A Test for Multiple Binary Endpoints with Continuous Latent Distribution in Clinical Trials

Takuma Ishihara¹ · Kouji Yamamoto²

Received: 9 July 2021 / Accepted: 23 September 2021 / Published online: 13 October 2021
© The Author(s) 2021

Abstract

In clinical trials, two or more binary responses obtained by dichotomizing continuous responses are often employed as multiple primary endpoints. Testing procedures for multiple binary variables with latent distribution have not yet been adequately discussed. Based on the association measure among latent variables, we provide a statistic for testing the superiority of at least one binary endpoint. In addition, we propose a testing procedure with a framework in which the trial efficacy is confirmed only when there is superiority of at least one endpoint and non-inferiority of the remaining endpoints. The performance of the proposed procedure is evaluated through simulations.

Keywords Latent variable · Non-inferiority · Superiority

Abbreviations

ACR	American College of Rheumatology criteria
ACR20	20% improvement in American College of Rheumatology criteria
ALRT	Approximate likelihood ratio test
DAS28	Disease Activity Score modified to include the 28 diarthrodial joint count
ESR	Erythrocyte sedimentation rate
4F-PCC	Four-factor prothrombin complex concentrate
ICH	International council for Harmonisation of technical requirements for pharmaceuticals for human use
INR	International normalized ratio
IUT	Intersection-union test
MLE	Maximum likelihood estimator

✉ Takuma Ishihara
tishiha@gifu-u.ac.jp

¹ Innovative and Clinical Research Promotion Center, Gifu University Hospital, Gifu, Japan

² Department of Biostatistics, Yokohama City University School of Medicine, Kanagawa, Japan

1 Introduction

In confirmatory clinical trials, several correlated binary response variables are used to assess the efficacy and safety of new treatments. The ICH E9 guideline [1] recommends that the primary endpoint should consist of only one variable that provides strong scientific evidence of treatment efficacy. However, in clinical trials for a variety of diseases, it is often useful to evaluate efficacy using multiple primary endpoints. For example, in clinical trials of patients with rheumatoid arthritis, the percentage of patients achieving short-term improvement of 20 percent in the American College of Rheumatology criteria (ACR20) and the percentage achieving long-term low disease activity [Disease Activity Score (DAS28-ESR) ≤ 3.2] are often used as primary endpoints (e.g., [2]). In clinical trials of patients with psoriasis, short- and long-term improvements are simultaneously assessed based on the percentage of patients with at least 75 percent improvement in the psoriasis area-and-severity index (DASI) score (e.g., [3]). In particular, binary endpoints are often used when it is more meaningful to diagnose improvement beyond clear standards rather than to assess the disease state using continuous variables. In such trials, we can consider that all primary endpoints are binary and often have a continuous latent distribution.

Most trials use multiple endpoints only to evaluate non-inferiority or superiority, but some trials have been conducted to confirm the non-inferiority and superiority of all endpoints. For example, a clinical trial to confirm the efficacy of four-factor prothrombin complex concentrate (4F-PCC) included two primary endpoints [4], namely the percentage of patients with a hemostatic effect and the percentage with a decrease in the international normalized ratio (INR). In the above trial, superiority was only evaluated if there was non-inferiority for both endpoints. When we confirm not only non-inferiority but also superiority, the use of the closed testing procedure [5] for the primary analysis is reasonable, and in this case, no adjustment is needed to control the type I error rate. Sozu et al. [6–8] have already proposed a testing method for dealing with several endpoints in a trial and a method for calculating the sample size. Their theory is based on the framework of recognizing a treatment effect when the superiority of all primary endpoints is confirmed. Such a setting of endpoints is called “co-primary endpoints”. In general, however, it is difficult to demonstrate the superiority of two or more endpoints because the power decreases as the number of endpoints increases. On the other hand, “multiple endpoints” are used in trials that recognize a treatment effect if it is superior to at least one of the endpoints. Developing a procedure that can confirm the superiority of at least one binary endpoint with latent distribution is a challenge for statisticians in the design and analysis of clinical trials. Thus, the aim of this paper was to define a testing procedure within a framework in which the efficacy of a test treatment is confirmed only when the superiority of the treatment relative to control is evidenced for at least one endpoint, and non-inferiority is demonstrated for the remaining endpoints.

For multiple continuous endpoints, Perlman and Wu [9] proposed a testing procedure that is applicable to the framework mentioned above. Nakazuru et al.

[10] proposed a more powerful testing procedure using the approximate likelihood ratio test (ALRT) defined by Glimm et al. [11]. However, there has been inadequate development of methods for multiple binary endpoints. In the same framework as this study, Ishihara and Yamamoto [12] proposed a method using multiple binary endpoints, however, they were unable to assume a latent distribution for the binary variables. A statistic for testing the superiority of binary co-primary endpoints assuming latent variables has been developed by Sozu et al. [6]. Therefore, we herein propose a testing procedure that is appropriate when all endpoints are binary and have a latent distribution. Our procedure is based on the intersection-union test (IUT), proposed by Nakazuru et al. [10], and a method by Sozu et al. [6] for estimating correlations between binary endpoints that have latent variables. In particular, we consider two statistics estimated under the null and alternative hypotheses for the test of superiority. Since there has not yet been any discussion on whether the statistics obtained under the null hypothesis or the alternative hypothesis are better in practical terms, another purpose of this study is to discuss this issue.

This article is structured as follows. In Sect. 2, we define several assumptions, including the hypothesis regarding the testing procedure and the association between correlated binary endpoints when taking latent variables into consideration. In Sect. 3, we give two IUT statistics for the superiority of at least one endpoint and the non-inferiority of the remaining endpoints when all endpoints are binary and have a latent distribution. In Sect. 4, we provide a numerical experiment using Monte Carlo simulation to illustrate the behavior of the power and type I error rate of the proposed test. Regarding the power, the proposed statistics was compared with the closed testing procedure. In Sect. 5, we provide the results of IUT based on an actual clinical trial. Finally, in Sect. 6, we summarize our findings and present concluding remarks. Based on the proposed statistics and conducted simulations, we suggest how to use the statistics in a real clinical trial.

2 Assumption and Hypotheses

2.1 Statistical Setting

In this article, we focus on a randomized clinical trial comparing p (≥ 2) endpoints with two treatment groups. There are n_1 subjects in the test group and n_2 subject in the control group. Let Y_{ijk} ($i = 1, 2; j = 1, \dots, p; k = 1, \dots, n_i$) denote the binary response variable of the j th primary endpoint of the i th treatment in the k th subject. Suppose that the vectors of binary response variables $\mathbf{Y}_{ik} = (Y_{i1k}, \dots, Y_{ipk})'$ ($i = 1, 2; k = 1, \dots, n_i$) are independently distributed as a p -variate Bernoulli distribution with $E(Y_{ijk}) = \pi_{ij}$, $V(Y_{ijk}) = \pi_{ij}(1 - \pi_{ij})$, and $\text{Corr}(Y_{ijk}, Y_{ij'k}) = \rho_{(i)jj'}$ for all $j \neq j'$, where the superscript $\{\cdot\}'$ denotes transpose. In this setting, the correlation coefficient $\rho_{(i)jj'}$ of the multivariate Bernoulli distribution is expressed as

$$\rho_{(i)ij'} = \frac{\phi_{(i)ij'} - \pi_{ij}\pi_{ij'}}{\sqrt{\pi_{ij}(1-\pi_{ij})}\sqrt{\pi_{ij'}(1-\pi_{ij'})}}, \quad (1)$$

where $\phi_{(i)ij'}$ is the joint probability of two response variables $(Y_{ijk}, Y_{ij'k})$. Note that the range of $\rho_{(i)ij'}$ is equal to or less than $(-1, 1)$ depending on the value of π_{ij} and $\pi_{ij'}$ (Bahadur [13]). That is, $\rho_{(i)ij'}$ is bounded below by

$$\max \left(-\sqrt{\frac{\pi_{ij}\pi_{ij'}}{(1-\pi_{ij})(1-\pi_{ij'})}}, -\sqrt{\frac{(1-\pi_{ij})(1-\pi_{ij'})}{\pi_{ij}\pi_{ij'}}} \right), \quad (2)$$

and above by

$$\min \left(\sqrt{\frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})}}, \sqrt{\frac{\pi_{ij'}(1-\pi_{ij})}{\pi_{ij}(1-\pi_{ij'})}} \right). \quad (3)$$

Furthermore, we assume that Y_{ik} are dichotomized random variables of continuous unobservable response $Z_{ik} = (Z_{i1k}, \dots, Z_{ipk})^t$ ($i = 1, 2; k = 1, \dots, n_i$). We also assume that Z_{ik} are independently distributed as a standardized p -variate normal distribution with $\text{Corr}(Z_{ijk}, Z_{ij'k}) = \gamma_{(i)ij'}$ for all $j \neq j'$. For each variable Z_{ik} , there is a single threshold $g_{ij} = \Phi^{-1}(1 - \pi_{ij})$ ($i = 1, 2; j = 1, \dots, p$) that partitions the latent distribution, where Φ^{-1} is the inverse function of the standard normal cumulative distribution function. Then, the binary response Y_{ijk} ($i = 1, 2; j = 1, \dots, p; k = 1, \dots, n_i$) can be defined as

$$Y_{ijk} = \begin{cases} 1, & Z_{ijk} \geq g_{ij} \\ 0, & Z_{ijk} < g_{ij}. \end{cases}$$

Set $X = (X_1, \dots, X_p)^t$ with $X_j = (\bar{Y}_{1j} - \bar{Y}_{2j})$ ($j = 1, \dots, p$), where $\bar{Y}_{ij}$ ($i = 1, 2; j = 1, \dots, p$) is the sample proportion for the j th endpoint of the i th treatment. Let the true proportion vector be the i th treatment $\pi_i = (\pi_{i1}, \dots, \pi_{ip})^t$ with difference of proportion $\Delta = (\delta_1, \dots, \delta_p)^t = \pi_1 - \pi_2$ and the covariance matrix Σ that is defined as follows:

$$\begin{aligned} \Sigma &= \Sigma^{(1)} + \Sigma^{(2)} \\ &= \frac{1}{n_1} \begin{pmatrix} \pi_{11}(1-\pi_{11}) & \cdots & \rho_{(1)1p} \sqrt{\pi_{11}(1-\pi_{11})} \sqrt{\pi_{1p}(1-\pi_{1p})} \\ \vdots & \ddots & \vdots \\ \rho_{(1)p1} \sqrt{\pi_{11}(1-\pi_{11})} \sqrt{\pi_{1p}(1-\pi_{1p})} & \cdots & \pi_{1p}(1-\pi_{1p}) \end{pmatrix} \\ &+ \frac{1}{n_2} \begin{pmatrix} \pi_{21}(1-\pi_{21}) & \cdots & \rho_{(2)1p} \sqrt{\pi_{21}(1-\pi_{21})} \sqrt{\pi_{2p}(1-\pi_{2p})} \\ \vdots & \ddots & \vdots \\ \rho_{(2)p1} \sqrt{\pi_{21}(1-\pi_{21})} \sqrt{\pi_{2p}(1-\pi_{2p})} & \cdots & \pi_{2p}(1-\pi_{2p}) \end{pmatrix}, \end{aligned}$$

where $\Sigma^{(i)} (i = 1, 2)$ is the covariance matrix of $(\bar{Y}_{i1}, \dots, \bar{Y}_{ip})^t$. Note that X is approximately normally distributed with mean Δ and covariance matrix Σ .

2.2 Hypotheses

Without loss of generality, we assume that test treatment superiority is recognized when the proportion of responses to the test treatment is greater than that to the control treatment. That is, a maximum value of $\delta_j (j = 1, \dots, p)$ greater than 0 indicates an improvement of at least one endpoint of the test treatment compared to the control treatment. In the framework dealt with in this study, a test treatment effect is recognized only when the null hypothesis for the superiority of at least one endpoint and the null hypothesis for all non-inferiority are all rejected simultaneously. In such a framework, the null hypotheses of superiority and non-inferiority are represented by a union. Therefore, we consider the combined hypothesis for the superiority of at least one endpoint and the non-inferiority of the remaining endpoints. We consider a null hypothesis H_0 and an alternative hypothesis expressed by

$$H_0 : \left\{ \max_{1 \leq j \leq p} \delta_j \leq 0 \right\} \cup \left\{ \min_{1 \leq j \leq p} (\delta_j + \epsilon_j) \leq 0 \right\} \text{ versus } H_1 : \text{not } H_0,$$

where $\epsilon_j > 0 (j = 1, \dots, p)$ is the non-inferiority margin of the j th endpoint that denotes a prespecified positive constant. Furthermore, the non-inferiority part of H_0 can be expressed as a union of null hypotheses of non-inferiority for individual endpoints, since it means that either $\delta_j + \epsilon_j (j = 1, \dots, p)$ is less than or equal to zero. Therefore, H_0 is also expressed as

$$H_0 \equiv H_0^{(0)} \cup \left\{ H_0^{(1)} \cup \dots \cup H_0^{(p)} \right\},$$

which defines the sub hypothesis of superiority “ $H_0^{(0)} : \max_{1 \leq j \leq p} \delta_j \leq 0$ ” and the sub hypothesis of non-inferiority “ $H_0^{(j)} : \delta_j \leq -\epsilon_j$ ”, for $j = 1, \dots, p$. $H_0^{(0)}$ is adaptable to the one-sided ALRT, and the IUT (Berger, 1982) [14] can be applied to test H_0 .

3 A New Test Statistics

To determine the IUT statistics, we need to estimate g_{ij} , π_{ij} and $\phi_{(i)jj'}$ considering the latent variable of Y_{ijk} . In Subsect. 3.1 below, we propose an estimation procedure for those parameters. While Sozu et al. (2010) used a sample proportion to obtain an estimator of g_{ij} , we propose a new procedure for estimating g_{ij} under the sub-null hypothesis $H_0^{(0)}$. In Subsect. 3.2, we propose a new testing procedure that extends the procedure proposed by Nakazuru et al. (2014) using the parameters estimated in Subsect. 3.1.

3.1 Proposed Estimating Procedure

For the sake of simplicity, the process of estimating parameters is divided into the following two steps.

3.1.1 Step 1. Estimating the Cut-Off Point g_{ij}

We assume that $\hat{g}_{ij}$ is the estimator of the latent cut-off point g_{ij} , and is estimated as $\hat{g}_{ij} = \Phi^{-1}(1 - \tilde{\pi}_{ij})$, where $\tilde{\pi}_{ij}$ is the maximum likelihood estimator (MLE) of the marginal probability derived by the p -variate Bernoulli distribution. Let the probability mass function of the p -variate Bernoulli distribution be

$$P(Y_{i1k} = y_{i1k}, \dots, Y_{ipk} = y_{ipk}) = \theta_{(i)0,0,\dots,0}^{\prod_{j=1}^p (1-y_{ijk})} \times \theta_{(i)1,0,\dots,0}^{y_{i1k} \prod_{j=2}^p (1-y_{ijk})} \\ \times \theta_{(i)0,1,\dots,0}^{(1-y_{i1k})y_{i2k} \prod_{j=3}^p (1-y_{ijk})} \times \dots \times \theta_{(i)1,1,\dots,1}^{\prod_{j=1}^p y_{ijk}},$$

where $\theta_{(i)0,0,\dots,0}, \dots, \theta_{(i)1,1,\dots,1}$ are joint probabilities when Y_{ik} takes values from $(0, \dots, 0), \dots, (1, \dots, 1)$, respectively, and $\theta_{(i)0,0,\dots,0} + \dots + \theta_{(i)1,1,\dots,1} = 1$. $\tilde{\pi}_{ij}$ can be expressed as $\tilde{\pi}_{ij} = \sum_{(s_1, s_2, \dots, s_p) \in S, s_j = 1} \hat{\theta}_{(i)s_1, s_2, \dots, s_p}$ using the estimator of $\theta_{(i)s_1, s_2, \dots, s_p}$, where $S = \{(s_1, s_2, \dots, s_p) | s_j = 0, 1, j = 1, \dots, p\}$ is a set whose elements consist of all pairs of response values.

In addition, $\tilde{\pi}_{ij}$ (and $\hat{g}_{ij}$) can be given in two ways depending on the estimation of $\theta_{(i)s_1, s_2, \dots, s_p}$ under the sub-null hypothesis $H_0^{(0)}$ or the sub alternative hypothesis $H_1^{(0)} : \text{not } H_0^{(0)}$. In particular, under the sub null hypothesis $H_0^{(0)}$, the Lagrange multiplier method is useful to obtain the MLE. On the other hand, under the sub-alternative hypothesis $H_1^{(0)}$, the estimator $\hat{\theta}_{(i)s_1, s_2, \dots, s_p}$ is obtained in a closed form as a sample proportion.

3.1.2 Step 2. Estimating the Joint and Marginal Probabilities

The estimator of the joint probability $\phi_{(i)jj'}$ in (1) is also given in two ways depending on $\hat{g}_{ij}$, which is obtained by the estimation of $\theta_{(i)s_1, s_2, \dots, s_p}$ constructing $\tilde{\pi}_{ij}$. $\hat{\phi}_{(i)jj'}$ can be given by

$$\hat{\phi}_{(i)jj'} = \text{Prob}(Z_{ij} \geq \hat{g}_{ij}, Z_{ij'} \geq \hat{g}_{ij'}) \\ = \int_{-\infty}^{\infty} \dots \int_{\hat{g}_{ij}}^{\infty} \dots \int_{\hat{g}_{ij'}}^{\infty} \dots \int_{-\infty}^{\infty} f(z_1, \dots, z_p; \hat{\gamma}_{(i)jj'}) dz_1 \dots dz_p \text{ for all } j \neq j',$$

where $f(z_1, \dots, z_p; \hat{\gamma}_{(i)jj'})$ is the joint density function of $\mathbf{Z}_{ik}$ and $z_1, \dots, z_p$ are random variables following the standard p -variate normal distribution wherein $\hat{\gamma}_{(i)jj'}$ is the Person's tetrachoric correlation (Pearson [15]) calculated from $(Y_{ij1}, \dots, Y_{ijn_i})$ and $(Y_{ij'1}, \dots, Y_{ij'n_i})$. Therefore, if the latency of the binary response is assumed to have a standardized multivariate normal distribution, $\hat{\phi}_{(i)jj'}$ is determined by $\hat{\gamma}_{(i)jj'}$ and the cut-off point given in Step 1. Furthermore, the estimator of the marginal probability π_{ij} constructing Σ and $\rho_{(i)jj'}$ in (1) should not be $\tilde{\pi}_{ij}$, which is obtained from

the p -variate Bernoulli distribution, so as to take into account the latent distribution function. Let $\hat{\pi}_{ij}$ denote the estimator of π_{ij} and be given by

$$\begin{aligned}\hat{\pi}_{ij} &= \text{Prob}(Z_{ij} \geq \hat{g}_{ij}) \\ &= \int_{-\infty}^{\infty} \cdots \int_{\hat{g}_{ij}}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \dots, z_p; \hat{\gamma}_{(i)jj'}) dz_1 \cdots dz_p \text{ for all } j.\end{aligned}$$

For example, with $p = 2$ endpoints, the estimator of $\phi_{(i)12}$ is written as

$$\hat{\phi}_{(i)12} = \int_{\hat{g}_{i2}}^{\infty} \int_{\hat{g}_{i1}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) dz_1 dz_2.$$

Furthermore, the marginal probabilities π_{i1} and π_{i2} are described as follows:

$$\begin{aligned}\hat{\pi}_{i1} &= \text{Prob}(Z_{i1} \geq \hat{g}_{i1}) = \int_{-\infty}^{\infty} \int_{\hat{g}_{i1}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) dz_1 dz_2, \\ \hat{\pi}_{i2} &= \text{Prob}(Z_{i2} \geq \hat{g}_{i2}) = \int_{-\infty}^{\infty} \int_{\hat{g}_{i2}}^{\infty} f(z_1, z_2; \hat{\gamma}_{(i)12}) dz_2 dz_1.\end{aligned}$$

Along with the estimation of the joint and marginal probabilities, the estimated covariance matrix $\hat{\Sigma}$ is defined as follows:

$$\begin{aligned}\hat{\Sigma} &= \hat{\Sigma}^{(1)} + \hat{\Sigma}^{(2)} \\ &= \frac{1}{n_1} \begin{pmatrix} \hat{\pi}_{11}(1 - \hat{\pi}_{11}) & \cdots & \hat{\phi}_{(1)1p} - \hat{\pi}_{11}\hat{\pi}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{\phi}_{(1)p1} - \hat{\pi}_{11}\hat{\pi}_{1p} & \cdots & \hat{\pi}_{1p}(1 - \hat{\pi}_{1p}) \end{pmatrix} \\ &\quad + \frac{1}{n_2} \begin{pmatrix} \hat{\pi}_{21}(1 - \hat{\pi}_{21}) & \cdots & \hat{\phi}_{(2)1p} - \hat{\pi}_{21}\hat{\pi}_{2p} \\ \vdots & \ddots & \vdots \\ \hat{\phi}_{(2)p1} - \hat{\pi}_{21}\hat{\pi}_{2p} & \cdots & \hat{\pi}_{2p}(1 - \hat{\pi}_{2p}) \end{pmatrix}.\end{aligned}$$

3.2 Proposed Test Statistics

We consider the following new IUT statistics to test hypothesis H_0 versus H_1 .

$$\begin{aligned}T^{(0)} &: \min(\bar{u}_A^2, \bar{u}_B^2) \text{ and} \\ T^{(j)} &: \frac{X_j + \epsilon_j}{\sqrt{\frac{\hat{\pi}_{1j}(1 - \hat{\pi}_{1j})}{n_1} + \frac{\hat{\pi}_{2j}(1 - \hat{\pi}_{2j})}{n_2}}} \text{ for } j = 1, \dots, p,\end{aligned}$$

where $\hat{\pi}_{ij}$ is the MLE of π_{ij} derived under the sub null hypothesis of non-inferiority $H_0^{(j)}$, and $T^{(j)}$ is a statistic commonly used in non-inferiority tests of binary endpoints (Farrington and Manning [16]). $T^{(0)}$ and $T^{(j)}$ are test statistics corresponding to the null hypotheses for superiority $H_0^{(0)}$ and non-inferiority $H_0^{(1)}, \dots, H_0^{(p)}$, respectively.

The proposed IUT rejects H_0 if and only if $T^{(0)} > c$ and $T^{(j)} > z_\alpha$, where c is a constant determined by

$$\sum_{j=0}^p \frac{p!}{j!(p-j)!} \frac{1}{2^p} \Pr(\chi_j^2 > c) = \alpha.$$

here, χ_j^2 denotes the χ^2 distribution with j degrees of freedom, and χ_0^2 is defined as the constant zero. α is the nominal significance level, and z_α is the upper 100α th percentile of the standard normal distribution. See Appendix for the derivation of $\bar{u}_A$ and $\bar{u}_B$.

Note that test statistics $T^{(0)}$ can be considered in two ways. One is provided by $\hat{\Sigma}$ derived under the sub-null hypothesis $H_0^{(0)}$, and the other is provided by $\hat{\Sigma}$ derived under the sub-alternative hypothesis $H_1^{(0)}$. Thus, we consider that there are also two types of IUT statistics. Let the IUT statistics using $T^{(0)}$ estimated under the sub-null hypothesis $H_0^{(0)}$ be the $T_0^{(0)}$ test type, and the those using $T^{(0)}$ estimated under the sub-null hypothesis $H_1^{(0)}$ be the $T_1^{(0)}$ test type.

4 Simulation Study

In order to evaluate the performance of the proposed IUT, we use Monte Carlo simulation to calculate the type I error rate and the powers of the $T_0^{(0)}$ and $T_1^{(0)}$ test types. In all simulations, we consider $n_1 = n_2 = 50, 100, 200$, $\epsilon_1 = \epsilon_2 = 0.2$, and $\alpha = 0.05$. The random numbers are generated from a standardized p -variate normal distribution, and response variables are obtained by dichotomizing random numbers using $\pi_i = (\pi_{i1}, \dots, \pi_{ip})^t$. The correlation between the latent variables assumes $\rho = 0, 0.4, 0.8$.

4.1 Type I Error Rate

We compare the type I error rate of the $T_0^{(0)}$ test type and the $T_1^{(0)}$ test type in the case $p = 2$. The generation of simulated data is repeated 1,000,000 times.

Table 1 shows the type I error rates for the two test types in the case $p = 2$. The type I error rate is greater than the nominal significance level for the $T_1^{(0)}$ test type when the correlation between the endpoints is zero with a large sample size. The $T_0^{(0)}$ test type is more conservative than the $T_1^{(0)}$ test type. The type I error rate is lower when π_{ij} is close to zero than when π_{ij} is 0.5. In the case where at least one of the differences in π_{ij} is less than zero and is within the non-inferiority margin, the type I error rate is less than when the difference in π_{ij} is zero, and it markedly decreases as the difference in π_{ij} becomes closer to the non-inferiority margin. The type I error rate is much smaller when there are inferior endpoints.

In particular, for the scenario where inflation is likely to occur in the case $p = 2$, we generated simulation data 100,000 times and also checked the type I error rate in the case $p = 3$.

Table 1 Type I error rates for $p = 2$

n	ρ	Type of test	$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$				
			$(0.5, 0.5),$ $(0.5, 0.5)$	$(0.1, 0.1),$ $(0.1, 0.1)$	$(0.4, 0.5),$ $(0.5, 0.5)$	$(0.35, 0.5),$ $(0.5, 0.5)$	$(0.25, 0.5),$ $(0.5, 0.5)$
50	0.8	$T_0^{(0)}$	0.042	0.028	0.023	0.011	0.0015
		$T_1^{(0)}$	0.047	0.025	0.026	0.012	0.0015
	0.4	$T_0^{(0)}$	0.038	0.021	0.011	0.004	0.0006
		$T_1^{(0)}$	0.043	0.022	0.013	0.005	0.0006
	0	$T_0^{(0)}$	0.034	0.010	0.007	0.002	0.0001
		$T_1^{(0)}$	0.038	0.010	0.008	0.002	0.0002
100	0.8	$T_0^{(0)}$	0.045	0.041	0.026	0.015	0.0010
		$T_1^{(0)}$	0.047	0.042	0.027	0.015	0.0010
	0.4	$T_0^{(0)}$	0.044	0.034	0.013	0.005	0.0003
		$T_1^{(0)}$	0.045	0.036	0.013	0.005	0.0003
	0	$T_0^{(0)}$	0.045	0.015	0.009	0.002	0.0001
		$T_1^{(0)}$	0.046	0.016	0.009	0.002	0.0001
200	0.8	$T_0^{(0)}$	0.045	0.046	0.022	0.011	0.0003
		$T_1^{(0)}$	0.046	0.047	0.022	0.011	0.0004
	0.4	$T_0^{(0)}$	0.044	0.045	0.013	0.005	0.0001
		$T_1^{(0)}$	0.045	0.046	0.013	0.006	0.0001
	0	$T_0^{(0)}$	0.049	0.033	0.012	0.003	<0.0001
		$T_1^{(0)}$	0.050	0.034	0.012	0.003	<0.0001

Table 2 Type I error rates for $p = 3$

n	ρ	Type of test	$(\pi_{11}, \pi_{12}, \pi_{13}), (\pi_{21}, \pi_{22}, \pi_{23})$	
			$(0.5, 0.5, 0.5),$ $(0.5, 0.5, 0.5)$	$(0.4, 0.5, 0.5),$ $(0.5, 0.5, 0.5)$
100	0.8	$T_0^{(0)}$	0.055	0.016
		$T_1^{(0)}$	0.057	0.016
	0.4	$T_0^{(0)}$	0.037	0.006
		$T_1^{(0)}$	0.038	0.006
	0	$T_0^{(0)}$	0.023	0.001
		$T_1^{(0)}$	0.023	0.001
200	0.8	$T_0^{(0)}$	0.072	0.020
		$T_1^{(0)}$	0.072	0.020
	0.4	$T_0^{(0)}$	0.058	0.006
		$T_1^{(0)}$	0.058	0.006
	0	$T_0^{(0)}$	0.043	0.001
		$T_1^{(0)}$	0.043	0.001

Table 2 shows the type I error rates for the two test types in the case $p = 3$. When the sample size is large and the correlations among the three endpoints is high, the type I error rate for both $T_0^{(0)}$ and $T_1^{(0)}$ is greater than the nominal significance level.

4.2 Power

We compare the powers of the $T_0^{(0)}$ and the $T_1^{(0)}$ test type for the proposed IUT, in the case $p = 2$. We also compare the power of the proposed IUT with that of a closed testing procedure that confirms the superiority of at least one of the two endpoints after the non-inferiority of both of the two endpoints is confirmed. The Bonferroni-corrected p-value (Bonferroni [17]) is used to test for superiority in the closed testing procedure. The generation of simulated data is repeated 100,000 times.

Table 3 shows the empirical powers of the proposed IUT and the closed testing procedure. The power of the $T_1^{(0)}$ test type is greater than that of the $T_0^{(0)}$ test type, and it becomes larger as the correlation between the endpoints increases with the small sample size. Even when the difference between the endpoints increases, the relationship between the power of the $T_0^{(0)}$ test type and that of the $T_1^{(0)}$ test type does not change much. On the other hand, as the sample size increases, the power of the $T_0^{(0)}$ test type becomes similar to that of the $T_1^{(0)}$ test type. Furthermore, the power of the proposed IUT is always greater than that of the closed testing procedure.

Even if the differences in π_{ij} are identical to each other, the power tends to be large when π_{ij} is close to zero, except when the correlation is zero and the sample size is small. On the other hand, if at least one endpoint is superior and the differences of all remaining endpoints are less than zero and within the non-inferiority margin, the power is lower than when all of the differences in π_{ij} are greater than zero. Furthermore, when there is an endpoint that is inferior, the power becomes quite small. The power does not monotonically increase or decrease depending on the correlation coefficient. The results for $p = 2$ and $p = 3$ in Subsect. 4.3 below show that as the number of endpoints that differ between the two groups increases, the power of the closed testing procedure is noticeably lower than that of the proposed IUT.

4.3 Power is Reduced when Non-Inferiority Test is Added to a Superiority Test

We also compare the performance of the proposed IUT and a test excluding the non-inferiority in the case of $p = 2$ and $p = 3$. The generation of simulated data is repeated 100,000 times.

Table 4 shows a power comparison between the proposed IUT and the superiority test alone for the case of $p = 2$. If all the differences in π_{ij} are greater than or equal to zero, as the sample size increases, regardless of the value of the correlation coefficient, the powers of the superiority test alone and the IUT remain similar. On the other hand, when the differences in π_{ij} between the two groups are partially within the non-inferiority margin, the power of IUT is low compared to the superiority test alone. When π_{ij} is close to zero, the powers of the superiority test alone and the IUT are more similar than when π_{ij} is close to 0.5.

Table 3 Empirical powers

n	ρ	Type of test	$(\pi_{11}, \pi_{12}), (\pi_{21}, \pi_{22})$					
			$(0.6, 0.5),$ $(0.5, 0.5)$	$(0.7, 0.5),$ $(0.5, 0.5)$	$(0.2, 0.1),$ $(0.1, 0.1)$	$(0.6, 0.6),$ $(0.5, 0.5)$	$(0.6, 0.4),$ $(0.5, 0.5)$	$(0.6, 0.25),$ $(0.5, 0.5)$
50	0.8	$T_0^{(0)}$ (proposed)	0.148	0.308	0.163	0.247	0.068	0.0022
		$T_1^{(0)}$ (proposed)	0.158	0.315	0.162	0.265	0.071	0.0023
		Closed testing	0.085	0.243	0.103	0.154	0.041	0.0018
	0.4	$T_0^{(0)}$ (proposed)	0.132	0.334	0.166	0.285	0.048	0.0023
		$T_1^{(0)}$ (proposed)	0.143	0.346	0.173	0.304	0.052	0.0025
		Closed testing	0.079	0.270	0.112	0.171	0.032	0.0017
	0	$T_0^{(0)}$ (proposed)	0.126	0.322	0.093	0.321	0.032	0.0009
		$T_1^{(0)}$ (proposed)	0.134	0.332	0.096	0.335	0.035	0.0011
		Closed testing	0.062	0.234	0.060	0.162	0.018	0.0007
	0.8	$T_0^{(0)}$ (proposed)	0.305	0.700	0.443	0.417	0.167	0.0017
		$T_1^{(0)}$ (proposed)	0.310	0.702	0.446	0.424	0.168	0.0017
		Closed testing	0.211	0.634	0.348	0.320	0.125	0.0016
100	0.4	$T_0^{(0)}$ (proposed)	0.265	0.686	0.446	0.489	0.113	0.0015
		$T_1^{(0)}$ (proposed)	0.269	0.691	0.456	0.496	0.114	0.0015
		Closed testing	0.203	0.638	0.365	0.358	0.094	0.0014
	0	$T_0^{(0)}$ (proposed)	0.268	0.684	0.275	0.575	0.082	0.0008
		$T_1^{(0)}$ (proposed)	0.271	0.689	0.282	0.581	0.082	0.0008
		Closed testing	0.186	0.618	0.217	0.379	0.066	0.0007
	0.8	$T_0^{(0)}$ (proposed)	0.536	0.955	0.780	0.665	0.315	0.0010
		$T_1^{(0)}$ (proposed)	0.538	0.955	0.782	0.669	0.317	0.0010
		Closed testing	0.402	0.949	0.708	0.546	0.302	0.0010
	0.4	$T_0^{(0)}$ (proposed)	0.493	0.960	0.780	0.757	0.269	0.0008
		$T_1^{(0)}$ (proposed)	0.497	0.961	0.784	0.760	0.275	0.0008
		Closed testing	0.402	0.951	0.718	0.605	0.250	0.0008
200	0	$T_0^{(0)}$ (proposed)	0.524	0.963	0.680	0.841	0.252	0.0008
		$T_1^{(0)}$ (proposed)	0.526	0.964	0.683	0.842	0.253	0.0008
		Closed testing	0.404	0.950	0.608	0.643	0.212	0.0007

Table 5 shows a power comparison between the proposed IUT and the superiority test alone for the case of $p = 3$. As with the case of $p = 2$, if all the differences in π_{ij} are greater than or equal to zero, adding a non-inferiority test does not reduce the power much when the sample size is large. Conversely, larger the number of endpoints for which the differences in π_{ij} are partially within the non-inferiority margin, the greater the reduction in power of the IUT.

Table 4 Evaluation of power reduction when adding the non-inferiority test for $p = 2$

n	Proportions		Type of test	Correlation		
	(π_{11}, π_{12})	(π_{21}, π_{22})		0	0.4	0.8
50	(0.6, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.126	0.132	0.148
			Only superiority	0.190	0.173	0.170
	(0.7, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.322	0.334	0.308
			Only superiority	0.545	0.511	0.425
	(0.2, 0.1)	(0.1, 0.1)	Superiority + non-inferiority	0.093	0.166	0.163
			Only superiority	0.101	0.179	0.169
	(0.6, 0.6)	(0.5, 0.5)	Superiority + non-inferiority	0.321	0.285	0.247
			Only superiority	0.358	0.302	0.251
	(0.6, 0.4)	(0.5, 0.5)	Superiority + non-inferiority	0.032	0.048	0.068
			Only superiority	0.142	0.137	0.131
	(0.6, 0.25)	(0.5, 0.5)	Superiority + non-inferiority	0.0009	0.0023	0.0022
			Only superiority	0.1257	0.0994	0.0709
100	(0.6, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.268	0.265	0.305
			Only superiority	0.313	0.286	0.311
	(0.7, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.684	0.686	0.700
			Only superiority	0.825	0.801	0.774
	(0.2, 0.1)	(0.1, 0.1)	Superiority + non-inferiority	0.275	0.446	0.443
			Only superiority	0.277	0.447	0.443
	(0.6, 0.6)	(0.5, 0.5)	Superiority + non-inferiority	0.575	0.489	0.417
			Only superiority	0.580	0.491	0.417
	(0.6, 0.4)	(0.5, 0.5)	Superiority + non-inferiority	0.082	0.113	0.167
			Only superiority	0.253	0.229	0.239
	(0.6, 0.25)	(0.5, 0.5)	Superiority + non-inferiority	0.0008	0.0015	0.0017
			Only superiority	0.2309	0.1558	0.0887
200	(0.6, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.524	0.493	0.536
			Only superiority	0.533	0.496	0.536
	(0.7, 0.5)	(0.5, 0.5)	Superiority + non-inferiority	0.963	0.960	0.955
			Only superiority	0.983	0.978	0.967
	(0.2, 0.1)	(0.1, 0.1)	Superiority + non-inferiority	0.680	0.780	0.780
			Only superiority	0.680	0.780	0.780
	(0.6, 0.6)	(0.5, 0.5)	Superiority + non-inferiority	0.841	0.757	0.665
			Only superiority	0.841	0.757	0.665
	(0.6, 0.4)	(0.5, 0.5)	Superiority + non-inferiority	0.252	0.269	0.315
			Only superiority	0.473	0.403	0.373
	(0.6, 0.25)	(0.5, 0.5)	Superiority + non-inferiority	0.0008	0.0008	0.0010
			Only superiority	0.4457	0.2663	0.1221

Table 5 Evaluation of power reduction when adding the non-inferiority test for $p = 3$

n	Proportions		Type of test	Correlation		
	$(\pi_{11}, \pi_{12}, \pi_{13})$	$(\pi_{21}, \pi_{22}, \pi_{23})$		0	0.4	0.8
50	(0.6, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.123	0.143	0.154
			Only superiority	0.241	0.209	0.174
	(0.6, 0.6, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.261	0.253	0.249
			Only superiority	0.409	0.335	0.272
	(0.6, 0.6, 0.6)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.479	0.401	0.311
			Only superiority	0.563	0.436	0.316
	(0.7, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.262	0.298	0.348
			Only superiority	0.579	0.514	0.479
	(0.2, 0.1, 0.1)	(0.1, 0.1, 0.1)	Superiority + non-inferiority	0.461	0.233	0.223
			Only superiority	0.487	0.245	0.235
	(0.6, 0.6, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.098	0.125	0.169
			Only superiority	0.346	0.313	0.326
	(0.6, 0.4, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.014	0.035	0.069
			Only superiority	0.140	0.126	0.133
100	(0.6, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.305	0.304	0.344
			Only superiority	0.370	0.333	0.349
	(0.6, 0.6, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.558	0.503	0.509
			Only superiority	0.624	0.540	0.521
	(0.6, 0.6, 0.6)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.791	0.655	0.536
			Only superiority	0.798	0.658	0.536
	(0.7, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.673	0.687	0.732
			Only superiority	0.848	0.817	0.796
	(0.2, 0.1, 0.1)	(0.1, 0.1, 0.1)	Superiority + non-inferiority	0.486	0.447	0.469
			Only superiority	0.493	0.448	0.469
	(0.6, 0.6, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.242	0.276	0.358
			Only superiority	0.568	0.532	0.558
	(0.6, 0.4, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.047	0.088	0.156
			Only superiority	0.247	0.211	0.217
200	(0.6, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.572	0.530	0.568
			Only superiority	0.581	0.533	0.568
	(0.6, 0.6, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.857	0.787	0.788
			Only superiority	0.864	0.791	0.789
	(0.6, 0.6, 0.6)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.964	0.879	0.771
			Only superiority	0.964	0.879	0.771
	(0.7, 0.5, 0.5)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.968	0.962	0.958
			Only superiority	0.985	0.977	0.966
	(0.2, 0.1, 0.1)	(0.1, 0.1, 0.1)	Superiority + non-inferiority	0.795	0.784	0.789
			Only superiority	0.795	0.784	0.789
	(0.6, 0.6, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.543	0.554	0.599
			Only superiority	0.834	0.778	0.755
	(0.6, 0.4, 0.4)	(0.5, 0.5, 0.5)	Superiority + non-inferiority	0.196	0.226	0.274

Table 5 (continued)

n	Proportions		Type of test	Correlation		
	$(\pi_{11}, \pi_{12}, \pi_{13})$	$(\pi_{21}, \pi_{22}, \pi_{23})$		0	0.4	0.8
			Only superiority	0.456	0.352	0.308

5 Numerical Example

We present the results of applying the proposed IUT to an actual trial which confirm the efficacy of 4F-PCC [4]. The clinical trial is a multicentre, randomized, open-label, phase III trial on patients aged 18 years or older needing rapid vitamin K antagonist reversal before an urgent surgical or invasive procedure. As mentioned in Sect. 1, this study includes two primary endpoints, that is, (i) the percentage of patients with a hemostatic effect defined as binary variable based on predicted blood loss, and (ii) the percentage with a decrease in the INR. The analyses were intended to evaluate, in a hierarchical fashion, first non-inferiority for both endpoints, then superiority if non-inferiority was achieved. Based on the result of the study (for details, see Fig. 3 in [4]), we consider the case of $\hat{\pi}_{11} = 0.9$ and $\hat{\pi}_{21} = 0.55$ for the percentage of the hemostatic effect, and $\hat{\pi}_{12} = 0.75$ and $\hat{\pi}_{22} = 0.1$ for the percentage of decreasing INR. The non-inferiority margin is set at 0.1 for both endpoints, and sample size of each group is $n_1 = 87$ and $n_2 = 81$. Since correlations between two variables cannot be derived from the reported results, we consider the case of $\hat{\gamma}_{(1)12} = \hat{\gamma}_{(2)12} = 0, 0.2, 0.4, 0.6, 0.8$. Since only the statistics calculated under the sub-null hypothesis $H_1^{(0)}$ from the reported information, Table 6 provides the value of $T_1^{(0)}$, $T^{(1)}$ and $T^{(2)}$ with results of IUT (rejected or accepted) at significance level $\alpha = 0.05$.

Although showing the non-inferiority for all endpoints had been the first step in the actual trial, even if the test had simultaneously taken into account showing the superiority for at least one endpoint, the same conclusion was obtained for any estimate of the correlation coefficient.

Table 6 Numerical example

n_1	n_2	$(\hat{\pi}_{11}, \hat{\pi}_{12})$	$(\hat{\pi}_{21}, \hat{\pi}_{22})$	$\hat{\gamma}_{(i)12}$	Statistic			Result
					$T_1^{(0)}$	$T^{(1)}$	$T^{(2)}$	
87	81	(0.9, 0.75)	(0.55, 0.1)	0.8	132.25	6.47	9.82	Reject $H^{(0)}$
				0.6	136.22	6.47	9.82	Reject $H^{(0)}$
				0.4	142.17	6.47	9.82	Reject $H^{(0)}$
				0.2	149.90	6.47	9.82	Reject $H^{(0)}$
				0	159.30	6.47	9.82	Reject $H^{(0)}$

6 Concluding Remarks

In this article, we developed a testing procedure for studies with multiple binary endpoints and a latent distribution. This was performed within a framework in which the efficacy of a test treatment is recognized when at least one endpoint demonstrates superiority and the remaining endpoints demonstrate non-inferiority. We derived two types of test statistics using cut-off points estimated under the sub null hypothesis $H_0^{(0)}$ and the sub-alternative hypothesis $H_1^{(0)}$, and these procedures were compared in a numerical experiment using a Monte Carlo simulation.

The numerical experiment clearly demonstrated that the $T_1^{(0)}$ test type was always more powerful than the $T_0^{(0)}$ test type. However, α -violation occurred in the $T_1^{(0)}$ test type when a sample size was large and the correlation coefficient was zero in the case $p = 2$. As the number of endpoints increased, we also found that α -violation was more likely to occur in the scenario where all differences of the endpoints were zero. On the other hand, α -violation did not occur when any of the endpoints was inferior. We believe that this will not be a fatal problem in practice because the framework of this study assumes that any of the endpoints are inferior. However, since α -violation is a serious issue in confirmatory clinical trials, it is necessary to develop a method that does not cause α -violation, or we should choose a non-inflationary test when the number of endpoints is large. Since there was not a large difference in power between the $T_0^{(0)}$ and $T_1^{(0)}$ test types, it may be reasonable to preferentially use the $T_0^{(0)}$ test type, especially if the correlation coefficients between the endpoints have not been investigated. Furthermore, this study showed a marked decrease in power as the number of differences in endpoints within the non-inferiority margin increased. Even if the number of superior endpoints was the same, power did not decrease when the number of endpoints was increased.

The power does not monotonically increase or decrease depending on the correlation coefficient between the endpoints. When determining the sample size, it is difficult to know whether the power will increase if the correlation is changed under the fixed difference in endpoints. A further non-negligible result is that type I error rates tend to increase when the correlation is low and the sample size is large. We also found that as the number of endpoints increases, the higher the correlation, the more likely the type I error rate will inflate. For a less problematic sample size determination, the correlations between endpoints should be accurately investigated before trial planning. However, in practice, it is difficult to estimate the correlation coefficients between endpoints during trial planning. Therefore, we recommend that power be simulated assuming several correlation coefficients under fixed proportions for each endpoint, and that the most conservative sample size among all scenarios be used in the trial. According to Offen et al. [18] and Sankoh et al. [19], the correlation coefficients between multiple endpoints in clinical trials are approximately equal to 0.4 and range from 0.2 to 0.8, which may help in setting up the simulation scenarios. For example, if we assume that the proportions of the two responses are 0.6 for the treatment group and 0.5 for the control group, the power will exceed 0.8 if the correlation

coefficient is zero, but the power will fall below 0.8 if the correlation coefficient is 0.4 to 0.8. Therefore, assuming a correlation of 0.8 in this case would be a conservative design.

Incidentally, like the proposed testing procedure, a closed testing procedure can be used for multiple endpoints where the familywise error rate is kept below the nominal significance level. In the framework of this study, the proposed IUT was shown to be more powerful than the closed testing procedure regardless of the correlation coefficient between endpoints, the difference between endpoints, the number of noticeably different endpoints, the sample size, and whether or not the proportions are close to zero. Although the closed testing procedure has a significant advantage in that it does not require control of the type I error rate in individual tests when there are inclusion relationships between null hypotheses, it may be more reasonable to use the proposed IUT in the framework of this study, where the superiority of at least one endpoint and the non-inferiorities of the remaining endpoints are confirmed simultaneously.

We also demonstrated a power reduction when the non-inferiority test was added to the superiority test. Our simulations showed that there was only a minimal decrease in power when the proportions of responses to the test treatment were all or somewhat higher than that to the control treatment. When the proportions were close to zero, the power was almost the same for the superiority test alone and the proposed IUT. By contrast, the power was reduced when the treatment effects were partially within the non-inferiority margin for a small sample size. In particular, the power decreased remarkably with the number of differences in the primary endpoints between the two groups that fell within the non-inferiority margin. Note that in such a situation, large sample size is needed to detect the difference. The development of more efficient methods with a higher power in such cases is required in the future.

Furthermore, the smaller the correlation coefficient, the lower the power of the proposed method in comparison to a procedure that tested only superiority. Therefore, in a primary analysis using the proposed testing procedure for certain sample sizes, assuming differences in proportions and correlations between endpoints, if all endpoints are binary and have a continuous latent distribution then it is ideal in practice to confirm not only the superiority of at least one endpoint, but also the non-inferiority of all remaining endpoints.

In the fields of economics, business, and education, a paradigm shift with methods that do not use p -values to establish statistical evidence continues to be proposed (Bhatti and Kim [20]). Currently, due to drug approval regulations, it will be unavoidable to determine the efficacy of a drug by the p value of the confirmatory trial. However, in the future, we will also need to develop methods that do not rely on p -values in order to conclude even in the exploratory phase.

Appendix

We consider an ALRT for $H_0^{(0)}$. Let A be the positive definite matrix such that $A'A = \hat{\Sigma}^{-1}$, where Σ^{-1} is the inverse matrix of Σ . The statistic $\mathbf{u}_A = (u_{A1}, \dots, u_{Ap})' = A\mathbf{X}$ is approximately distributed as a p -variate normal

distribution with mean $\mathbf{A}\Delta$ and covariance matrix $\mathbf{I}$ (the identity matrix). For simplicity, to represent $\mathbf{A}$ we use the set of eigenvectors multiplied by the square root of the corresponding eigenvalue, because $\mathbf{A}$ is not uniquely determined. Furthermore, according to the procedure of Nakazuru et al. (2014), $\mathbf{B}$ is defined as the matrix substituting the off-diagonal elements of $\mathbf{A}$ with their absolute values. Consider the two transformations such that

$$\begin{aligned}\mathbf{u}_A &\equiv (u_{A1}, \dots, u_{Ap})^t = \mathbf{A}\mathbf{X} \quad \text{and} \\ \mathbf{u}_B &\equiv (u_{B1}, \dots, u_{Bp})^t = \left(\frac{\det \mathbf{A}}{\det \mathbf{B}}\right)^{2/p} \mathbf{B}\mathbf{X}.\end{aligned}$$

In these assumptions, the ALRT rejects $H_0^{(0)}$ if and only if

$$T^{(0)} : \min(\bar{u}_A^2, \bar{u}_B^2) > c,$$

where $\bar{u}_A^2$ and $\bar{u}_B^2$ are defined by

$$\begin{aligned}\bar{u}_A^2 &= \sum_{j=1}^p \max(u_{Aj}, 0)^2, \\ \bar{u}_B^2 &= \sum_{j=1}^p \max(u_{Bj}, 0)^2.\end{aligned}$$

Acknowledgements The authors would like to sincerely thank the editor-in-chief, associate editor and referees for their valuable comments about our paper.

Funding The authors have solely funded the research by themselves.

Declarations

Conflict of interest The authors have declared no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. International Conference on Harmonization (ICH) of technical requirements for regulations of pharmaceuticals for human use: ICH Tripartite Guideline E-9 Documents (1998)
2. Smolen, J.S., Burmester, G.R., Combe, B., Curtis, J.R., Hall, S., Haraoui, B., van Vollenhoven, R., Cioffi, C., Ecoffet, C., Gervitz, L., Lonescu, L., Peterson, L., Fleischmann, R.: Head-to-head

- comparison of certolizumab pegol versus adalimumab in rheumatoid arthritis: 2-year efficacy and safety results from the randomised EXXELERATE study. *Lancet* **388**(10061), 2763–2774 (2016). [https://doi.org/10.1016/S0140-6736\(16\)31651-8](https://doi.org/10.1016/S0140-6736(16)31651-8)
3. Reich, K., Langlay, R.G., Papp, K.A., Ortonne, J.P., Unnebrink, K., Kaul, M., Valdes, J.M.: A 52-week trial comparing briakinumab with methotrexate in patients with psoriasis. *N. Engl. J. Med.* **365**(17), 1586–1596 (2011). <https://doi.org/10.1056/NEJMoa1010858>
 4. Goldstein, J.N., Refaai, M.A., Jr., Milling, T.J., Lewis, B., Goldberg-Alberts, R., Hug, B.A., Sarode, R.: Four-factor prothrombin complex concentrate versus plasma for rapid vitamin K antagonist reversal in patients needing urgent surgical or invasive interventions: a phase 3b, open-label, non-inferiority, randomised trial. *Lancet* **385**(9982), 2077–2087 (2015). [https://doi.org/10.1016/S0140-6736\(14\)61685-8](https://doi.org/10.1016/S0140-6736(14)61685-8)
 5. Marcus, R., Peritz, E., Gabriel, K.R.: On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660 (1976). <https://doi.org/10.2307/2335748>
 6. Perlman, M.D., Wu, L.: A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276–280 (2004). <https://doi.org/10.1111/j.0006-341X.2004.00159.x>
 7. Sozu, T., Sugimoto, T., Hamasaki, T.: Sample size determination in clinical trials with multiple co-primary binary endpoints. *Stat. Med.* **29**(21), 2169–2179 (2010). <https://doi.org/10.1002/sim.3972>
 8. Sozu, T., Sugimoto, T., Hamasaki, T.: Sample size determination in superiority clinical trials with multiple co-primary correlated endpoints. *J. Biopharm. Stat.* **21**(4), 650–668 (2011). <https://doi.org/10.1080/10543406.2011.551329>
 9. Sozu, T., Sugimoto, T., Hamasaki, T.: Sample size determination in clinical trials with multiple co-primary endpoints including mixed continuous and binary variables. *Biom. J.* **54**(5), 716–729 (2012). <https://doi.org/10.1002/bimj.201100221>
 10. Nakazuru, Y., Sozu, T., Hamada, C., Yoshimura, I.: A new procedure of one-sided test in clinical trials with multiple endpoints. *Jpn. J. Biom.* **35**, 17–35 (2014). <https://doi.org/10.5691/jjb.35.17>
 11. Glimm, E., Srivastava, M., Lauter, J.: Multivariate tests of normal mean vectors with restricted alternatives. *Commun. Stat. B: Simul. Comput.* **31**, 589–604 (2002)
 12. Ishihara, T., Yamamoto, K.: A testing procedure in clinical trials with multiple binary endpoints. *Commun. Stat. Theory Methods*. Advance Online Publication (2021)
 13. Bahadur, R.R.: In studies in item analysis and prediction, Stanford Mathematical Studies in the Social Sciences. In: Solomon H (ed) Stanford University Press, Stanford, pp 158–168 (1961)
 14. Berger, R.L.: Multiparameter hypothesis testing and acceptance sampling. *Technometrics* **24**, 295–300 (1982). <https://doi.org/10.1080/00401706.1982.10487790>
 15. Bonferroni, C.E.: Teoria statistica delle classi e calcolo delle probabilità, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze (1936)
 16. Farrington, C.P., Manning, G.: Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Stat. Med.* **9**, 1447–1454 (1990). <https://doi.org/10.1002/sim.4780091208>
 17. Pearson, K., III.: Mathematical contributions to the theory of evolution.—VIII. On the inheritance of characters not capable of exact quantitative measurement.—Part I. Introductory. Part II. On the inheritance of coat-colour in horses. Part III. On the inheritance of eye-colour in man. *Philos. Trans. R. Soc. A.* **195**, 1–47 (1900). <https://doi.org/10.1098/rsta.1900.0024>
 18. ...Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Muirhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J.D., Krishen, A., Liu, T., Ryder, S., Sankoh, A.J., Wang, J., Yeh, C.H.: Multiple co-primary endpoints: medical and statistical solutions. *Drug Inform. J.* **41**, 31–46 (2007). <https://doi.org/10.1177/009286150704100105>
 19. Sankoh, A.J., Huque, M.F., Russell, H.K., D'Agostino, R.B.: Global two-group multiple endpoint adjustment methods applied to clinical trials. *Ther. Innov. Regul. Sci.* **33**, 119–140 (1999). <https://doi.org/10.1177/009286159903300115>
 20. Bhatti, M.I., Kim, J.H.: Towards a new paradigm for statistical evidence in the use of p-value. *Econometrics* **9**(1), 2 (2021). <https://doi.org/10.3390/econometrics9010002>