Author/s:
Heritier, S;Lloyd, CJ;Lô, SN

Title:
Accurate p-values for adaptive designs with binary endpoints

Date:
2017-07-30

Citation:
Heritier, S., Lloyd, C. J. & Lô, S. N. (2017). Accurate p-values for adaptive designs with binary endpoints. Statistics in Medicine, 36 (17), pp.2643-2655. https://doi.org/10.1002/sim.7324.

Persistent Link:
https://hdl.handle.net/11343/292858

Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Accurate p-values for adaptive designs with binary endpoints

Stephane Heritier[a*], Chris J. Lloyd[b] and Serigne N. Lô[c,d]

Adaptive Designs (ADs) encompass all trials allowing various types of design modifications over the course of the trial. A key requirement for confirmatory ADs to be accepted by regulators is the strong control of the family-wise error rate (FWER). This can be achieved by combining the p-values for each arm and stage to account for adaptations (including but not limited to treatment selection), sample size adaptation and multiple stages. While the theory for this is novel and well-established, in practice these methods can perform poorly, especially for unbalanced designs and for small to moderate sample sizes. The problem is that standard stagewise tests have inflated type I error rate, especially but not only when the baseline success rate is close to the boundary and this is carried over to the adaptive tests, seriously inflating the FWER. We propose to fix this problem by feeding the adaptive test with second-order accurate p-values, in particular bootstrap p-values. Secondly, an adjusted version of the Simes procedure for testing intersection hypotheses that reduces the built-in conservatism is suggested. Numerical work and simulations show that unlike their standard counterparts the new approach preserves the overall error rate, at or below the nominal level across the board, irrespective of the baseline rate, stagewise sample sizes or allocation ratio.

Keywords: adaptive design; second-order test; combination test; bootstrap p-value, familywise error rate, Simes

1. Introduction

Over the past three decades, a lot of research effort has gone into developing methods to control type 1 error rates for adaptive phase II/III clinical trials; see [1–6] and references therein. A now well-established method involves feeding a set of component p-values, say for different treatment arms and stages, through a quite complicated algorithm. The key principles are multiplicity adjustment to adjust for selection, a combination function to pool information across several stages and the closed testing principle (CTP) to control familywise error rate (FWER). Practitioners might be forgiven for thinking the problem is solved. In theory it is but in practice it isn't. The purpose of this paper is to address the poor type 1 error control that easily arises when standard p-values are used to feed this algorithm. While the core ideas are general, we are mainly concerned with trials where the treatments are assessed using a binary endpoint. By standard p-values we mean those based on a normal approximation to the Z-statistic or likelihood ratio statistic. Especially when the control and treatment arms are unbalanced these standard methods have erratic properties even for quite large sample sizes and can lead to an adaptive test which is very liberal. On the other hand, multiplicity adjustment using the procedure of Simes [7] can be quite conservative which results in loss of power. The basic p-values used, and the Simes procedure, are the focus of this article. Specifically, we recommend replacing the approximate p-values with so-called parametric bootstrap p-values which are easy to calculate for moderate sample sizes and have excellent statistical properties [8]. Secondly, in combining p-values to test multiple hypotheses, we use an adjusted version of the Simes p-value that is less conservative than the standard Simes method for moderate correlations [9].

[a] *School of Public Health and Preventive Medicine, Monash University, Melbourne, VIC, Australia*
[b] *Melbourne Business School, University of Melbourne, Carlton, VIC, Australia*
[c] *Melanoma Institute Australia, North Sydney, NSW, Australia*
[d] *Sydney Medical School, The University of Sydney, Sydney, NSW, Australia*
[*] *Correspondence to: Stephane Heritier, Department of Epidemiology and Preventive Medicine, The Alfred Centre, 99 Commercial Road, Melbourne VIC 3004, Australia. Email: stephane.heritier@monash.edu*

Prepared using simauth.cls [Version: 2009/09/02 v2.00]

Section 2 provides an illustration of the dangers of feeding approximate p-values into either a two stage design with no selection or a single stage design with selection. Section 3 presents the three key elements of adaptive inference, namely multiplicity adjustment, combination functions and CTP. Section 4 presents standard tests and two alternative counterparts with higher order accuracy, one based on a second order approximation to the likelihood ratio (the so-called p-star statistic) and the second based on an exact calculation with estimated parameters assumed true (which is a parametric bootstrap). In section 5, we assess the performance of the procedures (level, power), implemented with both the standard Simes and its adjusted version. Section 6 illustrates the new approach in seamless phase II/III adaptive designs with regimen selection. Finally, possible extensions and a discussion are presented in Section 7.

2. Is type 1 error control really a problem?

In the dose selection problem, there are two key aspects of adaptive designs that need to be accounted for to ensure type 1 error control. The first is selection of the more promising treatment(s). This requires a multiple comparison adjustment if type 1 error is not to be severely inflated. The second is modifying design elements at an interim, for instance increasing the sample size in the remainder of the trial if results are not as good as anticipated. Again, this leads to severe type 1 error inflation if handled naively. A proper method is to combine the evidence from before and after the change via a so-called combination function. This method requires that the component p-values are "correct", ideally uniformly distributed but certainly stochastically no smaller than uniform. However, standard p-values are known to violate this condition, even for large sample sizes [10].

The problem is particularly severe for designs with unequal allocation ratio whose demand is growing [11, 12]. Rationale for this increased demand includes i) enrollment issues for placebo patients (see [13, 14] for an example in the adaptive design setting); ii) experimental therapy in short supply or expensive which pushes investigators to recruit more readily available controls [15, 16]; iii) a way to offset a higher dropout in other subgroups and other reasons [16]. Adaptivity itself can generate unbalanced designs as recommended in multi-arm multi-stage trials [17–19] or when a dose is added at an interim analysis and overall balance is desired at the end of the study.

While we are ultimately interested in designs that involve *both* selection and interim modification, it is pertinent to examine how the standard methods perform in two simpler designs, where the endpoint is binary. In our first design, there is a single treatment arm (so there is no selection) and two stages. For simplicity assume that the arm sizes are the same in both stages. We combine the two stagewise p-values from the pooled Z-test using the Fisher and normal combinations function (see next section). The left panel of Figure 1 shows the exact sizes of nominal 5% test for sample sizes of 30 in the control arm, 60 in the experimental arm at each stage. It has not been derived from the asymptotic distribution but rather calculated exactly. In the second design, there is only one stage but three treatment arms, each with a sample size of 60, to be tested against a common control with sample size 30. Each test is based on the likelihood ratio statistic. The treatment with the smallest p-value is chosen and we want to test whether it is effective. Some adjustment (such as the Bonferroni or Simes adjustment) is required to account for our data-based choice of the smallest p-value. If control of FWER is required then we must also apply the CTP. Details of these methods are in the next section. The right panel of Figure 1 shows the exact size of a nominal 5% test using these various adjustments.

Clearly, neither test preserves the nominal level of 5% across the whole range of baseline parameter values. The pooled z-test apparently performs better than the likelihood ratio (LR) tests for these sample sizes. The true size violates nominal most egregiously for more extreme success rates of the baseline but even for moderate baseline success rates, violation can be non-trivial. These problem remain across a range of conditions and statistics with sample sizes and we refer the reader to additional calculations and graphics in Figure 1 of supplementary material. Overall, practically significant size violations are commonplace even for sample sizes in the order of 200, which is a concern in regards to typical phase II sample sizes. While one may argue that we could, in practice, choose the underlying test to avoid the spikes, in zones where its performance is deemed acceptable (e.g. assuming a range of the parameter values and sample sizes), this is 'ad-hoc' and risky (if these assumptions turn out to be wrong). What is needed is a procedure that is stable across *all* baseline parameter values.

3. Elements of error control for adaptive designs

Consider K treatments or treatment variants which are to be compared to a control. We denote the probability of a favourable binary response (e.g. survival rate at a given timepoint) by $\pi_k; k = 0, ..., K$, with control coded as $k = 0$. The effect of treatment k is measured by θ_k (e.g. the difference in success rates $\pi_k - \pi_0$) and is assessed by testing the null hypothesis $H_{0,k} : \theta_k \leq 0$ that it is ineffective versus the alternative $H_{1,k} : \theta_k > 0$ that it is effective.

Let p_k be the p-value for testing $H_{0,k}$ based on comparing the outcomes for those given treatment k with those given control. Since they each use the same control data, these p-values are positively correlated. Critically for the theory to follow, they are each assumed to be stochastically no smaller than uniform, so that for any nominal size type 1 error control is assured for testing each individual treatment.

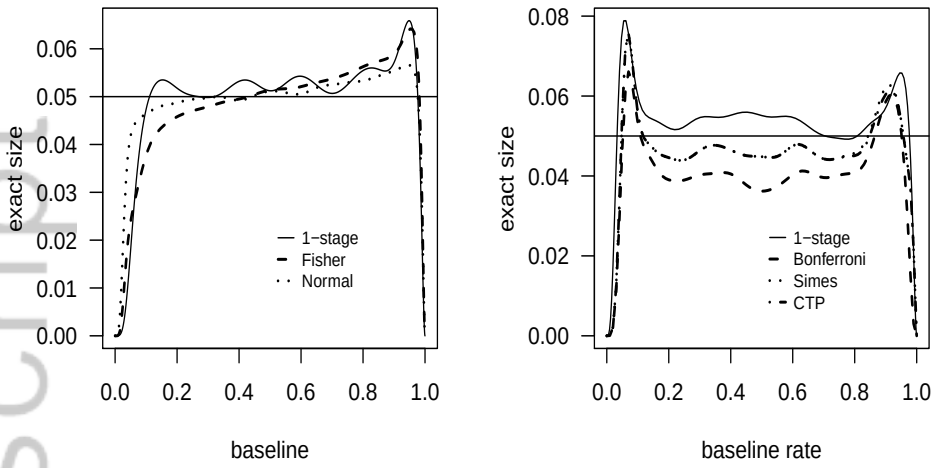

Figure 1. Size profiles of two elemental designs for $n_0 = 30, n_1 = 60, \alpha = 0.05$. *Left.* Two stage design with no selection by combining two stagewise pooled Z-tests using Fisher and normal combinations function. *Right.* Single stage three arm design selecting most significant treatment using LR test and three multiplicity adjustments.

3.1. Testing several treatments

To test whether any of the treatments are effective, the null hypothesis is the intersection $\cap_{k=1}^{K} H_{0,k}$ that no treatment is effective. The alternative is that at least one of the treatments is effective. The test is to be based only on the p-values $p_1, ..., p_K$. Naive methods lead to inflated type 1 error. There are several methods to account for this, known as multiplicity adjustments (e.g. Bonferroni, Hommel, Holm, Sidak, Dunnett tests among others, see [1]). We focus here on the Simes procedure [7] which is known to control type 1 error when the p_k are either independent or positively correlated [20, 21].

Letting $p_{(i)}$ denote the ith largest of the p_k, Simes' p-value is given by

$$p_{\text{SIMES}} = \min_{k \in \{1,...,K\}} \{K p_{(k)}/k\} \tag{1}$$

Note that $p_{\text{SIMES}} \leq K p_{(1)} = p_{\text{BONF}}$ and is therefore more powerful and less conservative than Bonferroni. The Simes construction can be applied to any subset $\mathcal{I}$ of $\{1, ..., K\}$, the only difference being that K is replaced by the cardinality of $\mathcal{I}$. Simes' p-value is still conservative (i.e. stochastically larger than uniform) under the null, especially for correlations in the 0.7-0.9 range, which is likely to occur for allocation ratios r around 3. Lloyd [9] showed numerically that the conservatism of the Simes p-value mainly depends on three parameters: first, the pairwise correlation of the component p-values, second the number of hypotheses being tested, and lastly the skewness of the underlying Z-statistics. Based on a great deal of numerical work, this regular dependence is modeled and an appropriate adjustment to the Simes p-value is made. All details are in [9], where simple code is also provided. This modified Simes p-value can be used instead of (1) though it cannot be formally proven to control the overall type I error.

3.2. Familywise error rate and CTP

The FWER is controlled in the strong sense if the probability of rejecting at least one true null hypotheses, is controlled, regardless of the configuration of false and true hypotheses. It is preferred to the weak control of type 1 error under the global null, considered as too soft a criterion by many researchers and regulators. The application of the CTP is necessary and sufficient [22] to preserve FWER in the strong sense. This principle requires that, to reject a single hypothesis $H_{0,k}$, all intersection hypothesis containing $H_{0,k}$ must be rejected. Using the notation $p_{\mathcal{I}}$ to denote the p-value for testing the intersection hypothesis associated with $\mathcal{I}$, the CTP p-value for testing $H_{0,k}$ is given by

$$\max\{p_{\mathcal{I}} : k \in \mathcal{I}\}. \tag{2}$$

For instance, when $K = 3$, the null hypothesis $H_{0,1}$ is decided by calculating p-values p_1, $p_{\{12\}}$, $p_{\{13\}}$ and $p_{\{123\}}$ and comparing the maximum of these with the nominal level α. Apart from p_1, the other three p-values will be based on the Simes (or modified Simes) construction.

3.3. Multiple stages and combination tests

Various authors including [1, 2, 4, 23] presented a variety of adaptive designs for seamless phase II/III clinical trials. Their approach follows the initial idea of [24] that allows data-dependent adaptations while preserving the integrity of the trial.

We refer to data prior to adaptation as stage 1 and after adaptation as stage 2. The control and experimental sample sizes are n_{i0} and n_{ik} $k = 1, \ldots, K$ respectively for stage $i = 1, 2$. We also assume a fixed allocation ratio $r_k = n_{ik}/n_{i0}$ across the stages for whatever treatments k are present in stage i. Analysing the whole data ignoring the adaptation leads to loss of error control. For instance, if a treatment looks promising one might want to reduce the number of subjects at stage 2 but this will inflate type 1 error unless properly handled. Temporarily dropping the subscript k, a single one-sided hypothesis H_0 is to be tested and we let p and q be the p-values from stage 1 and stage 2 data respectively. The key idea is to combine the two p-values through a so-called combination function $C(p, q)$ which is increasing in both p and q and whose distribution is uniform when p and q are independent and uniformly distributed. Formally, the null hypothesis is rejected if an appropriately chosen combination $C(p, q)$ is less than or equal to α [24–26].

Fisher [27] considered this issue and noted that $-2\log(pq)$ has the χ^2_4 distribution and so $C(p, q) = 1 - F_{\chi^2_4}(-2\log(pq))$ is uniform. This combination function remains popular. The most common competitor is the weighted inverse normal combination function [28, 29] where

$$C(p, q) = 1 - \Phi(w_1\Phi^{-1}(1 - p) + w_2\Phi^{-1}(1 - q)), \tag{3}$$

Φ is the standard normal distribution function, and w_i, $i = 1, 2$ are pre-specified positive weights such that $w_1^2 + w_2^2 = 1$. If the second stage involves sample size adaptation the weights must remain unchanged for the procedure to be valid. Denote by $Z_1 = \Phi^{-1}(1 - p)$ and $Z_2 = \Phi^{-1}(1 - q)$ the respective stagewise one-sided $Z-$statistics. Rejecting H_0 when $C(p, q) \leq \alpha$ in (3) is equivalent to rejecting the null using a weighted test statistic $Z = w_1Z_1 + w_2Z_2$.

In this construction, we have implicitly assumed that no formal stopping rule for futility or efficacy has been pre-specified at the end of the first phase (where typically an interim analysis is conducted). The calculation of the overall p-value for the combination test can however be easily modified to accommodate these constraints - see, for instance references [1, 26] for details.

A common choice for the weights is

$$w_i = \frac{\sqrt{n_{i0}}}{\sqrt{n_{10} + n_{20}}} \text{ for } i = 1, 2. \tag{4}$$

These weights, optimal for Gaussian endpoints as based on the square root of the originally planned information increments, are often used for binary endpoints for simplicity and avoid dependence on the parameter of interest. This weighting was chosen in the examples of Section 2 involving the inverse normal combination function and in the simulations presented in Section 3. In the balanced case, $n_{i0} = n_{ik} = n_i$ and (4) reduces to the well known formula $w_i = \sqrt{n_i}/\sqrt{n_1 + n_2}$ for $i = 1, 2$.

3.4. Application to adaptive dose selection

Our later numerical work will focus on a design that involves both multiple treatments and stagewise adaptation, where the adaptation includes selection of the most promising treatment(s). To illustrate suppose that, on the basis of effect size or statistical significance, a subset S of the treatments are carried forward to stage 2. At the end of stage 2, we want to test whether one of the selected treatments $k \in S$ is effective, accounting for the fact that it was selected from stage 1, for possible changes to sample sizes at stage 2, and controlling FWER. We have available p-values $p_1, ..., p_K$ from the first stage and $q_i : i \in S$ from the second stage. The formula (2) for the CTP p-value still applies except that each intersection hypothesis $\mathcal{I}$ is to be tested from an appropriate combination of evidence from both stages. From stage 1, we have the p-value $p_{\mathcal{I}}$ for for testing $\mathcal{I}$. For stage 2, some of the elements in $\mathcal{I}$ may not have been carried forward so we just apply a standard multiplicity p-value to those treatments in $\mathcal{I}$ that were carried forward i.e. to $\mathcal{I} \cap S$. To summarise, to test whether a selected treatments $k \in S$ is effective we reject $H_{0,k}$ if

$$C(p_{\mathcal{I}}, q_{\mathcal{I}\cap S}) \leq \alpha \text{ for all } \mathcal{I} : k \in \mathcal{I}, \tag{5}$$

and the combined p-value for hypothesis $H_{0,k}$ is the maximum of all such individual term $C(p_{\mathcal{I}}, q_{\mathcal{I}\cap S})$ over all intersections $\mathcal{I} : k \in \mathcal{I}$.

4. Standard and alternative stagewise p-values

The theory in section 3 requires all of stagewise p-values $p_1, ..., p_K$ and $q_j : j \in S$ to be either uniform or stochastically larger than uniform under the null hypothesis. Especially for binary outcomes, standard p-values can violate this condition badly, even for quite large sample sizes. In this section we describe the standard p-values and some more modern alternatives. For simplicity, we drop subscript i corresponding to stage and present the calculation on testing $H_0 : \theta_1 \leq 0$ without loss of generality.

4.1. Asymptotic p-values

Stagewise p-values feeding the combination tests are commonly based on the unpooled Z-test for proportions, denoted here by Z_U, the pooled test Z_P, or the signed root likelihood ratio ($SRLR$) test Z_L, which is the one-sided version of the likelihood ratio test (see [30], p128). The formal definitions of these test statistics are well known and are given in the Appendix. Each Z-statistic produces a p-value $\sup_{H_0} \Pr(Z \geq z|\pi_0, \pi_1)$ where z is the observed value. The standard p-values ignore dependence of the tail probability on (π_0, π_1) and instead approximate this by $1 - \Phi(z)$. For these so-called first-order p-values, the absolute error is of order $O(n^{-1/2})$ where n is the minimum of the two group sizes provided that the free parameter π_0 is not on the boundary. For small to moderate samples and for noncentral values of π_0 e.g. $\pi_0 \leq 0.10$ or $\pi_0 \geq 0.90$), asymptotic p-values are often inaccurate [31]). Specifically, they are not close to $\Pr(Z \geq z|\pi_0)$ and their actual distribution is far from $U[0, 1]$ under the null. For this reason, the standard statistics often come with a caveat that they should not be used if expected frequencies are below the arbitrary threshold of 5, while not specifying what is to be done in this case.

The size violations of the standard statistics are typically carried over to the combination p-value and also to the Simes and CTP p-values, as was illustrated in Figure 1. Potentially then, this can cause the adaptive testing procedure to breakdown and generate a type I error much larger than nominal. Such effects were demonstrated in [32] for Z_U and Z_L in the dose selection problem. Although Z_P was the chosen test statistic for this 2 : 1 design in favour of any experimental arm, it would have failed, had the design required more control patients than experimental ones. This feature has been overlooked due to the fact that balanced designs (1:1) are more commonly used in clinical trials masking the underlying issue of instability of standard testing procedures.

4.2. Second-order procedures

There are at least two alternative approaches to improving the size accuracy of p-values. The first is based on so-called higher order asymptotics, see for instance, [33–35]. The second, sometimes called parametric bootstrap [31], involves an exact calculation assuming unknown parameters equal their empirical values. Both approaches have absolute error of order $O(n^{-3/2})$ for continuous models and seemingly of order $O(n^{-1})$ for discrete models [31, 36]. Our proposal is to feed these second order p-values into the adaptive inference and to numerically examine the accuracy of the resulting p-values.

4.2.1. Modified likelihood ratio The modified likelihood ratio statistic is based on higher order asymptotic approximations to a conditional distribution. The theory underpinning this development is complex [37, 38] but the test statistic has a closed form:

$$Z_L^* = Z_L + Z_L^{-1} \log (Q/Z_L) \tag{6}$$

where Q is complicated in general but is given explicitly in the appendix. The corresponding p-value is simply $p^* = 1 - \Phi(z_L^*)$. Technically, Z_L^* is not defined when $Z_L = 0$ and Q is undefined when $Z_L = \infty$ though this is not a practical limitation.

4.2.2. Parametric bootstrap The general principle of bootstrap is to replace unknown parameters in some measure of statistical accuracy by their estimated value. In non-parametric settings this often leads to intractable expressions that require resampling based simulation. For parametric hypothesis testing, this need not be so. We start with the exact significance $\Pr(Z \geq z|\theta)$ of an observed value z of a test statistic Z. This depends on $\theta \in \mathcal{H}_0$ so we replace this unknown by its estimate $\tilde{\theta}_0$ under the null. Theoretical accounts may be found in [36,39], and the present application to binomial data has been explained and evaluated in [31]. In the specific case of two samples of binary data, let $Y_j : j = 0, 1$ be the sum of all individual binary responses for the n_j subjects in group j. The *exact* p-value based on some first order test statistic $Z(Y_0, Y_1) = z$ is

$$\Pr(Z(Y_0, Y_1) \geq z|\pi_0) = \sum_{(y_0,y_1):Z(y_0,y_1)\geq z} b(y_0; n_0, \pi_0) b(y_1; n_1, \pi_0)$$

where $b(y; n, \pi)$ denotes the general binomial probability function. The bootstrap p-value is $p_{\text{BOOT}} = \Pr(Z \geq z|\tilde{\pi})$ where $\tilde{\pi} = (y_0 + y_1)/(n_0 + n_1)$ and the corresponding z-statistic $Z_{\text{BOOT}} = \Phi^{-1}(1 - p_{\text{BOOT}})$. For sample sizes up to the thousands, it is possible to compute p_{BOOT} directly by enumeration without using simulation as is typically associated with bootstrap. In wider contexts, for example logistic regression, the bootstrap p-value can be efficiently simulated using importance sampling [40].

Looking at the definition above, p_{BOOT} in principle depends on the choice of statistic Z. Luckily this dependence is very slight so long as Z is one of the standard first order test statistics [31]. Throughout we base our bootstrap p-value on Z_L which has certain guaranteed monotonicity properties. These properties ensure that the maximum value of $\Pr(Z_L \geq z|\pi_0, \pi_1)$ over the null hypothesis that $\pi_1 \leq \pi_0$ is attained for $\pi_0 = \pi_1$. This guarantees that the bootstrap p-value as defined above is compatible with the general definition of size of a test.

It has been shown in [31] that p_{BOOT} is extremely close to uniformly distributed under the null, even for small sample sizes or baseline π_0 close to the boundary. So another terminology could be "quasi-exact" p-value but, for consistency, we stick to the name that is already established in this literature. The bottom line is that we expect the procedure to work even if small samples are chosen in stage 1 (phase II) of these two stage-designs.

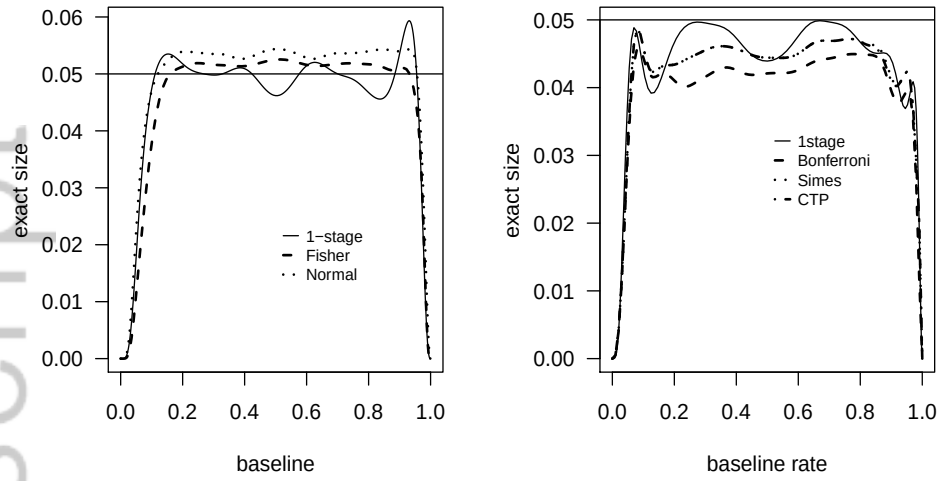

Figure 2. Size profiles of two second order procedures with $n_0 = 30, n_1 = 60, \alpha = 0.05$**.** *Left.* Two stage design with no selection using p^*. *Right.* Single stage three arm design selecting most significant treatment using bootstrap p-value and three multiplicity adjustments.

5. Numerical results

5.1. Size profile plot for simple models

As illustrated in Figure 1, for a two-stage model with no selection or a single stage model with selection of three treatments, it is feasible to calculate the size exactly as a function of the free baseline parameter. We call this the size *profile*. The left panel of Figure 2 displays size profiles for the two-stage design with samples sizes $n_0 = 30, n_1 = 60$ at each stage. The basic statistic is the modified likelihood ratio. The solid line gives the profile of the one stage test which appears to be somewhat better than the pooled z-statistic in Figure 1. The combined tests are displayed with broken lines and exceed nominal consistently (by a small amount). While this is just one example, theoretical work in [41] shows that the normal combination size profile is largely determined by the profile of the single stage test at a size much higher than the nominal α. The bottom line is that inference based on p^* looks better than using the pooled Z statistic but still violates nominal.

The right panel shows results for a single stage three arm design. The basic statistic here is p_{BOOT} and the single arm profile looks to control size extremely well. Again this is only one example, but extensive numerical work in [31] shows that p_{BOOT} is almost always very close to uniform. Consequently, the multiple comparison adjusted p-values all control size below nominal but tend to be conservative. This is nothing to do with p_{BOOT}. This is a well known property of Simes p-values: if the single arm p-values are exactly uniform then the final p-values will typically be larger than uniform (i.e. conservative) under the multivariate totally positive of order two (MTP2) condition [20] shared by commonly encountered multivariate distributions in multiple testing.

For more general adaptive designs, it is not feasible to calculate size profiles exactly. However, the results for the elemental designs we have just described give confidence that second order methods will control size well in more complex designs.

5.2. Level simulations

We now study the performance of the new proposals in the dose selection problem as described in section 3.4. Simulations have been carried out to compare the level of the combination test fed with the following 5 statistics: Z_U, Z_P, Z_L, Z_L^* and Z_{BOOT} described above. In all the simulations the most effective treatment (the one with smallest p-value) is chosen at interim. As background for the simulation design, we consider the 2-stage adaptive design described in [13, 32], a 2:2:2:2:1 study in paediatric dermatology where one (or two) treatment(s) had to be selected at the interim. The study allowed other adaptations (stopping for futility and sample size reassessment) but we will not consider these for simplicity. They did not come into effect in the original study either [14]. To go beyond this trial, we imagined that the allocation ratio is $r : 1$ per experimental:control arm with $r > 1$ indicating that more patients are randomised to an experimental arm and $r < 1$ the opposite The primary endpoint is binary (e.g. success versus failure at a particular time point). A number of scenarios were considered $\pi_0 = 0.07$ (very small), $\pi_0 = 0.10$ (to mimic the background study), and $\pi_0 = 0.25$ (central values). Further results for three different values of π_0 are given as supplementary material. Unlike our previous study we allowed the allocation ratio r to vary between $1/4$ and 4 although for most trials $1/3 \leq r \leq 3$ is more realistic. For simplicity we choose

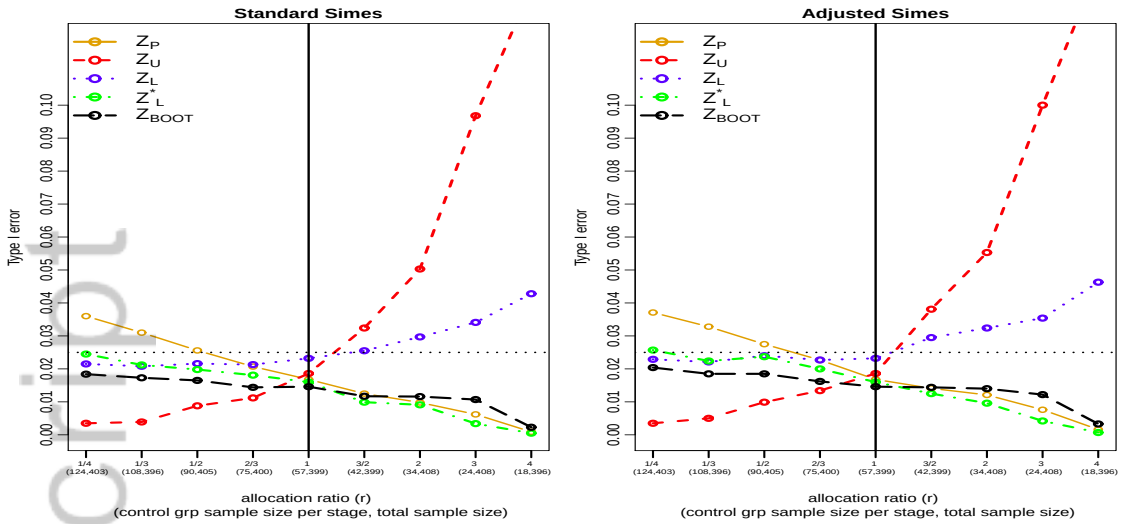

Figure 3. Type I error of various combination tests for a baseline success rate $\pi_0 = 0.07$

the same sample size in the control arm at stage 1 (phase II) and at stage 2 (phase III) but the results presented below are representative of what happens in general.

Figure 3 displays the type I error versus the allocation ratio for $\pi_0 = 0.07$ for a total sample size that remains roughly constant (within the range $[396 - 408]$ to be precise) for each scenario. A fixed overall sample size was chosen to mimic the common requirement of a maximum number of patients to be randomised, irrespective of the loss of power that results. For each allocation ratio, a pair is reported on the x-axis representing the number of control patients per stage and the total number of patients over the two stages. For all scenarios, the nominal size is $\alpha = 0.025$, and the type I error reported is based on $10,000$ runs using both standard Simes' and adjusted Simes' procedures. The overall test in each replication is computed by combining the stagewise p-values using the weighted inverse normal combination function in (3). Similar simulations were also conducted for $\pi_0 = 0.04, 0.10, 0.25$ and are presented in Figures 2-4 of the supplementary material.

The most striking feature in all these plots is that first-order combination tests cannot preserve their nominal level α and display an asymmetric pattern. While Z_P and Z_L are conservative for scenarios where $r > 1$, they become liberal for $r < 1$ (i.e. more patients are randomised to the control group); in contrast, Z_U has the opposite behaviour being extremely liberal (resp. conservative) when $r > 1$ (resp. $r < 1$). This undesirable feature is observed when the success rates are not central, including moderate values such as $\pi_0 = 0.10$, the success rate assumed in the original study. Similar performance is observed for both the Simes or adjusted Simes procedures. This issue is however attenuated for more central baseline rates (e.g. $\pi_0 = 0.25$) apart from Z_U which is totally unreliable for $r > 1$. We did not find any particular problem with balanced designs ($r = 1$), which seems to be in agreement with published results. Second-order combination tests, i.e. the ones based on p^* or the bootstrap p-value, clearly outperform their first-order counterparts and have good size accuracy for all allocation ratios. Although p^* seems to perform well in these simulations, it can break down for value close to the boundary (simple 2 sample test problem only) unlike the bootstrap p-value [31]. For this reason and the fact that the latter is computationally more attractive, our preference goes to the bootstrap p-value approach.

5.3. Power simulations

Most of the discussion so far has focused on the FWER control (as appropriate in confirmatory trials) for the new procedures introduced above. It is equally important to show that they also lead to powerful adaptive tests. Evidence exists that the bootstrap approach has generally good power in the standard 2 arm testing problem. This property is likely to be extended to combination tests. As it is invalid to compare the performance of testing procedures with unequal level, we computed power adjusted for size using ROC curves [42] for all combination tests introduced above. In this simulation of size 5000, all success rates were equal to 0.10 except in one arm (arm 4) where it was increased progressively. Figure 4 displays the empirical power versus the empirical size of all 5 adaptive tests for two alternatives $\pi_4 = 0.20$ and 0.30; panels (a) and (b) correspond to an allocation ratio $r = 1/2$ and panels (c) and (d) to $r = 2$. The overall sample over the two stages was fixed at roughly 400 (actually 405 and 408 for $r = 1/2$ and $r = 2$, the difference being due to the necessary choice of integers for each stage arm size).

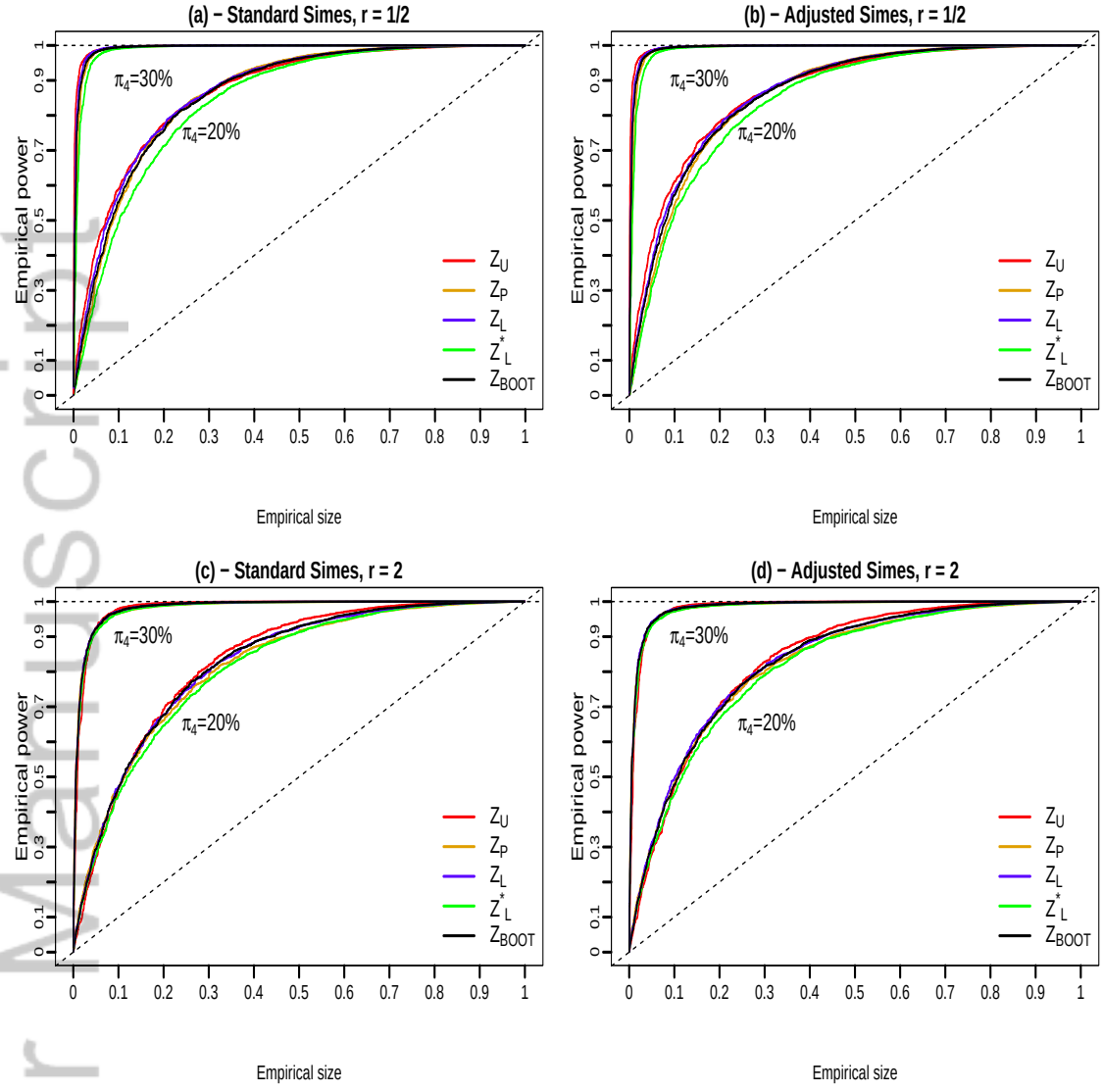

Figure 4. Empirical ROC curves for asymptotic and second-order adaptive tests

The plots show that both second-order procedures Z_L^* and Z_{BOOT} have generally good (size-adjusted) power compared to their standard counterparts with Z_{BOOT} being better than Z_L^* for $r = 1/2$. The key conclusion is that Z_{BOOT} does not sacrifice any of the information in the original statistics; it just calibrates the adaptive procedure correctly. All methods are generally more powerful for $r = 1/2$ than for $r = 2$. The adjustment to the Simes procedure (right side of Figure 4) shows a similar pattern with a slightly improved power of all combination tests for $\pi_4 = 0.20$ compared to the standard Simes'. Similar features were also observed for balanced designs $r = 1$ and other alternatives (results not shown).

6. Examples

6.1. A pediatric dermatology study

We illustrate the performance of the new procedures on the adaptive phase II/III trial Hemangiol in pediatric dermatology [13]. The objective of this trial in infants with infantile hemangioma (IH) was to chose a particular treatment among 4 possible regimens of oral propranolol and carry forward the selected treatment to the next stage. The primary endpoint was complete or nearly complete resolution of target IH at week 24 (W24). Heritier et al. [14] reported the interim analysis results favouring both 6-month regimens with a success rate of 37.5% and 62.8% for the 1 mg/kg/day and 3 mg/kg/day doses, respectively, compared to only 8% in the placebo group. The 3-month regimens did not perform well due to rebounds with respectively success rates of 9.8% and 7.7% for the two doses. All regimens were found safe, and not surprisingly, propranolol 3 mg/kg/day for 6 months was the selected regimen. Assume that the second best dose 1 mg/kg/day, 6 months is chosen due

to safety reasons. The combined p-value using Z_P to feed the combination test is strongly significant ($p < .0001$) and a similar result is obtained using the bootstrap approach. Note that it is actually possible to conduct this analysis thanks to overrun patients (i.e. the sponsor had enrolled all patients that would be needed to carry forward any of the propranolol regimens by interim analysis time [14]). Now, assume we conduct a sensitivity analysis similar to the one carried out in the original study, i.e. a partial reclassification of infants who switched to prohibited treatment prior to the W24 due to intolerance or worsening. Such a request was actually made by the sponsor but not reported upon. As described in the reference above, two cases arise: 1) if a stabilization or a worsening is confirmed by the centralized assessment or if the patient is withdrawn from study therapy for intolerance, the patient remains a failure; 2) If a stabilization or a worsening is not confirmed by the centralized assessment, half of the patients concerned in each treatment group are selected at random and their primary endpoint redefined as a success. In that case, the success rate jumps to 36% (resp 42.5%) in stage 1 and 20% (resp. 59.7%)in stage 2 in the placebo arm (resp. 1 mg/kg/day, 6 month regimen). Taking into account all re-evaluated stagewise p-values that we cannot disclose for confidentiality, the combined p-value is 0.0259 using the pooled z-test and 0.0368 using the bootstrap approach. Both p-values are similar as expected as we are in a situation ($r = 2$) where the standard approach is reliable.

6.2. A possible pancreatic cancer trial

We further illustrate the standard and new procedure in a situation where the pooled z-test is known to be liberal, i.e. when the baseline rate is small and the allocation ratio r smaller than 1, or in other words, more patients are randomised to the placebo group. A ratio of 2:1 in favour of the control arm has been recommended in multi-arm multi-stage (MAMS) adaptive designs such as STAMPEDE [17, 18], a prostate cancer trial, and PanACEA [19], a tuberculosis study. As data are not available for these two ongoing studies, we imagine a hypothetical 2-stage adaptive design in pancreatic cancer patients where 4 experimental treatments A-D are to be compared to standard of care (SOC), i.e. chemotherapy. Pancreatic cancer is a very aggressive form of cancer and a maximum of 20% of patients are eligible for surgery which improves survival. For the remaining 80% of our target population, the median survival rate is low, leading to a 15% survival rate at 1 year, especially as this type of cancer is typically diagnosed late. The primary endpoint is 12 month mortality although survival time is often used in such cancer studies.

We follow the MAMS recommendation and use a ratio of 2.5 to 1 in favour of SOC, possibly due to the cost of the new experimental treatments. The objective is to select the most favourable treatment at interim and proceed to the second stage for further evaluation. The number of patients per arm in the second stage (confirmation) would typically be larger than in the first stage but, to simplify this presentation, we assume 75/30 patients per, respectively, SOC/experimental arm in both stages 1 and 2. The combination function is the inverse normal function (3) with equal weight 0.707 and no stopping for futility or efficacy is considered here. We take the nominal size to be 2.5%.

Table 1. Results of a hypothetical four arm AD in pancreatic cancer with treatment selection at interim

Treatment	**Response**	**Patients**	**Survival**	Z_U	Z_P	Z_L	Z_L^*	Z_{BOOT}
Stage 1	y_i	n_i	**rate at 1 yr**	p-value	p-value	p-value	p-value	p-value
SOC	7	75	9.3%					
A	4	30	13.3%	0.2854	0.2727	0.2769	0.2690	0.2778
B	4	30	13.3%	0.2854	0.2727	0.2769	0.2690	0.2778
C	3	30	10.0%	0.8995	0.8524	0.8702	0.8456	0.4592
D	7	30	23.3%	0.0482	0.0283	0.0339	0.0341	0.0358
Stage 2								
SOC	12	75	16.0%					
D	9	30	30.0%	0.0677	0.0526	0.0576	0.0575	0.0663
Combined				0.0475	0.0227	0.0292	0.0294	0.0346

Suppose that the null hypothesis is true, with common probability of survival at 1 year $\pi_{SOC} = \pi_A = \pi_B = \pi_C = \pi_D = 0.15$. While all tests will likely accept the null, we know that the combination test based on Z_p is liberal under these conditions so there will be data sets where it falsely rejects the null while its bootstrap counterpart does not. Table 1 displays just such a typical outcome (where control counts are within a statistical deviation of the mean $11.3 = 75 \times 0.15$ and the null counts are within a statistical deviation of the mean $4.5 = 30 \times 0.15$). The upper section describes the first stage with four treatment arms. As treatment D appears to be most effective, A-C are stopped and only SOC and D proceed to the next stage for further evaluation. The lower section of the table describes this second stage on selected treatment D. The last row gives the combination test p-value.

For illustration, all five basic test statistics have been calculated, though we recommend only the bootstrap. The differences between these 5 methods are not trivial, especially for the smaller (one-sided) p-values which are of practical interest. The bootstrap p-value is a little larger than Z_P's which is often but not always the case. The p-value based on Z_U is much larger than the others, which is also common. Focusing on the test decision at the one-sided 2.5% level, the combination test based

on Z_P falsely declares treatment D effective ($p = 0.0227$) while all the other adaptive tests find insufficient evidence. The combined one-sided p-value for the unpooled z-test Z_U is as high as $p = 0.0475$ whereas the bootstrap's is $p = 0.0346$.

7. Discussion

We showed in this paper that standard combination tests used in ADs to preserve the FWER may fail to do so when the design is unbalanced and the baseline rates are noncentral (e.g. ≤ 0.10 or ≥ 0.90). The undesirable feature can persist for quite large sample sizes and is caused by poor accuracy of stagewise p-values feeding the combination tests. We focused on the dose selection problem in phase II/III seamless trials with binary endpoints but the problem is quite general and will affect other ADs (e.g. enrichment designs, treatment addition, designs with more than two stages, sample size reassessment etc.). This issue has been overlooked, perhaps due to the common use of balanced adaptive designs (equal allocation ratio) where tests seem better behaved. As the root of the problem lies in having stagewise p-values that can be quite far from $U(0,1)$ under the null, we propose instead to use second-order accurate p-values, e.g. bootstrap p-values. Their computation by enumeration is tractable for rather large sample sizes and very fast R-code is available from the authors. When covariates are available, an adjusted analysis may be performed using logistic regression. The inadequacies of standard methods persist though are less well known. It was shown in [40] that importance sampling provides an elegant solution to computing bootstrap p-values in generalised linear models, which makes the extension to the general AD setting straightforward. In this work, we did not consider Fisher's exact test since our entire approach has been unconditional and Fisher's test is known to be extremely conservative unconditionally. Barnard's unconditional test is also known to be conservative, for quite different reasons and this and other "maximisation type" tests are clearly outperformed by the bootstrap approach in standard testing problems [43]. In our view, there is no particular reason to continue to promote such procedures in the more complex AD setting. Limited simulations with count data (e.g. Poisson distributed random variables) conducted by the authors indicate that this shortcoming is not limited to binary endpoints (although the level distortion does not seem so dramatic for count data). Further investigation is needed to assess the importance of the problem and the possibility to extend the proposed approach to other types of endpoint. Finally, any second-order procedure that provides accurate one-sided p-values could also be considered. More research is needed to determine whether saddlepoint-based testing procedures [44–46] that have recently emerged can also be extended to this setting.

8. Appendix

Each of the p-values are based on a comparison of response rates for a treatment, denoted treatment 1 for simplicity, compared to a control. Again Y_0 (resp. Y_1) is the sum of all individual binary responses over n_0 (resp. n_1) subjects in the control (resp. treatment 1) group; y_0 and y_1 their observed values. The sample proportions are $\hat{\pi}_0 = y_0/n_0$ in the control arm and $\hat{\pi}_1 = y_1/n_1$ in the treatment arm. The pooled proportion is $\tilde{\pi} = (y_0 + y_1)/(n_0 + n_1)$. The test statistics mentioned in section 4 are defined as follows.

Unpooled Z-test:

$$Z_U = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_0(1-\hat{\pi}_0)/n_0}}. \tag{7}$$

Pooled Z-test:

$$Z_P = \frac{\hat{\pi}_1 - \hat{\pi}_0}{\sqrt{\tilde{\pi}(1-\tilde{\pi})(n_1^{-1} + n_0^{-1})}}. \tag{8}$$

Signed Root Likelihood ratio statistic:

$$Z_L = \text{sign}(\hat{\pi}_1 - \hat{\pi}_0)\sqrt{2(\ell(\hat{\pi}_0, \hat{\pi}_1) - \ell(\tilde{\pi}, \tilde{\pi}))}, \tag{9}$$

where the log-likelihood function of the data (y_0, y_1) is $l = \ell(\pi_0, \pi_1)$ given by:

$$l(\pi_0, \pi_1) = y_0 \log \pi_0 + (n_0 - y_0)\log(1-\pi_0) + y_1 \log \pi_1 + (n_1 - y_1)\log(1-\pi_1)$$

Modified statistic Z_L^*:

$$Z_L^* = Z_L + Z_L^{-1} \log(Q/Z_L) \tag{10}$$

where Q is in general complex. For testing the difference between two binomial random variables the general expression [31] for U shows that

$$Q = \frac{(\hat{\eta}_1 - \hat{\eta}_0)\sqrt{\hat{v}_1 \hat{v}_0}}{\sqrt{\tilde{\pi}(1-\tilde{\pi})(n_1^{-1} + n_0^{-1})}}$$

where $\hat{\eta}_i$ is the estimate of $\text{logit}(p_k)$ and $\hat{v}_k = \hat{\pi}_k(1-\hat{\pi}_k)$ for $k = 0, 1$. It is worth pointing out that U is undefined when a count $\hat{\pi}_k$ equals 0 or 1. In particular, it is not defined for the most extreme outcome that $\hat{\pi}_1 = 1$ and $\hat{\pi}_0 = 0$.

Supplementary Materials

Supplementary Figure 1 referenced in Section 2 and supplementary Figures 2-4 referenced in Section 5 are available with this paper on the journal website.

References

1. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **24**(34):3697–3714. DOI: 10.1002/sim.2389
2. Bretz F, Schmidli H, Koenig F, Racine A, Maurer W. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: General Concepts, *Biometrical Journal* 2006; **48**:623–634. DOI: 10.1002/bimj.200510232
3. Schmidli H, Bretz F, Koenig F, Racine A, Maurer W. Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: Applications and Practical Considerations. *Biometrical Journal* 2006; **48**:635–643. DOI: 10.1002/bimj.200510231
4. Jennison C, Turnbull BW. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics* 2007; **17**:1135–1161. DOI: 10.1080/10543400701645215
5. Stallard N, Todd S. Seamless phase II/III designs. *Statistical Methods in Medical Research* 2011; **20**(6):623–34. DOI: 10.1177/0962280210379035
6. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; **18**:1833–48. DOI: 10.1002/(SICI)1097-0258(19990730)18:14¡1833::AID-SIM221¿3.0.CO;2-3
7. Simes RJ. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* 1986; **73**:51–754. DOI: 10.2307/2336545
8. Lloyd CJ. Exact p-values for discrete models obtained by estimation and maximisation. *Australian New Zeland Journal of Statistics* 2008; **50**(4):329–345. DOI: 10.1111/j.1467-842X.2008.00520.x
9. Lloyd CJ. A practical ad hoc adjustment to the Simes p-value *Statistics and Probability Letters* 2012; **82**:1297–1302. DOI: 10.1016/j.spl.2012.03.009
10. Berger RL, Boos DD. P-values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**:1012–1016. DOI: 10.2307/2290928
11. Kutnetsova OM, Tymofyeyev Y. Covariate-adaptive randomisation with unequal allocation. In *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*, Ed. O. Sverdlov, CRC Press, 2015; 171–97.
12. Peckham E, Brabyn S, Cook L, Devlina T, Dumville J, Torgerson DJ. The use of unequal randomisation in clinical trials An update. *Contemporary Clinical Trials* 2015; **45**:113–122. DOI: 10.1016/j.cct.2015.05.017
13. Léauté-Labrèze C., Hoeger, P., Mazereeuw-Hautier, J., Guibaud, L., Baselga, E., Posiunas, G., Phillips, R.G. et al. *A randomized controlled trial of oral propranolol in infantile hemangioma*. New England Journal of Medicine 2015; **372**(8):735–746. DOI: 10.1056/NEJMoa1404710
14. Heritier S, Morgan-Bouniol C, Lô SN, Gautier S, Voisard JJ. . A single pivotal adaptive trial in infants with proliferating hemangioma: rationale, design challenges, experience and recommendations. In *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*. Ed. O. Sverdlov, CRC Press 2015; 453–82.
15. Torgerson DJ, Campbell MK. Economics notes: use of unequal randomisation to aid the economic efficiency of clinical trials. *British Medical Journal* 2000; **321**(7263):759.
16. Dumville JC, Hahn S, Miles JNV, Torgerson, DJ. The use of unequal randomization in clinical trials. *Contemporary Clinical Trials* 2006; **27**:1–12. DOI: 10.1016/j.cct.2005.08.003
17. Sydes MR, Parmar MKB, James ND, Clarke NW, Dearnaley DP, Mason MD, Morgan RC, SandersK, Royston P. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials* 2009; **10**:39–45, DOI: 10.1186/1745-6215-10-39.
18. James ND, Sydes MR et al. for the the STAMPEDE investigators. Addition of docetaxel, zoledronic acid, or both to fi rst-line long-term hormone therapy in prostate cancer (STAMPEDE): survival results from an adaptive, multiarm, multistage, platform randomised controlled trial. *The Lancet* 2016; **387**:1163–77, DOI: 10.1016/S0140-6736(15)01037-5.
19. Boeree MJ, Heinrich N. et al. on behalf of the PanACEA consortium. High-dose rifampicin, moxifl oxacin, and SQ109 for treating tuberculosis: a multi-arm, multi-stage randomised controlled trial *Lancet Infectious Diseases* 2017; **17**:39–49, DOI: 10.1016/S1473-3099(16)30274-2.
20. Sarkar S, Chang CK. The Simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association* 1997; **92**:1601–1608. DOI: 10.2307/2965431
21. Sarkar S. Probability inequalities of ordered MTP2 random variables: a proof of the Simes conjecture. *Annals of Statististics* 1998; **26**:494–504.
22. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660. DOI: https://doi.org/10.1093/biomet/63.3.655
23. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Tutorial in Biostatistics: Adaptive designs for confirmatory clinical trials. *Statisics in Medicine* 2009; **28**:1181–1217. DOI: 10.1002/sim.3538
24. Bauer P, Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:029–1041.
25. Bauer P. Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie*, 1989; **20**:130–148.
26. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statical Association* 2002; **97**:236–244. DOI: 10.1198/016214502753479374

27. Fisher RA. *Statistical Methods for Research Workers*, 4th Ed. Oliver and Boyd: London, 1932.
28. Mosteller F, Bush RR. *Selected quantitative techniques.* In *Handbook of Social Psychology.* Ed. G. Lindzey, Addison-Wesley: Cambridge MA, 1954; 289–334.
29. Lehmacher W, Wassmer G. Adaptive sample size calculation in group sequential trials. *Biometrics* 1999; **55**:1286–1290. DOI: 10.1111/j.0006-341X.1999.01286.x
30. Davison AC. *Statistical Models.* Cambridge University Press, 2003. DOI: https://doi.org/10.1017/CBO9780511815850
31. Lloyd CJ. Bootstrap and second-order tests of risk difference. *Biometrics* 2010; **66**:975–982. DOI: 10.1111/j.1541-0420.2009.01354.x
32. Heritier S, Lô SN, Morgan CC. An adaptive confirmatory trial with interim treatment selection: practical experiences and unbalanced randomisation. *Statistics in Medicine* 2011; **30**:1541–1554. DOI: 10.1002/sim.4179
33. Reid N. Asymptotics and the theory of inference. *Annal of Statistics* 2003; **31**:1695–1731.
34. Young GA, Smith RL. *Essentials of statistical inference.* Cambridge University Press, 2005.
35. Brazzale AR, Davison AC, Reid N. *Applied Asymptotics Case Studies in Small-Sample Statistics.* Cambridge University Press, 2007.
36. DiCiccio TJ, Young GA. Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika* 2008; **95**:747–58. DOI: https://doi.org/10.1093/biomet/asn011
37. Lugannani R, Rice S. Saddlepoint approximation to the distribution of the sum of independent random variables. *Advances in Applied Probability* 1980; **12**:475–490.
38. Barndorff-Nielsen O, Cox DR. *Inference and Asymptotics.* London. Chapman and Hall, 1994.
39. Lee SMS, Young GA. Parametric bootstrapping with nuisance parameters. *Statistics Probability Letters* 2005; **71**:143–153. DOI: http://dx.doi.org/10.1016/j.spl.2004.10.026
40. Lloyd CJ. Computing highly accurate or exact *p*-values using importance sampling. *Computational Statistics and Data Analysis* 2012; **56**(6):1784–1794. DOI: 10.1016/j.csda.2011.11.003.
41. Lloyd CJ. On the exact size of tests of treatment effects in multi-arm clinical trials. *Australian & New Zealand Journal of Statistics* 2014; **56**:359-369. DOI: 10.1111/anzs.12089.
42. Lloyd CJ. On comparing the accuracy of competing tests of the same hypotheses from simulation data. *Journal of Statistical Planning and Inference* 2005; **128**:97–508. DOI: 10.1016/j.jspi.2003.12.002
43. Lloyd CJ. Exact *p*-values for discrete models obtained by estimation and maximization. *Australian & New Zealand Journal of statistics* 2008; **50**(4):329–345. DOI: 10.1111/j.1467-842X.2008.00520.x.
44. Lô SN, Ronchetti E. Robust and accurate inference for generalized linear models. *Journal of Multivariate Analysis* 2009; **100**(9):2126–36. DOI: 10.1016/j.jmva.2009.06.012
45. Ma Y, Ronchetti E. Saddlepoint test in measurement error models. *Journal of the American Statistical Association* 2011; **106**(493):147–56. DOI: 10.1198/jasa.2011.tm10031.
46. Aeberhard WA, Cantoni E, Heritier S. Saddlepoint tests for accurate and robust inference on overdispersed count data. *Computational Statistics and Data Analysis* 2017; **107**:162-75. DOI: 10.1016/j.csda.2016.10.009.