Research Article

Statistics in Medicine

Received 16 April 2014, Accepted 27 January 2016 Published online 25 February 2016 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6911

Fallback tests for co-primary endpoints

Robin Ristl,[a] Florian Frommlet,[a] Armin Koch[b] and Martin Posch[a*†]

When efficacy of a treatment is measured by co-primary endpoints, efficacy is claimed only if for each endpoint an individual statistical test is significant at level α. While such a strategy controls the family-wise type I error rate (FWER), it is often strictly conservative and allows for no inference if not all null hypotheses can be rejected. In this paper, we investigate fallback tests, which are defined as uniform improvements of the classical test for co-primary endpoints. They reject whenever the classical test rejects but allow for inference also in settings where only a subset of endpoints show a significant effect. Similarly to the fallback tests for hierarchical testing procedures, these fallback tests for co-primary endpoints allow one to continue testing even if the primary objective of the trial was not met. We propose examples of fallback tests for two and three co-primary endpoints that control the FWER in the strong sense under the assumption of multivariate normal test statistics with arbitrary correlation matrix and investigate their power in a simulation study. The fallback procedures for co-primary endpoints are illustrated with a clinical trial in a rare disease and a diagnostic trial.

Keywords: multiple endpoints; multiple testing; diagonally trimmed Simes test; Rüger test; small populations

1. Introduction

In many settings it is not sufficient to show superiority of an experimental treatment compared with a control in a single endpoint, but multiple endpoints need to be considered to demonstrate benefit. In Alzheimer's disease, for example, the European Medicines Agency guideline [1] recommends a cognitive and a functional scale as co-primary endpoints, supported by a global assessment scale as secondary endpoint. Also in several rare diseases, evidence for several co-primary endpoints is required. For the development of medicines for Duchenne and Becker muscular dystrophy, the recent draft guidance recommends to pre-specify two co-primary endpoints from the domains motor functioning and muscle strength [2]. Furthermore, in specific settings also measures of cardiac or respiratory function are recommended as relevant co-primary endpoints. In some settings, more than two co-primary endpoints are considered. For example, in a recent study for the treatment of generalized seizures associated with Lennox–Gastaut syndrome, a rare but catastrophic pediatric epilepsy syndrome, three co-primary endpoints were defined to measure the reduction in different seizure types and seizure severity [3].

If the primary endpoints are defined as co-primary, efficacy has to be demonstrated in all endpoints to claim superiority of a treatment. To this end, each endpoint is tested by a single level α test, and the null hypothesis of no efficacy is rejected if all individual tests are significant. No adjustment of the individual significance level to correct for multiple testing is required in this setting [4]. It has been shown that requiring a proof of efficacy in several endpoints in general has a negative impact on the power of the study [5, 6]. In diseases with a high prevalence, such as Alzheimer or multiple sclerosis, this can be addressed by increasing the sample size, but in rare diseases, this may not be a feasible option.

[a] *Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[b] *Centre for Biometry, Medical Informatics and Medical Technology, Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany*
[*] *Correspondence to: Martin Posch, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria.*
[†] *E-mail: martin.posch@meduniwien.ac.at*

To improve the power of the test for multiple endpoints, it has been proposed to test the elementary endpoints at a local level larger than α [5, 7–9]. The resulting procedures do not control the family-wise type I error rate (FWER) on the whole null space: if there is no treatment effect in one endpoint but the effects in the other endpoints are very large, the FWER approaches the local level that was applied. Chuang-Stein *et al*. [7] argue that real treatment effects cannot be arbitrarily large such that it suffices to control the average type I error rate over a restricted nullspace. Kordzakhia *et al*. [9] propose a method where a combination test is applied that controls the FWER at level α over a restricted null space and at some larger level over the whole null space.

In this paper we start from the general question of how the classic co-primary endpoint test can be extended to enable inference even in situations where the primary objective is not met. What can be inferred in such a trial if only a subset of the endpoints meet the efficacy criterion? To address this question, we propose fallback tests for co-primary endpoints, defined as multiple testing procedures that have the same rejection region as the classic co-primary endpoint test for the simultaneous rejection of all null hypotheses, but allow one to reject elementary or intersection null hypotheses also if this objective is not achieved. Focusing on the setting of two or three co-primary endpoints we investigate examples of such fallback tests and study their power in a simulation study. By definition, fallback tests uniformly improve the classical and widely used co-primary endpoint test and allow for additional inference, without a penalty regarding the main objective to show a significant effect in all endpoints.

One approach to allow for inference if the main objective is not met is hierarchical testing [10, 11]. This approach, however, has the limitation that the co-primary endpoints need to be tested sequentially according to a pre-defined ordering. For the case of two and three co-primary endpoints, we propose alternative fallback testing procedures that do not rely on an ordering of the hypotheses and control the FWER in the strong sense under the assumption of multivariate normal test statistics with arbitrary covariance matrix. For the proof, we show a general result on the trivariate normal distribution which gives a uniform improvement of the Rüger test [12] of the intersection of three elementary hypotheses for multivariate normal test statistics.

A further application of the fallback tests is a procedure that has been proposed for the assessment of diagnostics. In this setting, studies with three readers that diagnose the same group of patients have been proposed. The diagnostic is considered as acceptable if the sensitivity and specificity of the diagnoses of two out of the three readers significantly exceed certain thresholds at the level α. We show that this is a valid test for the global null hypothesis that for all three readers the sensitivity or specificity (or both) lie below the threshold.

Note that the term fallback test has been used for a generalization of the hierarchical test for a primary and several ordered secondary endpoints that allows one to continue testing in the hierarchical order even if the test for the primary endpoint does not reject [13]. This comes at the cost that the individual tests cannot be performed at the full level α. Similar as the fallback test for co-primary endpoints considered here, the fallback test for primary and secondary endpoints allows for inference if the main study goal is not achieved.

The manuscript is structured as follows: In Section 2, we propose a general framework for fallback tests for co-primary endpoints and give examples of fallback tests for two and three co-primary endpoints. In Section 3, the power of the tests is investigated in a simulation study. In Section 4, we illustrate the application of fallback tests with a clinical trial in a rare disease and a diagnostic trial. Technical proofs are given in the Appendix.

2. Fallback tests for co-primary endpoints

Consider the test of n co-primary endpoints and the corresponding one-sided null hypotheses $H_1, \dots, H_n$ and let $X = (X_1, \dots, X_n)$ denote the vector of respective test statistics. The classical test for co-primary endpoints rejects all n elementary hypotheses $\left(\text{that is the union } \cup_{i=1}^{n} H_i\right)$ if $\min_{i=1,\dots,n} X_i \geqslant z_{1-\alpha}$. Otherwise, no hypothesis is rejected. Here, $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of the distribution of X_i under H_i (which we assume to be identical for all i). It is well known that this test controls the FWER at level α in the strong sense.

We define a *fallback test for co-primary endpoints* as a multiple testing procedure controlling the FWER in the strong sense that

1. rejects $\cup_{i=1}^{n} H_i$ if the vector of test statistics X lies in the rejection region of the classical co-primary endpoint test, given by $R^c = \{x \in \mathbb{R}^n : x_i \geqslant z_{1-\alpha}, \forall i \in \{1, \dots, n\}\}$,
2. and in addition allows one to reject elementary or intersection hypotheses also for some outcomes where the test statistics do not lie in R^c.

Thus, a fallback test allows one to reject (intersection) hypotheses even if not all $X_i \geqslant z_{1-\alpha}$, but has the same rejection region and power for $\cup_{i=1}^{n} H_i$ as the classic test. For the construction of such fallback tests, we employ the closed testing principle [14] to control the FWER in the strong sense at level α. Consider the family of null hypotheses $\mathcal{H} = \{H_I : I \subseteq \{1, \ldots, n\}\}$, where $H_I = \cap_{i \in I} H_i$. $\mathcal{H}$ includes the elementary hypotheses H_i and all their intersections H_I, stating that all hypotheses $H_i, i \in I$ hold. A closed test rejects an (intersection) hypothesis $H_I, I \subseteq \{1, \ldots, n\}$ at family-wise level α (i.e., controlling the FWER for the tests of the family $\mathcal{H}$) if all intersection hypotheses $H_J, J \supseteq I$ can be rejected by pre-defined tests at local level α. FWER control then follows by the closed testing principle. Note that in contrast to the local level α tests of intersection hypotheses, the closed test controls the FWER for all tests of $\mathcal{H}$, that is, including the intersection hypothesis tests.

Accordingly, we can define fallback tests by defining appropriate local level α tests for all intersection hypotheses $H_J, J \subseteq \{1, \ldots, n\}$. To this end, first note that the classical co-primary endpoint test can be written as closed test where the test of each intersection hypothesis $H_J, J \subseteq \{1, \ldots, n\}$ has rejection region $R_J^c = \{x \in \mathbb{R}^n : x_i \geqslant z_{1-\alpha}, \forall i \in J\}$. Thus, when constructing a fallback test as a closed test, the rejection regions of each intersection hypothesis H_J must contain R_J^c (requirement 1) and must be strictly larger at least for $J = \{1, \ldots, n\}$ (requirement 2).

A simple example of a fallback test is the *hierarchical test*: For this test, the n hypotheses have to be a priori ordered $H_1 \rightarrow H_2 \rightarrow \cdots \rightarrow H_n$ and tested in this sequence at local level α. Testing stops as soon as a hypothesis cannot be rejected. This hierarchical procedure rejects all H_i whenever the classical co-primary endpoint test does, but in addition allows one to reject elementary hypotheses in settings where not all test statistics exceed the critical value and is therefore a fallback test according to the above definition. Note that the hierarchical test can be written as a closed testing procedure defining rejection regions $R_J^h = \left\{x \in \mathbb{R}^n : x_{\min(J)} \geqslant z_{1-\alpha}\right\}$ for all intersection hypothesis tests. Thus, whenever the hierarchical test rejects an elementary hypothesis H_i also all intersection hypotheses H_I such that $i \in I$ are rejected by the hierarchical test.

A drawback of the hierarchical test is its dependence on an a priori ordering. If, for example, the first hypothesis test in the ordering does not reject, no other hypothesis can be tested. For the practically important cases of two and three co-primary endpoints, we discuss below alternative fallback tests that do not require to order the hypotheses.

The Hochberg [15] as well as the more powerful Hommel [16] tests are fallback tests for co-primary endpoints. This follows, because both are shortcuts of a closed test where the intersection hypotheses are tested based on the Simes inequality [17]: An intersection of n hypotheses is rejected if $p_{(k)} \leqslant k\alpha/n$ for some $k = 1, \ldots n$, where $p_{(1)} \leqslant \cdots \leqslant p_{(n)}$ denote the ordered elementary p-values. However, for multivariate normal and t-distributed test statistics, the Simes and consequently the Hochberg and Hommel tests (which are equivalent to the Simes test in the case of two endpoints) have only been shown to be conservative for non-negative correlations (and $\alpha \leqslant 0.5$) but do not control the level for arbitrary correlation structures [18–20]. To construct a fallback test for two co-primary endpoints that controls the type I error rate for arbitrary correlations, we propose a modification of the classical Simes test:

Diagonally trimmed Simes test for two hypotheses

1. If $\min(X_1, X_2) \geqslant z_{1-\alpha}$, reject both H_1 and H_2.
2. If $X_1 \geqslant z_{1-\alpha/2}$ and $X_1 + X_2 \geqslant 0$, reject H_1.
3. If $X_2 \geqslant z_{1-\alpha/2}$ and $X_1 + X_2 \geqslant 0$, reject H_2.

Theorem 1

Let $X = (X_1, X_2)$ denote a bivariate normal vector of test statistics with unit variances or a bivariate t-distributed vector of test statistics, with arbitrary correlation and mean vector (μ_1, μ_2), such that $\mu_i \leqslant 0, i \in \{1, 2\}$, if the null hypothesis H_i holds. Then the diagonally trimmed Simes test controls the family-wise type I error rate at level $\alpha \in [0, 1]$.

For the proof, see Appendix A. The definition of the bivariate t-distribution in Theorem 1 is as in [21] and applies to one-sample and two-sample t-tests for the means of two normally distributed endpoints. The diagonally trimmed Simes test is a modification of the trimmed Simes test proposed by Brannath *et al*. [21]. For $i = 1, 2$ the latter rejects H_i if either both elementary tests are significant at local level α or if the test for H_i is significant at local level $\alpha/2$ and the test for the other hypothesis is significant at local level $1 - \alpha/2$, that is, the effect estimate in the other endpoint does not point too far in the opposite direction. While the trimmed Simes test is also conservative for arbitrary correlations, its rejection region is not monotone in α such that for a specific data set, the null hypothesis may not be rejected at the

pre-set significance level but could be rejected at a lower level. This lack of α-consistency [22] makes it, for example, difficult to define p-values for such a test. The diagonally trimmed test is α-consistent and, in addition, uniformly more powerful than the trimmed Simes test by Brannath *et al*., because its rejection region is strictly larger.

For general closed testing procedures, multiplicity adjusted p-values for an elementary or intersection null hypothesis H can be defined as the maximum of the local p-values of all (intersection) hypotheses tests that contain H. Thus, to construct multiplicity adjusted p-values for the diagonally trimmed Simes test, we define the local p-value of the intersection hypothesis test on which the diagonally trimmed Simes test is based (Appendix A) by $p_{\{1,2\}} = \min\left(p_{(2)}, \max\left(2p_{(1)}, \mathbb{1}_{\{p_1+p_2>1\}}\right)\right)$, where $\mathbb{1}_{\{\cdot\}}$ denotes the indicator function and p_i ($p_{(i)}$) the (ordered) local unadjusted p-values. Then, the multiplicity adjusted p-values are given by $p_i^* = \max\left(p_i, p_{\{1,2\}}\right), i = 1, 2$ and $p^*_{\{1,2\}} = p_{\{1,2\}}$.

Next, consider the test of three co-primary endpoints. We assume that the test statistics are marginally normal distributed and the vector $X = \left(X_1, X_2, X_3\right)^T$ follows a general trivariate normal distribution with covariance matrix Σ with variances equal to 1. For all $i = 1, 2, 3$ for which the null hypothesis holds, we assume that the marginal means are 0.

A 2 out of 3 fallback tests for three co-primary endpoints

Let $\alpha \leqslant 0.5$. A fallback test for three endpoints is defined by the following procedure:

1. If $\min(X_1, X_2, X_3) \geqslant z_{1-\alpha}$, reject all elementary hypotheses H_1, H_2, H_3 and stop the procedure.
2. If there exist $i, j \in \{1, 2, 3\}$, $i \neq j$ such that $\min(X_i, X_j) \geqslant z_{1-\alpha}$, reject the intersection hypothesis $H_i \cap H_j$.
3. If $X_i \geqslant z_{1-\alpha/2}$ and there exists a $j \in \{1, 2, 3\}, i \neq j$ such that $\min(X_i, X_j) \geqslant z_{1-\alpha}$ and $\min(X_1, X_2, X_3) + X_i \geqslant 0$, reject H_i.

The classical test for three co-primary endpoints corresponds to the first step of the aforementioned procedure. Thus, the fallback test uniformly improves the classical test. Furthermore, it also allows one to make inference in settings where only two of the co-primary endpoints reach significance at the local α level. In these situations, it provides a 'proof of principle' by rejecting the intersection hypothesis, and one can conclude that at least one of the null hypotheses is false. Furthermore, an elementary hypothesis H_i is rejected by the fallback procedure, if at least two hypotheses can be rejected at local level α, H_i can be rejected at local level $\alpha/2$ and the sum of X_i and the smallest of the three test statistics is not negative. Note that in contrast to the hierarchical test, the 2 out of 3 test is not consonant such that for some outcomes, an intersection hypothesis but no elementary hypothesis may be rejected.

To show that the 2 out of 3 fallback test controls the FWER in the strong sense, we rewrite it as a closed testing procedure. For the test of the global null hypothesis, we set $R_{123} = \{x \in \mathbb{R}^3 : \text{there exist } i, j \in \{1, 2, 3\}, i \neq j \text{ such that } \min(x_i, x_j) \geqslant z_{1-\alpha}.\}$. Thus, the global intersection hypothesis is rejected if at least two of the three test statistics exceed $z_{1-\alpha}$. The rejection regions of the pairwise intersection hypotheses are given by R^s_{ij}, the rejection region of the intersection hypothesis test of the diagonally trimmed Simes test defined previously. See Appendix A for an illustration.

It is easy to see that the closed test defined by these local tests leads to the test decisions of the 2 out of 3 fallback test. It remains to show that for multivariate normal test statistics, the intersection hypothesis tests have local level α. For the pairwise intersection hypothesis tests based on the diagonally trimmed Simes test, this was shown in Theorem 1. For the test of the global null hypothesis, which rejects if at least two out of the three test statistics exceed the $1-\alpha$ quantile, type I error control follows by Theorem 2 on trivariate normally distributed random variables.

Theorem 2

Let $\alpha \leqslant 0.5$ and $X = (X_1, X_2, X_3)$ denote a trivariate normally distributed vector of marginally standard normally distributed random variables. Then the probability π that at least two of the three random variables take values greater or equal than $z_{1-\alpha}$ is bounded by α.

For special correlation structures, the result is easy to see: If the random variables are independent, the probability that at least two out of three X_i exceed the critical value is $\pi = 3\alpha^2 - 2\alpha^3$, and therefore $\pi \leqslant \alpha$, if $\alpha \leqslant 0.5$. If all three test statistics are perfectly positively correlated, the probability π is equal to α. For general correlation structures, we first show that all local extrema of π written as a function of the pairwise correlation coefficients lie on the boundary of the parameter space, where the correlation matrix is singular. Then we show that for all correlation structures on the boundary π is less or equal α. The proof is based on elementary geometric arguments and outlined in Appendix B. After the original

submission of this paper, we learned that van de Wiel *et al.* [23] in the context of model selection proved a corresponding more general result for n-dimensional multivariate normal statistics for a test that rejects if the median of the test statistics exceeds $z_{1-\alpha}$, using a result on concentration inequalities due to Massart [24].

Note that the proposed test for the global null hypothesis is a uniform improvement of the Rüger test for three hypotheses. Rüger studied tests for intersection hypotheses $H = H_1 \cap \ldots \cap H_n$ that reject if a pre-specified number k out of n individual hypotheses can be rejected at local level α and derived the upper bound $\alpha n/k$ for the type I error rate if no assumption on the joint distribution of test statistics is made [12]. Indeed, one can find dependence structures for test statistics for which this bound is sharp [25]. Thus, to control the type I error rate at level α, the Rüger test with $k = 2$ and $n = 3$ requires more stringent critical values $z_{1-2\alpha/3}$, while for trivariate normal test statistics less stringent critical values $z_{1-\alpha}$ can be applied.

To construct adjusted p-values for the 2 out of 3 fallback test, we define the local p-value for the global intersection hypothesis $p_{\{1,2,3\}} = \max\left(p_{(2)}, \mathbb{1}_{\{p_{(2)}>0.5\}}\right)$, where the indicator function accounts for the fact that Theorem 2 covers only significance levels $\leqslant 0.5$. The p-values for the intersection of two elementary hypothesis are defined as for the diagonally trimmed Simes test. Then, the multiplicity adjusted p-values for the elementary (and intersection) hypotheses $H_J, J \subseteq \{1,2,3\}$ are given by $p^*_J = \max_{I \subseteq \{1,2,3\}, J \subseteq I} p_I$.

3. Power characteristics

We investigate the power characteristics of fallback tests in the settings of two and three co-primary endpoints and consider the diagonally trimmed Simes test (for two co-primary endpoints), the 2 out of 3 fallback test (for three co-primary endpoints), and the hierarchical test with the ordering $H_1 \rightarrow H_2 (\rightarrow H_3)$. We also include the Hommel test that controls the FWER only for non-negative correlations. Note that for two and three endpoints, Hommel's test has the same rejection region for elementary hypothesis as a closed test that tests intersection hypotheses based on the Simes inequality. There is a subtle difference only in the case of three endpoints as the Hommel test does not allow for rejection of an intersection hypothesis when no elementary hypothesis was rejected. For comparison, two testing procedures that are not fallback tests (their rejection region does not contain the rejection region of the co-primary endpoint test), the Bonferroni–Holm test [26] and the resampling-based maxT step-down test [27] (which accounts for the correlation structure) are included in the simulation study.

The simulations were performed for equi-correlated normal test statistics with correlations $\rho \in \{0, 0.5, 0.75, 0.85, 0.9\}$, means $\delta_i, i = 1, 2(, 3)$ and standard deviation 1, where δ_i corresponds to the non-centrality parameter for the i-th endpoint such that, for example, in a z-test for the null hypothesis of zero mean difference between two independent groups of size n with common known variance σ_i^2, $\delta_i = \Delta_i / \left(\sigma_i \sqrt{2/n}\right)$, where Δ_i denotes the assumed mean difference. All considered tests control the FWER under these assumptions. All tests were performed at the nominal one-sided family-wise level $\alpha = 0.025$. The simulation was implemented using the package mvtnorm in R [28, 29]. In the simulations for the maxT test a sample size of 50 per group was assumed. For the hierarchical test, several power definitions co-incide: The power to reject H_1 is equal to the power to reject at least one elementary null hypothesis and equal to the power to reject at least one pairwise intersection hypothesis. It is 85.1% if $\delta_1 = 3$, 51.6% if $\delta_1 = 2$ and 2.5% if $\delta_1 = 0$. Because the hierarchical test is a fallback test, the power to reject all elementary hypotheses is the same as that of the other fallback tests investigated.

In the succeeding discussions, we refer to the power to reject all null hypotheses simultaneously as conjunctive power and the power to reject at least one elementary null hypothesis as disjunctive power [30].

3.1. Power characteristics of tests for two co-primary endpoints

For the setting of two co-primary endpoints, simulation results are shown in Table I, where the power to reject both elementary hypotheses, H_1 (which is the first hypothesis in the hierarchical test), H_2 or at least one elementary hypothesis is given.

In the considered scenarios, the diagonally trimmed Simes test and the Hommel test perform almost identically. For low correlations, all investigated tests have similar power characteristics. The disjunctive power of the diagonally trimmed Simes test is larger or equal (up to simulation error) than the power of the

Table I. Power in settings with two endpoints. Power to reject both elementary hypotheses ($H_1 \cup H_2$), specifically H_1, H_2, or at least one elementary hypothesis (any H_i) under alternatives with standardized effects $\delta = (\delta_1, \delta_2)$, assuming bivariate normal test statistics with variances equal to 1 and correlation ρ and global one-sided level $\alpha = 0.025$. The power is given in percent (100,000 simulation runs per scenario). The column Trimmed Simes refers to the diagonally trimmed Simes test.

δ	ρ	Test	$H_1 \cup H_2$	H_1	H_2	any H_i
(3, 0)	0	Trimmed Simes/Hommel	2.1/2.1	77.8/77.9	2.3/2.3	78.0/78.1
		Bonferroni–Holm/maxT	2.0/1.9	77.8/76.5	2.2/2.1	78.0/76.8
	0.5	Trimmed Simes/Hommel	2.5/2.5	77.6/77.7	2.5/2.5	77.6/77.7
		Bonferroni–Holm/maxT	2.5/2.5	77.7/77.3	2.5/2.5	77.7/77.3
	0.75	Trimmed Simes/Hommel	2.5/2.5	77.8/77.9	2.5/2.5	77.8/77.9
		Bonferroni–Holm/ maxT	2.5/2.5	77.9/78.5	2.5/2.5	77.9/78.5
	0.85	Trimmed Simes/Hommel	2.5/2.5	77.4/77.4	2.5/2.5	77.4/77.4
		Bonferroni–Holm/maxT	2.5/2.5	77.4/79.1	2.5/2.5	77.4/79.1
	0.9	Trimmed Simes/Hommel	2.5/2.5	77.8/77.8	2.5/2.5	77.8/77.8
		Bonferroni–Holm/maxT	2.5/2.4	77.8/79.7	2.5/2.4	77.8/79.7
(2, 3)	0	Trimmed Simes/Hommel	44.0/44.0	50.1/50.1	81.4/81.4	87.4/87.4
		Bonferroni–Holm/maxT	43.2/41.7	49.2/48.1	80.5/79.3	86.6/85.7
	0.5	Trimmed Simes/Hommel	48.3/48.3	50.2/50.2	80.0/80.0	81.9/81.9
		Bonferroni–Holm/maxT	47.5/47.0	49.4/49.1	79.2/78.9	81.1/81.0
	0.75	Trimmed Simes/Hommel	50.4/50.4	50.8/50.8	78.8/78.8	79.2/79.2
		Bonferroni–Holm/maxT	49.7/49.6	50.1/50.1	78.1/79.3	78.5/79.8
	0.85	Trimmed Simes/Hommel	51.3/51.3	51.4/51.4	78.2/78.2	78.3/78.3
		Bonferroni–Holm/maxT	50.9/50.3	51.0/50.4	77.8/79.9	77.8/80.1
	0.9	Trimmed Simes/Hommel	51.6/51.6	51.6/51.6	78.0/78.0	78.1/78.1
		Bonferroni–Holm/maxT	51.3/50.5	51.3/50.6	77.8/80.3	77.8/80.3
(3, 3)	0	Trimmed Simes/Hommel	72.5/72.5	84.1/84.1	83.9/83.9	95.5/95.5
		Bonferroni–Holm/maxT	71.9/70.3	83.5/82.4	83.3/82.4	94.9/94.5
	0.5	Trimmed Simes/Hommel	75.9/75.9	83.3/83.3	83.2/83.2	90.6/90.6
		Bonferroni–Holm/maxT	75.1/74.0	82.4/81.7	82.4/81.9	89.7/89.6
	0.75	Trimmed Simes/Hommel	78.7/78.7	83.0/83.0	83.0/83.0	87.4/87.4
		Bonferroni–Holm/maxT	77.6/76.8	81.9/81.9	81.9/81.8	86.3/86.9
	0.85	Trimmed Simes/Hommel	79.9/79.9	82.8/82.8	82.8/82.8	85.6/85.6
		Bonferroni–Holm/maxT	78.5/78.5	81.4/82.3	81.3/82.2	84.2/86.1
	0.9	Trimmed Simes/Hommel	81.2/81.2	83.1/83.1	83.1/83.1	85.0/85.0
		Bonferroni–Holm/maxT	79.4/79.5	81.3/82.5	81.3/82.4	83.1/85.4

Bonferroni–Holm test in all scenarios. For large correlations, the fallback tests gain some advantage over the Bonferroni–Holm test with respect to the disjunctive power because they reject if both test statistics exceed $z_{1-\alpha}$, while the Bonferroni–Holm test requires one test statistic to exceed the larger threshold $z_{1-\alpha/2}$. The maxT test shows similar characteristics as the Bonferroni–Holm test; however, it becomes more powerful to reject at least one hypothesis for large correlations, and it can be superior by up to 2 percentage points over the trimmed Simes test if in addition the effects in the two endpoints are different. However, for most scenarios, the fallback tests have larger disjunctive power than the Bonferroni–Holm or the maxT test. For the hierarchical test, this holds only if the there is a treatment effect in the first endpoint in the hierarchy.

In the setting where there is a strong treatment effect in one but a lower effect in the other endpoint, the conjunctive power drops considerably compared with the case where there is a strong effect in both endpoints. However, the disjunctive power ranges still between 78.1% and 87.4% (depending on the correlation) for the diagonally trimmed Simes test. For the non-hierarchical tests, the disjunctive power is decreasing, and the conjunctive power is increasing with increasing correlation between the two endpoints.

The disjunctive power for the hierarchical test depends only on the effect of the first endpoint in the hierarchy. In the scenario with $\delta_1 = 2$ and $\delta_2 = 3$, the non-hierarchical tests have far greater disjunctive power than the hierarchical test. For $\delta_1 = \delta_2 = 3$, they have greater disjunctive power (with a difference up to 10 percentage points) than the hierarchical test when the correlation is small to moderate. Only for large correlations the hierarchical test has slightly more disjunctive power compared with the other tests.

However, the power to reject specifically H_1 is always largest for the hierarchical test. Note that for the hierarchical test, the power to reject H_2 is the same as the power to reject $H_1 \cup H_2$. It follows that the diagonally trimmed Simes test has larger power to reject H_2 than the hierarchical test, because both are fallback tests.

3.2. *Power characteristics of tests for three co-primary endpoints*

For the setting of three co-primary endpoints, simulation results are shown in Table II. Similar to the setting for two co-primary endpoints, the conjunctive power increases with increasing correlation. It is equal for all fallback tests and lower for the Bonferroni–Holm test and the maxT test. The difference in power is however moderate, between 2 and 3 percentage points for most scenarios.

By definition the Hommel test is for all alternatives at least as powerful as the Bonferroni–Holm test [16]. In the scenarios for three endpoints covered in our simulation study, the improvement of

Table II. Power in settings with three endpoints. Power to reject all elementary hypotheses ($\cup_{i=1}^{3} H_i$), at least one union of two hypotheses (any $H_i \cup H_j$), at least one elementary hypothesis (any H_i), at least one pairwise intersection hypothesis (any $H_i \cap H_j$), or H_1 under alternatives with standardized effects $\delta = (\delta_1, \delta_2, \delta_3)$, assuming trivariate normal test statistics with variances equal to 1 and equal correlations ρ and global one-sided level $\alpha = 0.025$. The power is given in percent (100,000 simulation runs per scenario).

δ	ρ	Test	$\cup_{i=1}^{3} H_i$	any $H_i \cup H_j$	any H_i	any $H_i \cap H_j$	H_1
(3,0,0)	0	2 out of 3 Fallback/Hommel	0.1/0.1	1.9/1.9	4.0/73.4	4.2/73.4	3.8/73.0
		Bonferroni–Holm/maxT	0.0/0.0	1.9/1.9	73.3/72.0	73.3/72.0	72.9/71.5
	0.5	2 out of 3 Fallback/Hommel	0.5/0.5	2.4/2.4	4.5/72.8	4.6/72.8	4.5/72.8
		Bonferroni–Holm/maxT	0.4/0.4	2.3/2.4	72.8/72.8	72.8/72.8	72.8/72.8
	0.75	2 out of 3 Fallback/Hommel	0.9/0.9	2.2/2.2	4.1/72.7	4.1/72.7	4.1/72.7
		Bonferroni–Holm/maxT	0.8/0.9	2.1/2.5	72.7/75.0	72.7/75.0	72.7/75.0
	0.85	2 out of 3 Fallback/Hommel	1.3/1.3	2.1/2.1	3.8/72.6	3.8/72.6	3.8/72.6
		Bonferroni–Holm/maxT	1.1/1.2	1.9/2.5	72.6/76.5	72.6/76.5	72.6/76.5
	0.9	2 out of 3 Fallback/Hommel	1.5/1.5	2.1/2.1	3.5/72.8	3.5/72.8	3.5/72.8
		Bonferroni–Holm/maxT	1.2/1.3	1.8/2.5	72.8/77.3	72.8/77.3	72.8/77.3
(3,3,0)	0	2 out of 3 Fallback/Hommel	1.9/1.9	60.8/60.9	72.4/93.2	73.1/93.2	66.5/77.0
		Bonferroni–Holm/maxT	1.7/1.6	60.4/58.3	92.6/91.8	92.6/91.8	76.5/74.9
	0.5	2 out of 3 Fallback/Hommel	2.5/2.5	65.5/65.5	75.0/87.3	75.9/87.3	70.2/76.4
		Bonferroni–Holm/maxT	2.5/2.4	65.2/64.5	86.6/86.5	86.6/86.5	75.9/75.5
	0.75	2 out of 3 Fallback/Hommel	2.5/2.5	69.1/69.1	77.4/83.3	78.6/83.3	73.3/76.1
		Bonferroni–Holm/maxT	2.5/2.5	68.7/69.7	82.3/83.9	82.3/83.9	75.4/76.8
	0.85	2 out of 3 Fallback/Hommel	2.4/2.4	70.9/70.9	78.4/81.2	79.8/81.2	74.6/76.1
		Bonferroni–Holm/maxT	2.4/2.5	70.3/72.8	79.9/83.3	79.9/83.3	75.2/78.1
	0.9	2 out of 3 Fallback/Hommel	2.5/2.5	72.3/72.3	79.0/80.2	80.9/80.2	75.8/76.4
		Bonferroni–Holm/maxT	2.5/2.5	71.6/74.5	78.7/82.9	78.7/82.9	75.3/78.7
(2,3,3)	0	2 out of 3 Fallback/Hommel	37.3/37.3	75.9/75.9	85.0/95.8	85.5/95.8	47.6/48.3
		Bonferroni–Holm/maxT	35.7/34.2	74.1/72.4	95.1/94.5	95.1/94.5	46.7/45.5
	0.5	2 out of 3 Fallback/Hommel	45.9/45.9	72.1/72.1	79.9/88.2	80.9/88.2	49.0/49.1
		Bonferroni–Holm/maxT	44.4/43.4	70.3/69.9	87.2/87.3	87.2/87.3	47.7/47.0
	0.75	2 out of 3 Fallback/Hommel	50.1/50.1	71.9/71.9	78.9/83.9	80.1/83.9	50.7/50.7
		Bonferroni–Holm/maxT	48.9/48.4	70.4/72.1	82.7/84.7	82.7/84.7	49.5/49.3
	0.85	2 out of 3 Fallback/Hommel	51.3/51.3	72.4/72.4	79.0/81.5	80.4/81.5	51.5/51.4
		Bonferroni–Holm/maxT	50.5/49.8	71.0/73.6	80.1/83.5	80.1/83.5	50.6/50.0
	0.9	2 out of 3 Fallback/Hommel	51.7/51.7	72.8/72.8	79.1/80.2	80.9/80.2	51.7/51.7
		Bonferroni–Holm/maxT	51.2/50.5	71.7/75.4	78.6/83.5	78.6/83.5	51.2/50.6
(3,3,3)	0	2 out of 3 Fallback/Hommel	61.4/61.4	88.5/88.5	93.7/98.4	94.0/98.4	81.3/82.8
		Bonferroni–Holm/maxT	60.0/58.1	86.9/85.6	97.9/97.7	97.9/97.7	81.7/80.6
	0.5	2 out of 3 Fallback/Hommel	69.5/69.5	83.0/83.0	88.3/92.5	89.0/92.5	80.2/81.6
		Bonferroni–Holm/maxT	67.8/66.4	81.1/80.8	91.6/91.6	91.6/91.6	80.1/79.6
	0.75	2 out of 3 Fallback/Hommel	74.5/74.5	81.2/81.2	85.7/87.8	86.9/87.8	80.5/81.2
		Bonferroni–Holm/maxT	72.2/71.8	78.7/80.2	86.2/88.1	86.2/88.1	79.1/80.1
	0.85	2 out of 3 Fallback/Hommel	77.2/77.2	81.2/81.2	84.8/85.7	86.2/85.7	81.0/81.4
		Bonferroni–Holm/maxT	74.4/74.6	78.1/80.3	83.6/86.5	83.6/86.5	78.7/80.3
	0.9	2 out of 3 Fallback/Hommel	78.5/78.5	80.9/80.9	84.0/84.0	85.7/84.0	81.2/81.2
		Bonferroni–Holm/maxT	75.2/76.6	77.4/80.9	81.3/85.9	81.3/85.9	78.0/81.2

the Hommel test over the Bonferroni–Holm test is up to 3.5 percentage points, for the scenarios with uncorrelated test statistics the improvement is up to 1.8 points.

In settings where the alternative holds for only one endpoint, the conjunctive power is below the significance level for all considered testing procedures because all tests control the FWER in the strong sense in the studied settings. However, the Hommel, Bonferroni–Holm, maxT, and the hierarchical test (the latter only if the hypothesis for which the alternative holds is the first in the hierarchical order) have a substantial power to reject the elementary hypothesis for which the alternative holds. The 2 out of 3 fallback test has a power below 5% in this setting as it requires a significant effect in two endpoints.

If the alternative holds for exactly two endpoints, the conjunctive power is again bounded by the significance level for all testing procedures. In this setting, the power to reject any elementary or pairwise intersection null hypothesis is increasing with the correlation for the 2 out of 3 fallback test but decreasing for the Hommel, the Bonferroni–Holm, and the maxT test.

While for uncorrelated test statistics the Bonferroni–Holm and the Hommel test are considerably more powerful to reject any elementary or intersection hypothesis than the 2 out of 3 fallback test, this relation is changed for strong correlations where the 2 out of 3 fallback test has a small advantage over the Bonferroni–Holm test and performs very similar to the Hommel test. The maxT test is less affected by increasing correlation and for large correlations is superior to the other tests. The performance of the hierarchical test depends only on the effect size of the first endpoint in the hierarchical order. If the null hypothesis for this endpoint holds, the power to reject an elementary (and thus an intersection) null hypothesis is only α; if the alternative with effect size $\delta_1 = 3$ holds for the first endpoint, the hierarchical test is superior only if the correlation between endpoints is large. However, as expected, the hierarchical test has the largest power for rejecting specifically the first hypothesis in the hierarchy.

The power to reject the union of two null hypotheses is similar for all studied non-hierarchical tests. For the hierarchical test, the power to reject at least two out of three null hypotheses in the three endpoint setting is the same as the power to reject both null hypotheses using any fallback test in the two endpoint scenarios (Table I). If there is a treatment effect in the first two but not in the third endpoint, the hierarchical test is more powerful than the other tests.

In the setting where for all three endpoints the alternative holds, the power characteristics of the investigated non-hierarchical tests for rejecting an elementary hypothesis or an intersection hypothesis are similar to the setting where two endpoints show an effect. However, in these settings, the 2 out of 3 fallback test is more powerful than the hierarchical test to reject any intersection hypothesis for all considered correlations. The improvement compared with the hierarchical test is substantial in the scenario where the effect size of the first endpoint in the hierarchy is low.

In the investigated scenarios where the alternative holds for all three endpoints, the power to reject at least two elementary null hypotheses simultaneously is identical (up to simulation error) for the 2 out of 3 fallback test and the Hommel test. For small correlations or homogeneous effects in all endpoints, this power is larger than that of the maxT test. The power of the hierarchical test for rejecting at least two null hypotheses is substantially lower than for the other tests unless the correlation is large.

To summarize, the fallback tests preserve the conjunctive power of the classical test for co-primary endpoints and have substantial power to reject some (intersection) null hypothesis. The Bonferroni–Holm test and even more the maxT test can have some advantage when the main aim is to reject at least one elementary null hypothesis. However, both the 2 out of 3 fallback test and the Hommel test have for most scenarios better power to reject at least two out of three hypotheses. The latter, however, has been shown to control the FWER for non-negative correlations only.

4. Clinical trial applications

4.1. A clinical trial in Lennox–Gastaut syndrome

Lennox–Gastaut syndrome is a rare pediatric epilepsy syndrome, which is characterized by multiple types of seizures, high seizure frequency, and a high rate of seizure related injury. Glauser *et al*. [3] describe a randomized controlled trial for the treatment of Lennox–Gastaut syndrome using the anti-epilepticum rufinamide. Three co-primary endpoints were specified for this trial, the percent change from baseline in total seizure frequency per 28 days, the percent change from baseline in the sum of tonic and atonic seizures per 28 days and an evaluation score rating the global improvement in seizure severity. All three co-primary endpoints were compared between the treatment and the placebo group using the Wilcoxon

rank sum test. A total of 123 patients completed the trial. Given the asymptotic normality of the rank sum statistic, the application of the 2 out of 3 fallback test is justified.

This trial serves as an example for a disease setting in which the fallback test can be applied to increase the number of possible conclusions without any additional cost. As numerical example assume that one-sided p-values $p_1 = 0.01, p_2 = 0.02$, and $p_3 = 0.03$ were observed. In this small sample scenario, the 2 out of 3 fallback test would allow one to reject H_1 at family-wise level $\alpha = 0.025$. The adjusted p-values for the elementary hypotheses tests in the 2 out of 3 fallback procedure are $p_1^* = 0.02, p_2^* = 0.03, p_3^* = 0.03$. The classic test for three co-primary endpoints and also the Bonferroni–Holm test could not reject any null hypothesis in this example.

4.2. An application to diagnostic trials

As a second application of the 2 out of 3 fallback test procedure, we consider the validation of medical diagnostic tools. For diagnostic procedures, such as many imaging tools, which involve human judgment, the European Medicines Agency advises to test the tool with more than one reader [31]. As a consequence, studies including three readers are often performed that diagnose the same group of patients. Then for each of the readers, sensitivity and specificity are estimated as the sample proportions of correctly classified diseased or healthy patients. For each reader $i = 1, 2, 3$, the null hypotheses $H_{se,i} : q_{se,i} = \gamma_{se}, H_{sp,i} : q_{sp,i} = \gamma_{sp}$ are tested against one-sided alternatives, where $q_{se,i}, q_{sp,i}$ denote reader i's sensitivity and specificity, respectively, and γ_{se}, γ_{sp} some pre-specified thresholds. As there is an inherent trade-off between sensitivity and specificity, it is concluded that the ratings of a specific reader i meet the quality requirements only if both $H_{se,i}$ and $H_{sp,i}$ are rejected, that is, if the hypothesis $H_i = H_{se,i} \cup H_{sp,i}$ can be rejected. The hypotheses H_1, H_2, H_3 can be rejected at local level α if $T_i = \min\left(Z_{se,i}, Z_{sp,i}\right) \geqslant z_{1-\alpha}$, where $Z_{se,i}, Z_{sp,i}$ denote the one sample z-test statistics for proportions. To adjust for multiplicity, we apply the 2 out of 3 fallback test to the test statistics T_i. By a Corollary to Theorem 2 (see Appendix C for the technical details), this test asymptotically controls the FWER even though the T_i are not multivariate normal.

The 2 out of 3 fallback test allows one to conclude that the diagnostic tool meets the quality requirements for all three readers if all six z-statistics exceed the critical level. If the test statistics for only two readers meet the criterion, one can conclude that the tool satisfies the criterion for at least one of the two and if the rejection rules of step 3 of the procedure are met, it is possible to conclude which of the two readers was successful or if both were successful. Such decision procedures have been proposed to EMA in the past, and Theorem 2 allows one to quantify the level of evidence such a testing procedure provides. Tests for more readers can in principle be constructed following the scheme outlined in Section 2 using a test with rejection regions for each intersection hypothesis H_I that contain R_I^c in a closed testing procedure.

The example illustrates that the 2 out of 3 fallback test can be applied also in cases where the normality assumption for the test statistics is not met, but a dominating multivariate normal distribution for the test statistics exists.

5. Discussion

In clinical trials with co-primary endpoints, the intended claim of efficacy in all endpoints cannot be made, if not all corresponding elementary hypotheses tests can be rejected at the local significance level α. Even in such a failed trial, it is of interest to make best use of the collected data by making at least partial claims on the efficacy in some of the endpoints, while controlling the overall type I error rate. However, performing additional multiple testing procedures post hoc when the original co-primary endpoint test did not reject will in general inflate the FWER. In contrast, fallback tests have the same rejection region for the test regarding the main trial objective (proof of efficacy in all co-primary endpoints) but allow one to reject elementary or intersection hypotheses also if this objective is not achieved, while controlling the FWER in the strong sense. Thus, fallback tests are uniform improvements of the classical co-primary endpoint tests. They allow one to perform the classical co-primary endpoint test but also to test additional (weaker) claims in a trial with multiple endpoints.

A general fallback test, which is applicable for any number of endpoints and without any assumptions on the dependence structure of the test statistics, is the hierarchical test. A limitation of the hierarchical test is the requirement to test the endpoints sequentially according to a pre-defined ordering. If a hypothesis cannot be rejected, all hypotheses later in the ordering cannot be tested anymore. For the setting

of two and three co-primary endpoints, we investigated fallback tests that do not require to specify an ordering and control of the FWER for multivariate normal test statistics.

For two co-primary endpoints, we propose the diagonally trimmed Simes test as a fallback test with FWER control. This test is equivalent to a procedure where one performs a co-primary endpoint test first, and in the case of a negative outcome switches to a slightly modified Bonferroni test. For three co-primary endpoints, we propose a 2 out of 3 fallback test that rejects a pairwise intersection null hypothesis if two of the hypotheses can be rejected at local level α. One can then conclude that for at least one of these two hypotheses, the alternative hypothesis holds. If one of the two hypotheses can be additionally rejected at local level $\alpha/2$ and the effect size of the remaining endpoint does not indicate a strong detrimental effect also the corresponding elementary hypothesis can be rejected.

The assumption of multivariate normality of the test statistics is satisfied in many testing scenarios, at least asymptotically, due to the multivariate central limit theorem [32]. The diagonally trimmed Simes test for two endpoints has FWER control also under the assumption of multivariate t-distributed statistics (see [21] and Appendix A). However, so far, we were not able to prove a corresponding result for the 2 out 3 fallback test.

By definition, the fallback tests preserve the conjunctive power of the classical co-primary endpoint test and therefore have optimal power for the aim to reject all elementary null hypotheses simultaneously. However, this power still may be low if sample sizes are small. When following a fallback strategy, the next best result is to reject at least all but one of the elementary hypotheses. In the setting of three endpoints, a further fallback step is to reject a single elementary hypothesis or at least the global intersection null hypothesis, which still may provide a valuable proof of principle of drug efficacy.

As the fallback tests uniformly improve the classic co-primary endpoint tests by allowing for additional rejections of elementary or intersection hypotheses, they should be preferred over the classical co-primary endpoint test. But should they also be preferred over non-fallback tests as the Bonferroni or the maxT tests? By construction, fallback tests have a larger conjunctive power. For two endpoints, the diagonally trimmed Simes test outperforms the Bonferroni–Holm test also regarding the disjunctive power in all considered scenarios. This is true also for the maxT test unless correlations are large and in addition the effect sizes differ across endpoints. In the setting of three endpoints, the trade-off of conjunctive versus disjunctive power will determine the choice of the test. Because the disjunctive power of the Bonferroni Holm and maxT test may be substantially larger than the power of the 2 out of 3 test and comes at a comparably small cost in conjunctive power, the latter tests are preferable if the disjunctive power is important, especially in the setting where only for one null hypothesis the alternative holds. In contrast, if rejection of all (or at least two out of three) null hypotheses is the predominating goal, the 2 out of 3 test should be preferred. It has not only uniformly larger conjunctive power but also larger power to reject at least two null hypotheses, with the exception of the maxT test in scenarios where both the correlation is large and effect sizes differ across endpoints.

As the correlation of endpoints is typically unknown, we derived fallback tests that control the FWER for multivariate normal test statistics with arbitrary correlation structures. Note that even if the correlations were known, the conjunctive power of the co-primary endpoint and fallback tests cannot be improved by applying relaxed critical boundaries. This holds because the critical boundaries cannot fall below the $1-\alpha$ quantile of the normal distribution in order to control the FWER in the strong sense. Under the assumption of known correlations (or large sample sizes such that the correlation can be reliably estimated), one could, however, relax critical values in intersection hypotheses tests of the fallback tests and thus, for example, increase the disjunctive power of the tests. For the application to smaller samples, a promising approach to test intersection hypotheses may be the application of non-parametric permutation-based fallback test to account for the unknown dependence structure of test statistics.

For settings where also tests for secondary endpoints need to be included in the confirmatory testing strategy, the fallback tests can be extended to control the FWER for the resulting larger family of tests. For example, such a test can be constructed following a hierarchical testing strategy and defining (in addition to the fallback test for the co-primary endpoints) a multiple testing procedure that controls the FWER for the secondary endpoints only. If the tests of the secondary endpoints are only performed if all co-primary hypotheses can be rejected, the resulting overall procedure controls the FWER for the total family of tests (co-primary and secondary).

Estimation after testing co-primary endpoints has received little attention so far. It is easy to see that classical level $1-\alpha$ confidence bounds, computed after the co-primary endpoint test rejected all hypotheses, do not have a simultaneous coverage probability of $1-\alpha$. Applying results on simultaneous confidence intervals for closed tests [33, 34], simultaneous confidence bounds for the classical co-primary

endpoint test can be defined: If the co-primary endpoint test rejects all null hypotheses, the bounds are given by the maximum of the respective Bonferroni-adjusted confidence bounds and the parameter values tested in the elementary null hypotheses. Otherwise, all bounds are set to $-\infty$. Simultaneous confidence bounds for fallback tests can be derived with the partitioning principle following the approach in [35]. For the hierarchical test, they have been proposed in [33], and simultaneous confidence bounds for the other considered fallback tests are a topic of further research. However, as many simultaneous confidence intervals for closed testing procedures, the confidence intervals for the co-primary endpoint test, the hierarchical test, and other fallback tests may not be informative, that is, for some outcomes, they may not give additional information than the rejection of the null hypotheses.

In clinical trials for small populations, it may be justified to relax the requirement to control the FWER in the strong sense. If we modify the 2 out of 3 fallback test such that an elementary null hypothesis H_i can be rejected if H_i and one further null hypothesis can be rejected at local level α, the resulting procedure will control the k-FWER rate at level α for $k = 2$, that is, the probability to erroneously reject two or more null hypothesis [36]. Furthermore, the procedure will still have weak FWER control at level α and strong FWER control at level 2α. This can be seen by writing the procedure as closed test where the global null hypothesis is tested as in the 2 out of 3 fallback test, but the pairwise intersection hypotheses $H_i \cap H_j, i \neq j$ are rejected if either X_i or X_j exceed $z_{1-\alpha}$.

k-out-of-n type procedures have also been proposed for gatekeeping tests of ordered families of hypotheses [37], but these differ from the 2 out of 3 fallback tests considered here. In the gatekeeping procedures, k out of n hypotheses in the primary family must be rejected by a multiple test that satisfies a so-called k-separability condition before hypotheses in the next family can be tested. In contrast, in the 2 out of 3 fallback test, a necessary condition to reject any elementary null hypothesis at multiple level α is that two hypotheses tests must be significant at local level α.

Formal proof of efficacy in phase III drug development usually requires that a pre-specified null hypothesis is rejected to demonstrate that an experimental drug is able to introduce a well-described treatment effect under well-defined clinical conditions, and a well-defined dose and mode of administration. In some instances, two, and in only rare instances, three co-primary endpoints are required to describe efficacy of a treatment. In these instances, all three null hypotheses would require to be rejected, because all variables cover important aspects of treatment benefit. Failure to reject one of these would be understood as leaving doubt that all aspects of treatment efficacy have been appropriately demonstrated.

The situation changes dramatically if treatments for rare, and in some instances, very rare diseases are under investigation. European legislation on orphan medicinal products (regulation (EC) no. 141/2000) states already in the introduction that patients suffering from rare conditions should be entitled to the same quality of treatment as other patients (Introduction, paragraph (2)) and continues that patients with such (i.e., orphan) conditions deserve the same quality, safety, and efficacy in medicinal products as other patients (Introduction, paragraph (7)). As a logical consequence, it is further stated that orphan medicinal products should therefore be submitted to the normal evaluation process (ibid). This implies that drugs for orphan diseases should be tested and evaluated according to the same standards that apply to drugs in the treatment of more frequent conditions. This leads to the well-known obstacles that the required sample size for proper assessment of a drug in a certain indication depends on the size of the treatment effect and variability of the outcome, but not on the rarity of the disease.

As soon as sample size is limited, difficulties will arise, if for orphan medicines the same amount of evidence is mandated as for 'normal' drug development. For obvious reasons, emphasis must be put on high-quality study designs to arrive at estimates for the treatment effect that are unbiased, whereas often significant trial findings cannot be made a relevant yardstick if in a reasonable time frame less than 100 patients, or even less than 50 patients can be recruited.

Nevertheless, seeing a pre-specified hypothesis being rejected, or having controlled a type I error for a hypothesis at a pre-specified significance level, is often of high value, and from this background, fallback procedures as presented here can be of high importance in research for orphan medicines: if in the presented example significant findings could only be demonstrated in two out of three co-primary endpoints according to prespecified rules, this trial would have to be considered non-successful leaving open all the ambiguity, whether the two significant findings are spurious and in reality treatment groups do not differ. The proposed fallback procedure comes as a trick and at no direct cost allowing the latter suspicion to be falsified, because at least an intersection hypothesis or even an elementary hypothesis can be rejected. This may then constitute a better basis for a discussion about efficacy of an orphan drug than the usual approach to consider all findings as exploratory only and to try to base conclusions solely on trends in a variety of target variables.

Appendix A: Proof of Theorem 1

The diagonally trimmed Simes test may be regarded as a closed testing procedure where $H_1 \cap H_2$ is rejected if the vector of test statistics X is in the rejection region $R = \left\{x \in \mathbb{R}^2 : \min(x_1, x_2) \geqslant z_{1-\alpha} \text{ OR } (\max(x_1, x_2) \geqslant z_{1-\alpha/2} \text{ AND } x_1 + x_2 \geqslant 0)\right\}$. By the closed testing principle, FWER control follows if we can show type I error rate control under $H_1 \cap H_2$. Further, as the type I error rate will be maximal for $\mu_1 = \mu_2 = 0$, it is enough to consider the test under this point null hypothesis.

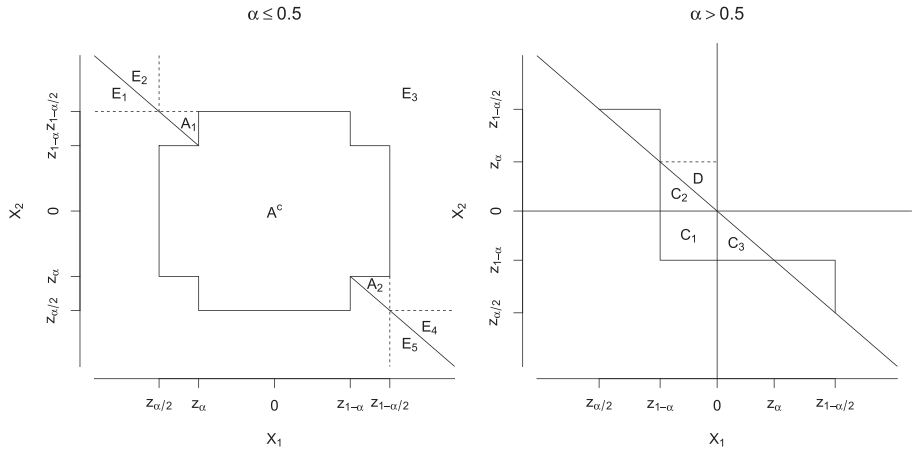

Figure A1. Left panel: case $\alpha \leqslant 0.5$. The rejection region of the Simes test is $\cup_{i=1}^{5} E_i$, the rejection region of the trimmed Simes test [21] is E_3, and the rejection region of the diagonally trimmed Simes test is $\cup_{i=2}^{4} E_i$. The halved rejection region of the two-sided Simes test is $A' = \cup_{i=2}^{4} E_i \cup A_1 \cup A_2$. A^c is the complement of the rejection region of the two-sided Simes test. Right panel: case $\alpha > 0.5$. See text for notation.

For the case $\alpha \leqslant 0.5$, the proof follows the argument of Theorem 4.1 in [21]. See Figure A1 for an illustration. Let $A = \left\{x \in \mathbb{R}^2 : \min(|x_1|, |x_2|) \geqslant z_{1-\alpha} \text{ OR } \max(|x_1|, |x_2|) \geqslant z_{1-\alpha/2}\right\}$ be the rejection region of the two-sided Simes test at level 2α. The two-sided Simes test for two hypotheses is conservative for bivariate normal test statistics [18] and for bivariate t-distributed test statistics [21] with arbitrary correlation. Due to the conservatism of the two-sided Simes test $P(A) \leqslant 2\alpha$ under the point null hypothesis, and due to the symmetry of the joint distribution of X, the halved region $A' = A \cap \left\{x \in \mathbb{R}^2 : x_1 + x_2 \geqslant 0\right\}$ has probability mass $P(A') = P(A)/2 \leqslant \alpha$. Because $R \subseteq A'$, FWER control for the diagonally trimmed Simes test follows for all levels $\alpha \leqslant 0.5$.

For the case $\alpha > 0.5$, first note that the only part of R not contained in the region $B = \{x \in \mathbb{R}^2 : x_1 + x_2 \geqslant 0\}$ is the set $C = \{x \in \mathbb{R}^2 : \min(x_1, x_2) \geqslant z_{1-\alpha}\} \cap B^c$. C is the union of the disjoint sets $C_1 = \{x \in \mathbb{R}^2 : z_{1-\alpha} \leqslant x_1 < 0, z_{1-\alpha} \leqslant x_2 < 0\}$, $C_2 = \{x \in \mathbb{R}^2 : x_1 \geqslant z_{1-\alpha}, x_2 \geqslant 0\} \cap B^c$ and $C_3 = \{x \in \mathbb{R}^2 : x_1 \geqslant 0, x_2 \geqslant z_{1-\alpha}\} \cap B^c$. Let $D = \{x \in \mathbb{R}^2 : x_1 \leqslant 0, x_2 \leqslant z_\alpha\} \cap B$ be the mirror set of C_3 with respect to the origin. By the symmetry of the joint distribution of X, $P(C_3) = P(D)$. Therefore, $P(C) = P(C_1 \cup C_2 \cup D) = P(\{x \in \mathbb{R}^2 : z_{1-\alpha} \leqslant x_1 \leqslant 0, z_{1-\alpha} \leqslant x_2 \leqslant z_\alpha\}) \leqslant P(z_{1-\alpha} \leqslant X_1 \leqslant 0) = \alpha - 0.5$. Now $P(R) \leqslant P(B) + P(C) \leqslant 0.5 + \alpha - 0.5 = \alpha$. See Figure A1 for an illustration.

Remark 1

The proof works identically under the more general assumptions similar to those made in [21], Theorem 4.1, that is, the distribution of X is continuous and symmetric under the point null hypothesis $\mu_1 = \mu_2 = 0$ in the sense that (X_1, X_2) has the same joint distribution as $(-X_1, -X_2)$, the marginal distribution functions $F_{X_i}(x_i, \mu_i), i = 1, 2$ are non-decreasing in μ_i, and the two-sided Simes test is conservative. The marginal distributions F_i need not be identical, if the p-value-based condition $p_1 + p_2 \leqslant 1$ is used instead of the test statistics-based condition $X_1 + X_2 \geqslant 0$. Note however that the theorem does not cover the case of the weighted trimmed Simes test considered in [21].

Remark 2

For the case $\alpha < 0.5$, a one-sided test rejecting $H_1 \cap H_2$ if $X \in A'$ has level not greater than α and is uniformly more powerful than the trimmed Simes test and the diagonally trimmed Simes test, as the

rejection region of either is contained in A'. However, such a test is not α-consistent, and it is not monotone in the test statistics X_1 and X_2.

Appendix B: Proof of Theorem 2

The theorem is shown by studying the gradient of π with respect to the pairwise correlation coefficients of the test statistics $\left(\rho_1, \rho_2, \rho_3\right)$. We first show (Lemmata 1 and 2) that the local extrema of π as a function of the correlation coefficients are located on the boundary of the parameter space, as defined in the succeeding discussions. Then (Lemma 3) we prove that for all correlation coefficients on this boundary, the type I error rate is less or equal than α. This proves the theorem.

Assumption 1
Let $\mathbf{X} = \left(X_1, X_2, X_3\right)^T$ denote a trivariate normally distributed random vector of standardized test statistics such that they have mean 0 and the correlation matrix Γ and covariance matrix Σ coincide. We use the notation

$$\Gamma = \begin{pmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_3 \\ \rho_2 & \rho_3 & 1 \end{pmatrix} = \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

The probability to falsely reject the global intersection hypothesis $H = H_1 \cap H_2 \cap H_3$ is given by

$$\pi = P(X_1 > d, X_2 > d) + P(X_1 > d, X_3 > d) + P(X_2 > d, X_3 > d) - 2P(X_1 > d, X_2 > d, X_3 > d)\,, \quad \text{(B.1)}$$

where d denotes the $1-\alpha$ quantile of the standard normal distribution. We assume $\alpha \leqslant 0.5$ and so $d \geqslant 0$. Note that for $d = 0$, it is trivial that $\pi = \alpha = 0.5$ due to the symmetry of the multivariate normal density with respect to its mean. Therefore, $\pi \leqslant \alpha$ remains to be shown for $d > 0$.

Definition 1
The boundary $\mathcal{B}$ of the parameter space is defined as the set of vectors (ρ_1, ρ_2, ρ_3) for which Γ is positive semidefinite but does not have full rank.

Lemma 1
Let (i, j, k) denote any permutation of the indices $(1, 2, 3)$. Then, under the assumptions of the theorem and assuming Γ is positive definite and $d > 0$,

$$\frac{\partial \pi}{\partial \rho_i} = 0 \quad \text{if and only if} \quad \rho_j + \rho_k = 1 + \rho_i\,.$$

Proof
W.l.o.g. assume that $(i, j, k) = (1, 2, 3)$. Let $\phi_{X_1X_2X_3}(x_1, x_2, x_3)$ denote the trivariate normal density of $\mathbf{X}$ and $\phi_{X_iX_j}(x_i, x_j)$ the binormal density of X_i and X_j, for $1 \leqslant i \neq j \leqslant 3$. In what follows, we drop the subscripts if the notation is clear from the function arguments.

Now,

$$\begin{aligned} \frac{\partial \pi}{\partial \sigma_{12}} = & \int_{-\infty}^{\infty}\int_{d}^{\infty}\int_{d}^{\infty} \frac{\partial^2}{\partial x_1 \partial x_2}\phi\left(x_1, x_2, x_3\right) dx_1 dx_2 dx_3 + 0 + 0 \\ & - 2\int_{d}^{\infty}\int_{d}^{\infty}\int_{d}^{\infty} \frac{\partial^2}{\partial x_1 \partial x_2}\phi\left(x_1, x_2, x_3\right) dx_1 dx_2 dx_3 \qquad \text{(B.2)} \\ = & \int_{-\infty}^{\infty} \phi\left(d, d, x_3\right) dx_3 - 2\int_{d}^{\infty} \phi\left(d, d, x_3\right) dx_3\,. \qquad \text{(B.3)} \end{aligned}$$

where in (B.2) Plackett's identity $\frac{\partial \phi}{\partial \sigma_{i,j}} = \frac{\partial^2 \phi}{\partial x_i \partial x_j}$ is used [38]. Dividing (B.3) by $\phi_{X_1X_2}(d, d) > 0$ does not affect the sign and gives the expression of conditional densities

$$\begin{aligned}\frac{1}{\phi_{X_1X_2}(d,d)}\frac{\partial\pi}{\partial\sigma_{12}} &= \int_{-\infty}^{\infty}\frac{\phi\left(d,d,x_3\right)}{\phi_{X_1X_2}(d,d)}dx_3 - 2\int_{d}^{\infty}\frac{\phi\left(d,d,x_3\right)}{\phi_{X_1X_2}(d,d)}dx_3 \\ &= \int_{-\infty}^{\infty}\phi\left(x_3|x_1=d,x_2=d\right)dx_3 - 2\int_{d}^{\infty}\phi\left(x_3|x_1=d,x_2=d\right)dx_3\end{aligned} \tag{B.4}$$

The conditional density of the multivariate normal distribution is again a normal density and therefore symmetric. Thus, the right-hand side of (B.4) is zero iff d is equal to the expected value $\mu = E(x_3|x_1 = d, x_2 = d)$. That is

$$\mu = \Sigma_{12}\Sigma_{22}^{-1}\begin{pmatrix}1\\1\end{pmatrix}d = \frac{\sigma_{13}+\sigma_{23}}{1+\sigma_{12}}d$$

where $\Sigma_{12} = (\sigma_{13}, \sigma_{23})$ and $\Sigma_{22} = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{pmatrix}$ are submatrices of the covariance matrix Σ and we used that $E(X) = 0$ and $\sigma_{11} = \sigma_{22} = \sigma_{33} = 1$. Now, dividing the equation $d = \mu$ by d gives $\sigma_{13} + \sigma_{23} = 1 + \sigma_{12}$ which is equivalent to

$$\rho_2 + \rho_3 = 1 + \rho_1$$

completing the proof of Lemma 1.

□

Lemma 2
Under the assumptions of the theorem, all local extrema of π in (ρ_1, ρ_2, ρ_3) are on the boundary of the parameter space $\mathcal{B}$.

Proof
For every local extremum which is not on the boundary, it holds that $\frac{\partial\pi}{\partial\rho_i} = 0, \forall i = 1, 2, 3$. According to Lemma 1, this implies that the correlation coefficients solve a system of three equations $\rho_j + \rho_k = 1 + \rho_i$ with (i, j, k) being permutations of $(1, 2, 3)$ with $j < k$. The unique solution is $(\rho_1, \rho_2, \rho_3) = (1, 1, 1)$, which is located on the boundary, which proves the lemma.

□

Lemma 3
Under the assumptions of the theorem and the condition that $(\rho_1, \rho_2, \rho_3) \in \mathcal{B}$, $\pi \leqslant \alpha$ holds.

Proof
We can rewrite the probability to reject H from equation (B.1) as

$$\pi = P(X_1 > d) - P(X_1 > d, X_2 < d, X_3 < d) + P(X_1 < d, X_2 > d, X_3 > d) \tag{B.5}$$

Note that $P(X_1 > d) = \alpha$ and therefore according to (B.5), it holds that $\pi \leqslant \alpha$ is equivalent to

$$P(X_1 > d, X_2 < d, X_3 < d) \geqslant P(X_1 < d, X_2 > d, X_3 > d) \tag{B.6}$$

We next look at the geometry of the boundary $\mathcal{B}$. First, define the four points

$$\mathcal{P}_1 := (1,1,1),\ \mathcal{P}_2 := (1,-1,-1),\ \mathcal{P}_3 := (-1,1,-1),\ \mathcal{P}_4 := (-1,-1,1)\ ,$$

which are the corners of a regular tetrahedron with six edges $\mathcal{E}_{ij}$ connecting $\mathcal{P}_i$ and $\mathcal{P}_j$, $1 \leqslant i < j \leqslant 4$. It is easy to see that all six edges including the four corners lie on $\mathcal{B}$. Furthermore, for all points on the edges $\mathcal{E}_{1j}, j \in \{2, 3, 4\}$, one has $\pi = \alpha$. For example, a typical point of $\mathcal{E}_{12}$ is of the form $(1, \rho, \rho)$, which implies that $X_1 = X_2$, and equality in equation (B.6) is trivial. On the other hand, similar considerations show that for points on the remaining three edges $\mathcal{E}_{ij}$, $1 < i < j$ (excluding the corners) one has $\pi < \alpha$.

The rest of the surface $\mathcal{B}$ consists of four distinct curved faces which are separated by the six edges of the tetrahedron. We denote by $\mathcal{F}_i$ the face which is opposite to $\mathcal{P}_i, i \in \{1, 2, 3, 4\}$. It remains to show that for all points on those four faces, one also has $\pi < \alpha$. In fact, it is sufficient to consider only $\mathcal{F}_1$ and one of the remaining three faces, because due to symmetry, the functional values of π on $\mathcal{F}_2$, $\mathcal{F}_3$, and $\mathcal{F}_4$ are identical.

Let us consider the following parametrization of the four faces. Without restriction of generality, we begin with fixing the third coordinate, $\rho_3 = c \in (-1, 1)$. Intersecting $\mathcal{B}$ with the hyperplane $\rho_3 = c$ yields an ellipse defined by

$$\rho_1^2 + \rho_2^2 - 2c\rho_1\rho_2 + c^2 - 1 = 0, \tag{B.7}$$

where we used that on $\mathcal{B}$ the determinant Δ of Γ is zero. We will parameterize the ellipse using the parameter $t = \rho_1$ and thus obtain for the two branches of the ellipse the equations

$$\rho_2(t, c) = ct \pm \sqrt{(1 - c^2)(1 - t^2)}$$

Given the parametrization, we rewrite

$$\begin{aligned}\pi_+(t, c) &:= \pi\left(t, ct + \sqrt{(1 - c^2)(1 - t^2)}, c\right)\\ \pi_-(t, c) &:= \pi\left(t, ct - \sqrt{(1 - c^2)(1 - t^2)}, c\right)\end{aligned}$$

Now we observe that the ellipse touches four edges of the tetrahedron discussed previously and is therefore divided into four sections which directly correspond to the four different faces of the surface. Specifically, we have the following

$\mathcal{F}_1$: $\pi_-(t, c)$ for $-1 < t < -c$
$\mathcal{F}_2$: $\pi_+(t, c)$ for $-1 < t < c$
$\mathcal{F}_3$: $\pi_-(t, c)$ for $-c < t < 1$
$\mathcal{F}_4$: $\pi_+(t, c)$ for $c < t < 1$

Due to the symmetry argument mentioned previously, it is now sufficient to consider, for example, $\mathcal{F}_1$ and $\mathcal{F}_3$, which conveniently are both parameterized with π_-.

The Cholesky decomposition of the correlation matrix is $\Gamma = LL^T$ with

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \rho_1 & \sqrt{1 - \rho_1^2} & 0 \\ \rho_2 & \frac{\rho_3 - \rho_1\rho_2}{\sqrt{1-\rho_1^2}} & \frac{\sqrt{\Delta}}{\sqrt{1-\rho_1^2}} \end{pmatrix} \tag{B.8}$$

For the irregular case where Γ has not full rank, $\Delta = 0$ and so the last column of L is zero. We can therefore write the multivariate normal distribution of X subject to $\Delta = 0$ using the Cholesky decomposition as

$$X = CZ, \quad \text{with} \quad C = \begin{pmatrix} 1 & 0 \\ \rho_1 & \sqrt{1 - \rho_1^2} \\ \rho_2 & \frac{\rho_3 - \rho_1\rho_2}{\sqrt{1-\rho_1^2}} \end{pmatrix}, \; Z = \mathcal{N}_2\left(0, I_2\right) \; . \tag{B.9}$$

Given our parametrization of the faces $\mathcal{F}_1$ and $\mathcal{F}_3$ with $\pi_-(t, c)$, it follows that at this part of the boundary, (B.6) is equivalent to

$$\begin{aligned}&P\left(Z_1 > d, tZ_1 + \sqrt{1 - t^2}Z_2 < d, \rho_2 Z_1 + \gamma Z_2 < d\right)\\ \geqslant\; &P\left(Z_1 < d, tZ_1 + \sqrt{1 - t^2}Z_2 > d, \rho_2 Z_1 + \gamma Z_2 > d\right)\end{aligned} \tag{B.10}$$

with $\rho_2 = \left(ct - \sqrt{(1 - c^2)(1 - t^2)}\right)$ and $\gamma = \left(c\sqrt{1 - t^2} + t\sqrt{1 - c^2}\right)$.

The probabilities of (B.10) are defined by three linear conditions, which correspond to the following three lines in the two-dimensional plane of (Z_1, Z_2):

g_1: $Z_1 = d$
g_2: $Z_2 = \frac{d - tZ_1}{\sqrt{1-t^2}}$
g_3: $Z_2 = \frac{d - \left(ct - \sqrt{(1-c^2)(1-t^2)}\right)Z_1}{c\sqrt{1-t^2} + t\sqrt{1-c^2}}$

To finish the proof, it is crucial to understand some geometric properties of these three lines. We note that S_{12}, the point where g_1 and g_2 intersect, lies at $Z_2 = d\frac{1-t}{\sqrt{1-t^2}}$, whereas S_{13}, the intersection point between g_1 and g_3, is at $Z_2 = d\frac{1-ct+\sqrt{(1-c^2)(1-t^2)}}{c\sqrt{1-t^2}+t\sqrt{1-c^2}}$.

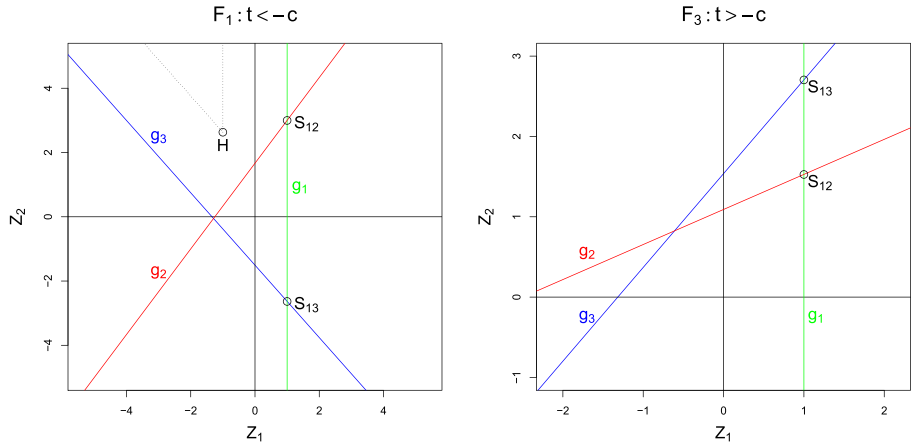

Figure B1. Typical situations for π_-: Left panel $\mathcal{F}_1$ with $c = 0.2$ and $t = -0.8$, right panel $\mathcal{F}_3$ with $c = 0.9$ and $t = -0.4$. In both cases, $d = 1$.

Let us first consider $\mathcal{F}_3$ for which the typical situation is depicted in the right panel of Figure B1. Elementary computations show that when $t > -c$, it always holds that S_{12} is below S_{13}, while the gradient of g_2 is smaller than that of g_3. In particular, this implies that

$$P\left(Z_1 > d, tZ_1 + \sqrt{1-t^2}Z_2 < d, \rho_2 Z_1 + \gamma Z_2 < d\right) = P\left(Z_1 > d, tZ_1 + \sqrt{1-t^2}Z_2 < d\right) ,$$

because due to the geometric constellation for $t > -c$, the condition implied by g_3 has become empty. Now the random vector (X_1, X_2) is bivariate normal with means 0 and variances 1, and thus, it is clear that

$$P\left(Z_1 > d, tZ_1 + \sqrt{1-t^2}Z_2 < d\right) = P\left(X_1 > d, X_2 < d\right) = P(X_1 < d, X_2 > d) .$$

Finally, one has

$$P\left(X_1 < d, X_2 > d\right) > P\left(X_1 < d, X_2 > d, X_3 > d\right) = P\left(Z_1 < d, tZ_1 + \sqrt{1-t^2}Z_2 > d, \rho_2 Z_1 + \gamma Z_2 > d\right) .$$

In geometric terms, g_3 cuts out a triangle with positive probability mass from the upper left area $\{X_1 < d, X_2 > d\}$. Thus, for $\mathcal{F}_3$, we have established the inequality (B.6).

For $\mathcal{F}_1$, the left panel of Figure B1 illustrates the typical situation. For $t < -c$, it is always true that S_{12} is above S_{13}, and that the gradient of g_2 is larger than that of g_3. Also note that geometrically, the condition $X_3 < d$ means that Z_2 is above the line g_3 (because of $\nu < 0$ the orientation changes). Therefore,

$$P(X_1 > d, X_2 < d, X3 < d) = P(X_1 > d, X_2 < d) - P(X_1 > d, X_3 > d)$$

and as before, we have

$$P(X_1 > d, X_2 < d) = P(X_1 < d, X_2 > d) .$$

It follows that to establish (B.6), it suffices to show that

$$P(X_1 > d, X_3 > d) < P(X_1 < d, X_2 > d, X3 < d)$$

where the right-hand side is the probability of the area left of g_1 and above g_2 and g_3. In geometrical terms, the key idea is to mirror the area $\{X_1 > d, X_3 > d\}$ at the origin and show that the mirror image is a proper subset of $\{X_1 < d, X_2 > d, X3 < d\}$ (compare Figure B1). This will certainly be the case

when $H := -S_{13}$, the mirror image of S_{13}, lies above g_2, which can be easily established by showing that $g_2(-d) < -g_3(d)$, and when the slope of g_2 is greater than the slope of g_3, which also holds.

More algebraically, we have thus shown the following:

$$\begin{aligned} P(X_1 > d, X_3 > d) = &\quad P(X_1 < -d, X_3 < -d) \\ = &\quad P(X_1 < -d, X_2 > d, X_3 < -d) \\ < &\quad P(X_1 < d, X_2 > d, X_3 < d) \end{aligned}$$

The first equality holds because of the symmetry of the bivariate standard normal distribution, and it corresponds to the mirroring of g_1 and g_3. The second equality follows from the fact that H lies above g_2, and thus, the condition $X_2 > d$ is empty, and the final inequality is trivial. □

Appendix C

To prove that the fallback test controls the FWER at level α when applied to the test statistics T_1, T_2, T_3, we show that the test for the global null hypothesis $H = H_1 \cap H_2 \cap H_3$ has level α (for all other intersection hypothesis tests the result is trivial). If H holds, for every $i = 1, 2, 3$, either $H_{se,i}$ or $H_{sp,i}$ (or both) holds. Let $Z = (Z_{s_1,1}, Z_{s_2,2}, Z_{s_3,3})$ denote the corresponding vector of test statistics, where $s_i \in \{se, sp\}$ corresponds to the respective true null hypothesis (if both are true, set $s_i = se$). Asymptotically, Z is multivariate normal, and therefore, Theorem 1 applies. Because $(T_1, T_2, T_3) \leqslant Z$, this also follows for the proposed test.

Acknowledgements

We wish to thank Ferran Torres for pointing us to the clinical trial in Lennox–Gastaut syndrome and Christian Gartner and Thomas Lang for their valuable comments on the diagnostic trial example. This work has been funded by the FP7-HEALTH-2013-INNOVATION-1 project Advances in Small Trials Design for Regulatory Innovation and Excellence (ASTERIX) grant agreement no. 603160.

References

1. CHMP. *Guideline on medicinal products for the treatment of Alzheimers disease and other dementias*, 2008. doc. ref. cpmp/ewp/553/95 rev. 1.
2. CHMP. *Draft guideline on the clinical investigation of medicinal products for the treatment of Duchenne and Becker muscular dystrophy*, 2013.
3. Glauser T, Kluger G, Sachdeo R, Krauss G, Perdomo C, Arroyo S. Rufinamide for generalized seizures associated with Lennox–Gastaut syndrome. *Neurology* 2008; **70**(21):1950–1958.
4. CHMP. *Points to consider on multiplicity issues in clinical trials*, 2002.
5. Offen W, Chuang-Stein C, Dmitrienko A, Littman G, Maca J, Meyerson L, Muirhead R, Stryszak P, Baddy A, Chen K, Copley-Merriman K, Dere W, Givens S, Hall D, Henry JD, Jackson D, Krishen A, Liu T, Ryder S, Sankoh AJ, Wang J, Yeh CH. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of america. *Drug Information Journal* 2007; **41**(1):31–46.
6. Lucadamo A, Accoto N, De Martini D. Power estimation for multiple co-primary endpoints: a comparison among conservative solutions. *Italian Journal of Public Health* 2012; **9**(4).
7. Chuang-Stein C, Stryszak P, Dmitrienko A, Offen W. Challenge of multiple co-primary endpoints: a new approach. *Statistics in Medicine* 2007; **26**(6):1181–1192.
8. Chuang-Stein C, Dmitrienko A, Offen W. Discussion of some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**(1):14–21.
9. Kordzakhia G, Siddiqui O, Huque M.F. Method of balanced adjustment in testing co-primary endpoints. *Statistics in Medicine* 2010; **29**(19):2055–2066.
10. Bauer P. Multiple testing in clinical trials. *Statistics in Medicine* 1991; **10**(6):871–890.
11. König F, Bauer P, Brannath W. An adaptive hierarchical test procedure for selecting safe and efficient treatments. *Biometrical Journal* 2006; **48**(4):663–678.
12. Rueger B. Das maximale Signifikanzniveau des Tests: Lehne H0 ab, wenn k unter n gegebenen Tests zur Ablehnung fuehren. *Metrika* 1978; **25**(1):171–178.
13. Wiens BL, Dmitrienko A. The fallback procedure for evaluating a single family of hypotheses. *Journal of Biopharmaceutical Statistics* 2005; **15**(6):929–942.
14. Marcus R, Eric P, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**(3):655–660.
15. Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**(4):800–802.
16. Hommel G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* 1988; **75**(2): 383–386.

17. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**(3):751–754.
18. Samuel-Cahn E. Is the Simes improved Bonferroni procedure conservative? *Biometrika* 1996; **83**(4):928–933.
19. Block HW, Savits TH, Wang J. Negative dependence and the Simes inequality. *Journal of Statistical Planning and Inference* 2008; **138**(12):4107–4110.
20. Block HW, Savits TH, Wang J, Sarkar SK. The multivariate-t distribution and the Simes inequality. *Statistics & Probability Letters* 2013; **83**(1):227–232.
21. Brannath W, Bretz F, Maurer W, Sarkar S. Trimmed weighted Simes' test for two one-sided hypotheses with arbitrarily correlated test statistics. *Biometrical Journal* 2009; **51**(6):885–898.
22. Hommel G, Bretz F. Aesthetics and power considerations in multiple testing–a contradiction? *Biometrical Journal* 2008; **50**(5):657–666.
23. van de Wiel MA, Berkhof J, van Wieringen WN. Testing the prediction error difference between 2 predictors. *Biostatistics* 2009; **10**(3):550–560.
24. Massart P. *Concentration Inequalities and Model Selection*. Springer: Berlin, 2007.
25. Caraux G, Gascuel O. Bounds on distribution functions of order statistics for dependent variates. *Statistics & Probability Letters* 1992; **14**(2):103–105.
26. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
27. Westfall PH, Young SS. *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley & Sons: New York, 1993.
28. R Core Team. *R: A language and environment for statistical computing*, R Foundation for Statistical Computing: Vienna, Austria, 2013. http://www.R-project.org/ [Last Accessed on 11 January 2016].
29. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. *mvtnorm: Multivariate normal and t distributions*, 2014. http://CRAN.R-project.org/package=mvtnorm, r package version 0.9-9997 [Last Accessed on 11 January 2016].
30. Bretz F, Hothorn T, Westfall P. *Multiple comparisons using R*. CRC Press: Boca Raton, 2010.
31. CHMP. *Appendix 1 to the guideline on clinical evaluation of diagnostic agents (cpmp/ewp/1119/98 rev. 1) on imaging agents*, 2009.
32. Van der Vaart AW. *Asymptotic Statistics*. Cambridge University Press: Cambridge, 2000.
33. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in medicine* 2008; **27**(24):4914–4927.
34. Guilbaud O. Simultaneous confidence regions corresponding to Holm's step-down procedure and other closed-testing procedures. *Biometrical Journal* 2008; **50**(5):678–692.
35. Hayter AJ, Hsu JC. On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association* 1994; **89**(425):128–136.
36. Lehmann EL, Romano JP. Generalizations of the familywise error rate. *Annals of Statistics* 2005; **33**(3):1138–1154.
37. Xi D, Tamhane AC. A general multistage procedure for k-out-of-n gatekeeping. *Statistics in Medicine* 2014; **33**(8): 1321–1335.
38. Plackett RL. A reduction formula for normal multivariate integrals. *Biometrika* 1954; **41**(3–4):351–360.

