldots, n_R\right)$$

$$UL_N = \frac{Q_{(n_E n_R + 1 - C_\alpha)}}{\hat{\Delta}}, \quad R_j \leq R_j', \left(R_j, R_j' = 1, \ldots, n_R\right)$$

respectively.

3.2. Modified Munzel method using unweighted relative effect

In section 2.2.2, we introduced the concept of a 'usual rank', which is calculated based on the total number of observations. Pseudo rank, on the other hand, is slightly different. It represents an unweighted rank, where we consider the total number of groups instead of the total number of observations. We refer to relative effects calculated using pseudo rank as 'fixed relative effects' because they are not influenced by the number of observations. The concept of unweighted relative effect was suggested by Brunner and Puri (1996) for the first time, and then asymptotic properties of unweighted relative effect was demonstrated by Domhof (2001). Brunner *et al.* (2021) cautioned that the weighted relative effect can vary according to the sample ratio, making it an unstable measure for use in non-inferiority trials. Therefore, we opted to use the pseudo rank method, which remains stable and is not affected by the sample ratio.

Table 1: Performances for clustered DIC data after 1000 iterations

Distribution	Normal distribution		Gamma distribution($\beta = 2$)	Exponential distribution
Parameters	(μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		$(\alpha_E, \alpha_R, \alpha_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$(\lambda_E, \lambda_R, \lambda_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$
Scenarios	(12, 14, 0) (2, 2, 2)	(12, 14, 0) (2, 1, 1)	(4, 5, 1) (8, 10, 2) (16, 20, 4)	(0.5, 0.25, 5) (2, 4, 0.2) (4, 16, 0.04)
	(14, 14, 0) (2, 2, 2)	(14, 14, 0) (2, 1, 1)	(5, 5, 1) (10, 10, 2) (20, 20, 4)	(0.25, 0.25, 5) (4, 4, 0.2) (16, 16, 0.04)

Table 2: Sample ratio and size of simulated scenarios

Ratio $n_E : n_R : n_P$	Small samples (n_E, n_R, n_P)	Large samples (n_E, n_R, n_P)
1 : 1 : 1	(20, 20, 20)	(50, 50, 50)
1 : 2 : 1	(20, 40, 20)	(50, 100, 50)
2 : 2 : 1	(40, 40, 20)	(100, 100, 50)
2 : 1 : 1	(40, 20, 20)	(100, 50, 50)

Let $U = (1/K)\sum_{i=1}^{k} F_i$ denote the unweighted mean distribution. As mentioned above, it is not affected by total number of observations. Similar to usual rank, respective empirical version of function U is $\hat{U} = (1/K)\sum_{i=1}^{k} \hat{F}_i$. And let $R_{ij}{}^{\varphi}$ represents the pseudo rank of X_{ij} among all k treatment groups, then $R_{ij}{}^{\varphi}$ is defined as

$$R_{ij}{}^{\varphi} = \frac{1}{2} + N\hat{U}\left(X_{ij}\right) = \frac{1}{2} + \frac{N}{k}\Sigma_{l=1}^{n_r} c\left(X_{ij} - X_{rl}\right),$$

where $\hat{U}(x) = (1/k)\Sigma_{r=1}^{k} F_r(x)$ denotes the unweighted mean of empirical distribution functions. It can be also estimated consistently by the simple plug-in estimator

$$\hat{p}_i^{\varphi} = \int U d\hat{F}_i = \frac{1}{N}\left(\bar{R}_{i\cdot}^{\varphi} - \frac{1}{2}\right),$$

where $\bar{R}_{i\cdot}^{\varphi} = (1/n_i)\Sigma_{j=1}^{n_i} R_{ij}^{\varphi}$ and $\hat{U} = (1/k)\Sigma_{r=1}^{k}\hat{F}_r$ denotes the unweighted mean of empirical distributions $\hat{F}_1, \ldots, \hat{F}_r$. The value p_i^{φ} quantifies an effect of the distribution F_i with respect to the unweighted mean distribution U. Fixed relative effect $\hat{p}_i^{\varphi}$ is also related to mean of the pseudo rank $\bar{R}_{i\cdot}^{\varphi}$. The only difference between relative effect and fixed relative effect lies in the replacement of $\hat{W}(X_{ij})$ with $\hat{U}(X_{ij})$. Therefore, the application unweighted relative effect to Munzel method is simply substituting (weighted) relative effect to unweighted relative effect. The null hypotheses are $H_{0E} : (p^{\varphi}{}_E - p^{\varphi}{}_R)/P^{\varphi}{}_R - p^{\varphi}{}_p \leq -\delta$ and $A_0 : P^{\varphi}{}_R - P^{\varphi}{}_P \leq Q_1$. The corresponding alternative hypotheses are $H_{1E} : (p^{\varphi}{}_E - p^{\varphi}{}_R)/p^{\varphi}{}_R - p^{\varphi}{}_p > -\delta$ and $A_1 : P^{\varphi}{}_R - P^{\varphi}{}_P > Q_1$, respectively. The two-sided $(1-\alpha)\times 100\%$ confidence interval for ratio $(p^{\varphi}{}_E - p^{\varphi}{}_R)/p^{\varphi}{}_R - p^{\varphi}{}_P$ is exactly identical with just substituting p_i to $p^{\varphi}{}_i$, $i = E, R, P$.

4. Simulation study

4.1. Simulation scheme

To assess the performance of the testing methods introduced in the previous sections, we conducted an extensive simulation study. The criteria for demonstrating its performance are empirical level and

Table 3: Assay sensitivity margins (M_1 and Q_1) and NI margins (M_2 and Q_2) of each distribution and testing method

	Parametric method		Nonparametric method	
	M_1	M_2	Q_1	Q_2
Normal distribution	6	2.5	0.6	0.3
Gamma distribution	4	2	0.2	0.15
Exponential distribution	2.5	1.5	0.4	0.2

statistical power. We aim to assess how the performance of the testing methods vary under various situations. We generate the data from three distributions (normal, gamma and exponential distribution). For the normal distribution, we consider both equal variance and unequal variance cases. In case of the gamma distribution, we used two different shape parameters (α) while keeping the rate parameter fixed ($\beta = 2$). Similarly, for the exponential distribution, we employed two rate parameters (λ). We also consider the four types of sample ratios reflecting the real-world situations. Additionally, sample size is also an important factor, and thus we have set the ratio 1 as equivalent to 20 and 50 people in small sample and large samples, respectively. Sample ratio and sample size variation is displayed in Table 2. The distribution of data used in simulation is displayed in Table 1.

For normal distribution case, we consider a total of four scenarios considering both homogeneous variance and heterogeneous variance. For gamma and exponential distribution cases, two scenarios are used in simulation study.

The first step of evaluating performance of testing methods is checking the empirical level. We calculated and compared to nominal level before proceeding to the statistical power comparison. We iterate a total of 10,000 times for each testing method, and thus the empirical level is considered valid when if falls within the interval $[0.0457, 0.0543]$. If the empirical level deviates from this interval, the comparison of statistical power becomes unreliable. After confirming that the empirical level is satisfied, statistical power is calculated. As in Hinda and Tango (2011), the power of the testing procedure is defined as

$$\text{Power} = \Pr\left\{T_E > t_{\alpha/2}\,(n_E + n_R - 2) \cap T_A > t_{\alpha/2}\,(n_R + n_P - 2) \mid H_{0E}, A_0\right\}.$$

Setting the NI trial and assay sensitivity margin is also crucial in NI trial. The margins of parametric testing methods denoted by M_1 and M_2 are the same as in the previous studies. Subsequently, the corresponding margins of nonparametric testing methods, labeled as Q_1 and Q_2 , are chosen based on the well-known property

$$p_i = \int HdF_i = \frac{1}{N}\sum_{h=1}^{k} n_h \cdot \Phi\left(\frac{\mu_i - \mu_h}{\sigma_h\sqrt{2}}\right).$$

It means that Φ is almost linear around 0.5, i.e., it is approximately linearly connected to set the nonparametric effect similar to parametric effect. Thus, gamma and exponential distribution are also applied to its property using normal approximation. The margins for each testing method and distribution are demonstrated in Table 3.

Additionally, an assay sensitivity test must be conducted before initiating an NI trial. The NI trial is performed only after confirming assay sensitivity. If assay sensitivity is not established, the NI trial is not conducted and the simulation is terminated. Thus, the simulation steps are follows:

1) Generate data from each distribution with predefined mean, variance, or parameters.

Table 4: Levels for NI trial hypothesis testing in case of sample ratio $(n_E, n_R, n_P) = (1 : 1 : 1)$: small samples, nominal level 0.05, from 10,000 random iterations

Distribution	(μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{n_E}{N}, \frac{n_R}{N}, \frac{n_P}{N}\right), \frac{1}{d}$	M-UR	MM-PR	M-PK	HT-P
Normal dist.	(12, 14, 0) (2, 2, 2)	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \frac{1}{3}$	0.0493	0.0493	0.0521	0.0497
	(14, 14, 0) (2, 2, 2)		0.0492	0.0492	0.0518	0.0504
	(12, 14, 0) (2, 1, 1)		0.0508	0.0508	0.0511	0.0501
	(14,14,0) (2,1,1)		0.0510	0.0510	0.0520	0.0496
	$(\alpha_E, \alpha_R, \alpha_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
Gamma dist. ($\beta = 2$)	(4, 5, 1) (8, 10, 2) (16, 20, 4)		0.0513	0.0513	0.0465	0.0307
	(5, 5, 1) (10, 10, 2) (20, 20, 4)		0.0507	0.0507	0.0462	0.0314
	$(\lambda_E, \lambda_R, \lambda_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
Exponential dist.	(0.5, 0.25, 5) (2, 4, 0.2) (4, 16, 0.04)		0.0489	0.0489	0.0527	0.0256
	(0.25, 0.25, 5) (4, 4, 0.2) (16, 16, 0.04)		0.0494	0.0494	0.0522	0.0263

M-UR = Munzel-usual rank; MM-PR = modified Munzel-pseudo rank; M-PK = modified Park-Kim; HT-P = Hida-Tango parametric

2) Apply each testing method to the data and determine whether the hypothesis is rejected or accepted.

3) Repeat these steps 10,000 times for each method.

4) Calculate the empirical level and statistical power of each testing method.

All simulation studies were conducted using programming version 4.1.2 (R project homepage: http://www.r-project.org). The author developed the codes for generating rank statistics and implementing all testing methods from scratch.

4.2. Simulation results

To improve the readability of the tables, the abbreviations have been employed in the simulation result tables. Abbreviations are as follows: 'HT-P' (Hida-Tango parametric method), 'M-PK' (modified Park-Kim method), 'M-UR' (Munzel method using usual rank), 'MM-PR' (modified Munzel method using pseudo rank).

Table 5: Levels for NI trial hypothesis testing in case of sample ratio $(n_E, n_R, n_P) = (2:2:1)$: small samples, nominal level 0.05, from 10,000 random iterations

Distribution	(μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{n_E}{N}, \frac{n_R}{N}, \frac{n_P}{N}\right), \frac{1}{d}$	M-UR	MM-PR	M-PK	HT-P
Normal dist.	(12, 14, 0) (2, 2, 2)	$\left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right), \frac{1}{3}$	0.0480	0.0481	0.0513	0.0508
	(14, 14, 0) (2, 2, 2)		0.0513	0.0495	0.0480	0.0506
	(12, 14, 0) (2, 1, 1)		0.0503	0.0494	0.0519	0.0497
	(14, 14, 0) (2, 1, 1)		0.0505	0.0500	0.0522	0.0493
Gamma dist. ($\beta = 2$)	$(\alpha_E, \alpha_R, \alpha_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
	(4, 5, 1) (8, 10, 2) (16, 20, 4)		0.0516	0.0513	0.0470	0.0333
	(5, 5, 1) (10, 10, 2) (20, 20, 4)		0.0518	0.0506	0.0473	0.0343
Exponential dist.	$(\lambda_E, \lambda_R, \lambda_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
	(0.5, 0.25, 5) (2, 4, 0.2) (4, 16, 0.04)		0.0513	0.0483	0.0521	0.0276
	(0.25, 0.25, 5) (4, 4, 0.2) (16, 16, 0.04)		0.0495	0.0509	0.0525	0.0289

M-UR = Munzel-usual rank; MM-PR = modified Munzel-pseudo rank; M-PK = modified Park-Kim; HT-P = Hida-Tango parametric

4.2.1. Empirical level for each testing method

To assess the performance of different testing methods in determining empirical levels, we conducted simulations for scenarios involving an experimental drug. Four testing methods, as described in Sections 2 and 3, were compared. While we cannot present all simulation result tables here, we highlight key findings below.

In scenarios where sample ratios remain consistent across the experimental drug, reference drug, and placebo, both the M-UR and MM-PR methods exhibit identical relative effects. Consequently, these methods demonstrate the same empirical level. For instance, in Table 4, when the sample ratios are $1:1:1$, the empirical levels of M-UR and MM-PR across various distributions and parameters are consistent. All nonparametric testing methods demonstrate valid empirical levels in both normal and non-normal settings. Conversely, the parametric testing method proves to be invalid in all non-normal situations. Particularly, when data are generated under gamma and exponential distributions, the empirical level of the parametric testing method falls significantly below 0.0457, the lower limit of the nominal level of 0.5. The result demonstrates consistency when sample ratios are not equal across the experimental drug, reference drug, and placebo (see Table 5).

4.2.2. Statistical power of each testing method

The statistical power analysis reveals insights into the performance of each testing method under various conditions. When the sample ratio remains consistent across the experimental drug, reference drug, and placebo, the relative effects using usual rank and pseudo rank are identical. Consequently, as shown in Table 6, the statistical power of M-UR and MM-PR mirrors their empirical levels.

In Table 6, additional information about significant results is presented. Parametric testing methods demonstrate poor statistical power when data are generated from non-normal distributions, such as gamma and exponential distributions, with statistical power close to 0.5. We also compared statistical powers among different sample ratios. Four sample ratios were simulated, with a sample ratio of $2:2:1$ exhibiting the highest statistical power. Although not tabulated here, sample ratios of $2:1:1$, $1:2:1$, and $1:1:1$ follow in descending order. The statistical power of the $2:2:1$ sample ratio is detailed in Table 7.

In the case of a sample ratio of $1:1:1$, the parametric testing method demonstrates the highest statistical power under normal distribution, while the M-PK method exhibits the lowest. Conversely, in non-normal situations, the parametric testing method shows the lowest statistical power, with the order established as $P < M-PK < M-UR = MM-PR$ (see Table 6). Under a non-equal sample ratio of $2:2:1$, the statistical power of M-UR and MM-PR differs, with the MM-PR method exhibiting higher statistical power than M-UR in both normal and non-normal situations. The order established is $P < M-PK < M-UR < MM-PR$ (see Table 7).

5. Conclusion

We have presented various NI trial methods, including both parametric and nonparametric approaches. Additionally, we modified Park-Kim method using two sample Hodes-Lehmann estimator. Also, in nonparametric method using relative effect, we applied unweighted relative effect which uses the pseudo rank. Pseudo rank is calculated only with the number of the treatment groups, not affected by the sample ratio of each treatment group. So, it is a fixed measure while usual rank is changed its value according to the sample ratio and thus yields unstable measure.

Now we summarize the major findings from our simulation studies.

1. Parametric testing methods demonstrate superiority under normal distribution, while among nonparametric methods, MM-PR exhibits the highest statistical power. Conversely, in non-normal scenarios, parametric methods falter, while MM-PR consistently display superior performance.

2. Notably high statistical power is observed in scenarios with a ratio of 2 for experimental drugs and the reference drug, and a ratio of 1 for the placebo.

3. Testing method based on unweighted relative effect (MM-PR) consistently outperform those based on weighted relative effect (M-UR).

4. Statistical power tends to be higher in situations with larger sample sizes. Particularly, the MM- PR method shows comparable performance to parametric methods in normal cases, while excelling in both normal and non-normal scenarios.

5. In small sample scenarios, the MM-PR method exhibits satisfactory statistical power, highlighting its effectiveness, especially in trials with limited sample availability, such as those for rare diseases.

For ease of comparison, we converted Table 7 into a bar graph (Figure 1). Scenarios 1 to 8 represent different distributions listed in Table 7. For example, Scenario 1 is a normal distribution with means

Table 6: Statistical power for NI trial hypothesis testing in case of sample ratio $(n_E, n_R, n_P) = (1 : 1 : 1)$: small samples, nominal level 0.05, from 10,000 random iterations

Distribution	(μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{n_E}{N}, \frac{n_R}{N}, \frac{n_P}{N}\right), \frac{1}{d}$	M-UR	MM-PR	M-PK	HT-P
Normal dist.	(12, 14, 0) (2, 2, 2)		0.850	0.850	0.552	0.911
	(14, 14, 0) (2, 2, 2)		0.857	0.857	0.565	0.917
	(12, 14, 0) (2, 1, 1)		0.857	0.857	0.565	0.917
	(14, 14, 0) (2, 1, 1)		0.863	0.863	0.572	0.923
Gamma dist. ($\beta = 2$)	$(\alpha_{E,}, \alpha_R, \alpha_P)$ $(\mu_{E,}, \mu_R, \mu_P)$ $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \frac{1}{3}$	M-UR	MM-PR	M-PK	HT-P
	(4, 5, 1) (8, 10, 2) (16, 20, 4)		0.883	0.883	0.617	0.507
	(5, 5, 1) (10, 10, 2) (20, 20, 4)		0.908	0.908	0.625	0.512
Exponential dist.	$(\lambda_E, \lambda_R, \lambda_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
	(0.5, 0.25, 5) (2, 4, 0.2) (4, 16, 0.04)		0.872	0.872	0.497	0.482
	(0.25, 0.25, 5) (4, 4, 0.2) (16, 16, 0.04)		0.898	0.898	0.504	0.493

M-UR = Munzel-usual rank; MM-PR = modified Munzel-pseudo rank; M-PK = modified Park-Kim; HT-P = Hida-Tango parametric.

(12, 14, 0) and variances (2, 2, 2), while Scenario 6 is a gamma distribution with shape parameters (5, 5, 1), means (10, 10, 2), and variances (20, 20, 4). Each method is represented by different colors. The HT-P method (light grey) performed well in Scenarios 1–4 (normal distributions) but was inferior in Scenarios 5–8 (non-normal distributions). The MM-PR method maintained its position as the second best in Scenarios 1–4 and outperformed in Scenarios 5–8.

We anticipate that our proposed MM-PR method will be particularly useful in rare disease clinical trials, where sample sizes are limited, owing to its minimal susceptibility to data distribution. Additionally, our ongoing research into nonparametric testing methods for cases involving multiple experimental drugs aims to provide further insights into the complexities of relative effect estimation and variance-covariance estimation. All simulation studies were conducted using programming version 4.1.2 (R project homepage: http://www.r-project.org). The author developed the codes for generating rank statistics and implementing all testing methods from scratch.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208882, No. NRF-2022M3J6A1063595). This research was also supported and funded by the Korean National Police Agency [Project Name: Advancing the Appraisal Techniques of Forensic Entomology / Project Number: PR10-04-000-22].

Table 7: Statistical power for NI trial hypothesis testing in case of sample ratio $(n_E, n_R, n_P) = (2 : 2 : 1)$: small samples, nominal level 0.05, from 10,000 random iterations

Distribution	(μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{n_E}{N}, \frac{n_R}{N}, \frac{n_P}{N}\right), \frac{1}{d}$	M-UR	MM-PR	M-PK	HT-P
Normal dist.	(12, 14, 0) (2, 2, 2)		0.848	0.859	0.557	0.913
	(14, 14, 0) (2, 2, 2)		0.860	0.869	0.575	0.925
	(12, 14, 0) (2, 1, 1)		0.851	0.863	0.565	0.919
	(14, 14, 0) (2, 1, 1)		0.867	0.875	0.584	0.930
Gamma dist. ($\beta = 2$)	$(\alpha_E, \alpha_R, \alpha_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$	$\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}\right), \frac{1}{3}$	M-UR	MM-PR	M-PK	HT-P
	(4, 5, 1) (8, 10, 2) (16, 20, 4)		0.892	0.903	0.629	0.519
	(5, 5, 1) (10, 10, 2) (20, 20, 4)		0.913	0.920	0.635	0.523
Exponential dist.	$(\lambda_E, \lambda_R, \lambda_P)$ (μ_E, μ_R, μ_P) $(\sigma_E^2, \sigma_R^2, \sigma_P^2)$		M-UR	MM-PR	M-PK	HT-P
	(0.5, 0.25, 5) (2, 4, 0.2) (4, 16, 0.04)		0.876	0.884	0.514	0.492
	(0.25, 0.25, 5) (4, 4, 0.2) (16, 16, 0.04)		0.904	0.910	0.517	0.503

M-UR = Munzel-usual rank; MM-PR = modified Munzel-pseudo rank; M-PK = modified Park-Kim; HT-P = Hida-Tango parametric.

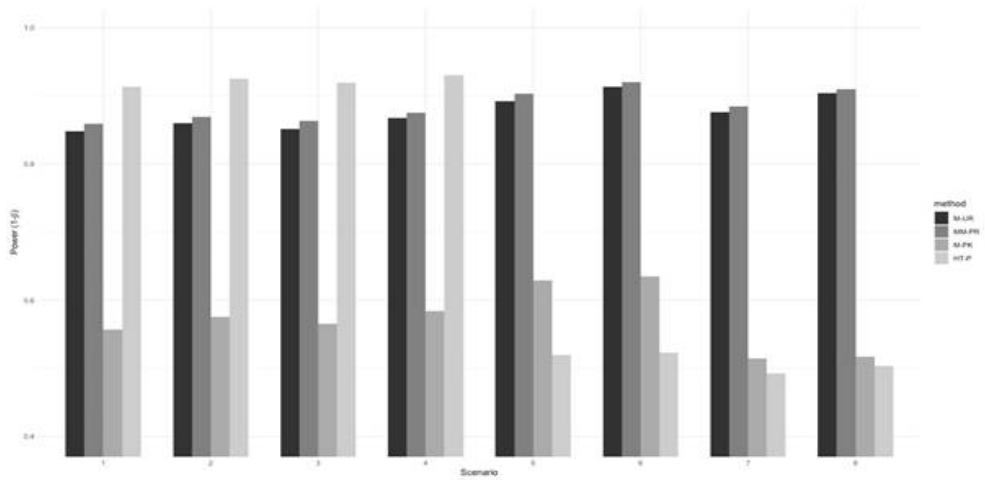

Figure 1: *Bar graph comparing statistical power across methods (based on Table 7).*

References

Brunner E, Bathke AC, and Konietschke F (2018). *Rank and Pseudo-rank Procedures for Independent Observations in Factorial Designs*, Springer International Publishing, Cham, Switzerland.

Brunner E, Konietschke F, Bathke AC, and Pauly M (2021). Ranks and Pseudo-ranks—Surprising results of certain rank tests in unbalanced designs, *International Statistical Review*, **89**, 349–366.

Brunner E and Puri ML (1996). 19 nonparametric methods in design and analysis of experiments, *Handbook of Statistics*, **13**, 631–703.

Domhof S (2001). Nichtparametrische relative Effekte (Ph D. thesis), University of Göttingen, Götting en.

Guideline IHT (2000). Choice of control group and related issues in clinical trials E10. Choice. E, 10, Available from: www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf

Hida E and Tango T (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin, *Statistics in Medicine*, **30**, 224–231.

Huang LC, Wen MJ, and Cheung SH (2015). Noninferiority studies with multiple new treatments and heterogeneous variances, *Journal of Biopharmaceutical Statistics*, **25**, 958–971.

Munzel U (2009). Nonparametric non-inferiority analyses in the three-arm design with active control and placebo, *Statistics in Medicine*, **28**, 3643–3656.

Park S and Kim D (2014). Nonparametric method for a non-inferiority test using confidence interval, *The Korean Journal of Applied Statistics*, **27**, 833–842.

U.S. Food and Drug Administration (2016). Non-inferiority clinical trials to establish effectiveness - guidance for industry, FDA guidance for industry, Available from: www.fda.gov/regulatory-information/search-fda-guidancedocuments/non-inferiority-clinical-trials-establish-effectiveness-guidance-industry

Received April 18, 2024; Revised September 20, 2024; Accepted October 02, 2024