RESEARCH ARTICLE

# Many-to-one comparisons after safety selection in multi-arm clinical trials

**Gerald Hlavin[1], Lisa V. Hampson[2,3], Franz Koenig[1]***

**1** Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria, **2** Statistical Innovation, Advanced Analytics Centre, AstraZeneca, Cambridge, United Kingdom, **3** Department of Mathematics & Statistics, Lancaster University, Lancaster, United Kingdom

* franz.koenig@meduniwien.ac.at

OPEN ACCESS

**Citation:** Hlavin G, Hampson LV, Koenig F (2017) Many-to-one comparisons after safety selection in multi-arm clinical trials. PLoS ONE 12(6): e0180131. https://doi.org/10.1371/journal.pone.0180131

**Editor:** Wenge Guo, New Jersey Institute of Technology, UNITED STATES

**Received:** January 24, 2017

**Accepted:** June 10, 2017

**Published:** June 26, 2017

**Data Availability Statement:** This is a methodological paper which is not based on empirical data sets.

**Funding:** This project [GH, FK] has received funding from the European Union's Seventh Framework Programme [FP7 2007-2013] under grant agreement no. 602552 for research, technological development, and demonstration (IDEAL - Integrated Design and Analysis of small population group trials). This funder provided support in the form of salaries for authors [GH, FK], but did not have any additional role in the

## Abstract

In phase II platform trials, 'many-to-one' comparisons are performed when *K* experimental treatments are compared with a common control to identify the most promising treatment(s) to be selected for Phase III trials. However, when sample sizes are limited, such as when the disease of interest is rare, only a single Phase II/III trial addressing both treatment selection and confirmatory efficacy testing may be feasible. In this paper, we suggest a two-step safety selection and testing procedure for such seamless trials. At the end of the study, treatments are first screened on the basis of safety, and those deemed to be sufficiently safe are then taken forwards for efficacy testing against a common control. All safety and efficacy evaluations are therefore performed at the end of the study, when for each patient all safety and efficacy data are available. If confirmatory conclusions are to be drawn from the trial, strict control of the family-wise error rate (FWER) is essential. However, to avoid unnecessary losses in power, no type I error rate should be "wasted" on comparisons which are no longer of interest because treatments have been dropped due to safety concerns. We investigate the impact on power and FWER control of multiplicity adjustments which correct efficacy tests only for the number of safe selected treatments instead of adjusting for all *K* null hypotheses the trial begins testing. We derive conditions under which strict control of the FWER can be achieved. Procedures using the estimated association between safety and efficacy outcomes are developed for the case when the correlation between endpoints is unknown. The operating characteristics of the proposed procedures are assessed via simulation.

## 1 Introduction

Clinical trials are often complex and costly undertakings, and many patients may be required to reach a conclusion on whether a new therapy is efficacious. One way to reduce the logistical burden of drug development is to simultaneously study *K* novel treatments within a single multi-arm trial which also includes a common control arm. Comparing treatments with the same control group can yield substantial reductions in sample size because only one control is

study design, data collection and analysis, decision to publish, or preparation of the manuscript. LVH is an employee of AstraZeneca and Lancaster University, and her contribution was partly funded by the Medical Research Council MRC (grant MR/M013510/1). AstraZeneca (AZ) provided financial support in the form of salaries for authors [LVH], but AZ and MRC did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

**Competing interests:** This project [GH, FK] has received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement no 602552 (IDEAL - Integrated Design and Analysis of small population group trials). LVH is an employee of AstraZeneca and Lancaster University, and her contribution was partly funded by the Medical Research Council MRC (grant MR/M013510/1).

needed compared with the $K$ control groups that would be needed by $K$ independent two-arm trials [1]. Such Phase II platform trials are particularly appealing as an efficient way to identify effective treatments for small populations, when reasonable numbers of patients can be challenging to recruit [2, 3]. In principle, multi-arm trials are appropriate for evaluating not only different treatments but also different patient subgroups, doses, etc.

Once promising treatments have been identified, they are then traditionally taken forwards to separate, confirmatory, Phase III trials. In this paper, we consider the design of a single-stage Phase II/III trial intended to address both treatment selection *and* confirmatory efficacy testing. Traditionally, confirmatory trials must ensure the family-wise error rate (*FWER*), defined as the probability of claiming efficacy for at least one ineffective treatment, is controlled at a suitably small level, which we denote by $\alpha_{nom}$ [4] (for example, $\alpha_{nom} = 0.05$ or $\alpha_{nom} = 0.025$ for two-sided and one-sided tests, respectively), for every configuration of ineffective treatments. The chance of making a false claim of efficacy for at least one treatment will increase as $K$ increases if the null hypothesis associated with each experimental treatment is naively tested at local level $\alpha_{nom}$ [5]. Multiple comparison procedures are regularly applied to correct the local significance level at which each hypothesis test is performed to ensure that strong control of the *FWER* level is maintained. When the primary efficacy endpoint is continuous, a standard parametric procedure which accounts for the correlation structure induced by the common control is the Dunnett test [6]. Under this and other methods (see, for example, [7]), the reduction in the local significance level becomes more severe as the number of comparisons increases [8–10].

Safety is of paramount importance in all phases of drug development. Suppose that at the end of the Phase II/III trial, treatments may be abandoned due to safety concerns before being tested for efficacy. Should one or more treatment arms be dropped in this way, when testing the efficacy of the remaining treatments relative to control it would seem natural to perform multiplicity corrections based only on the number of remaining treatments. Intuitively, it seems unreasonable to spend parts of the overall nominal significance level on treatments for which null hypotheses are not tested.

We will demonstrate that, in general, such a procedure (which selects safe treatments in a data-driven way and then performs a multiplicity correction based on the remaining treatments) will not control the *FWER*. However, we state sufficient conditions under which this procedure does control the *FWER*; this will provide reassurance to clinical trialists should they adopt designs satisfying these criteria. Furthermore, we propose a testing procedure which maintains adequate FWER control should the sufficient conditions fail to hold. Throughout this manuscript we restrict attention to the case where toxicity and efficacy are both measured by normally distributed variables (where lower toxicity but higher efficacy measurements are more favourable), a framework already used in similar settings [11–13].

Continuous safety parameters are often closely connected to the treatment of a patient. For example spironolactones that are used in the treatment of patients with high blood pressure with spironolactones, attention has to be paid on blood potassium levels, where increased levels can lead to abnormal heart rhythms and in severe cases to cardiac arrests [14]. Another example concerns treatments with agomelatine, a melatonergic antidepressants. Raised levels of transaminases which indicate liver damage, are a contraindication and a prespecified reason to discontinue a treatment using Valdoxan (a trade name of agomelatine) [15]. Another relevant continuous safety parameter in drug development is the assessment of the QT-time, a time interval in the electrical cycle of the heart [16]. A QT-time prolongation is especially relevant in psychoactive drugs [17]. In all of the aforementioned examples, certain continuous parameters are important to judge the safety and therefore the usefulness of a medicine.

On the other hand multi arm trials are increasingly demanded in modern clinical research [1–3]. In [18] an extensive list on multi arm trials published in 2009 is included, where 221 trials with 3 or more arms reported to have a parallel group design. Another example of a multi arm trial is [19], a study of Eplerenone (a selective aldosterone blocker intended for treatment of hypertension), where 417 patients were randomized to 6 Eplerone arms, 1 Spironolactone arm, and placebo. There the comparisons to placebo (or aldosterone) in serum aldosterion levels, total plasma renin, and active plasma renin were done using the Dunnett test. As an important tolerability endpoint, mean changes in blood potassium levels were measured. A further example is a study of mixed amphetamine salts with extended releae for the treatment of Attention-Deficit/Hyperactivity Disorder (ADHD) [20]. Efficacy comparisons of 3 treatment arms to placebo were again done using the Dunnett test. Besides adverse events, safety analysis included the measurements several continuous safety variable, e.g. mean colesterol decreases and ECG measurements like the aforementioned QT-time.

These examples show the potential of uses of our two-step approach: the structured combination of multi-arm trials with safety considerations. In S1 Example, a numerical example will illustrate this approach.

In Section 2, we present a motivating example which illustrates the testing problem at hand and introduces the terminology used throughout this article. In Section 3, general notation is defined and the bivariate normal data model is presented. The concept of maximum *FWER* inflation is explained in Section 4. There we also present simulation results which show how maximum FWERs vary with the correlation between normally distributed toxicity and efficacy measurements, to see what can go wrong by performing safety selection before efficacy testing. For the procedure basing multiplicity corrections on the number of selected treatments, it is proven that the maximum *FWER* inflation is monotonically increasing for decreasing correlations, and therefore the maximum occurs when the correlation is −1. Then the main result of this paper is presented, which states that so long as toxicity and efficacy have a non-negative correlation, safety selection can be applied without inflating the *FWER*. The proof of this result can be found in the Appendix. In Section 5, adjustments are proposed for the case when a positive (or zero) correlation cannot be assumed. A simulation study in Section 6 compares different methods with a conservative procedure which always corrects the local signficance levels of hypothesis tests for all *K* pre-specified comparisons.

## 2 Motivating example

Suppose we have three treatment groups indexed by $i \in \{1, 2, 3\}$ with corresponding expected efficacy outcomes $\mu_1$, $\mu_2$, and $\mu_3$, which we wish to compare with a common control (indexed by $i = 0$) on which the mean efficacy outcome is $\mu_0$. Treatment groups may represent, for example, different treatment regimens, drugs, or drug levels. For $i \in \{1, 2, 3\}$, the elementary null-hypothesis $H_i$: $\mu_i - \mu_0 \leq 0$ which states that treatment $i$ is not more efficacious than control, will be tested against the alternative $H_i^A : \mu_i - \mu_0 > 0$. We reject $H_i$ if the corresponding test statistic exceeds a pre-specified threshold denoted by $b_i$. Let us assume that the FWER is to be controlled at a pre-specified level $\alpha_{\text{nom}}$. Then, the decision thresholds $b_1$, $b_2$, and $b_3$ have to be chosen such that under any configuration $J \subseteq \{1, 2, 3\}$ of true null-hypotheses (indexing treatments which are not efficacious relative to control), the probability of rejecting at least one true null-hypothesis is less or equal to $\alpha_{\text{nom}}$. For ease of presentation, suppose that to achieve FWER control we apply the simple Bonferroni correction. Then we split the nominal level $\alpha_{\text{nom}}$ equally among all elementary null-hypotheses $H_i$, i.e., each $H_i$ is tested separately at local significance level $\alpha_{\text{nom}}/3$. In subsequent sections, we will consider the Dunnett test to correct for multiple hypothesis testing; however, even under this approach, the arguments presented in this example remain essentially unchanged.

Safety, as well as efficacy, is an important issue in clinical drug development. If a treatment appears to be unacceptable on the basis of observed toxicity data, should it still be considered for efficacy testing? If the answer is no, then the next question is how can we select on the basis of safety and still control the FWER for efficacy testing? To address this, let us continue with our example and suppose that at the end of the trial, treatment 3 is considered unsafe. Then it would seem unnatural to test the remaining hypotheses $H_1$ and $H_2$ at local significance level $\alpha_{\text{nom}}/3$, since this wastes 1/3 of $\alpha_{\text{nom}}$ on a treatment comparison that is no longer of interest. Although a Bonferroni correction based on the initial number of planned treatment-control comparisons would be valid in terms of FWER control, regardless of how treatments are selected, it may result in a conservative procedure. A more natural and less conservative approach would be to correct for the number of treatment-control comparisons that are actually performed after safety selection; in our example, this corresponds to testing $H_1$ and $H_2$ at a multiplicity-adjusted level of $\alpha_{\text{nom}}/2$. We will refer to the former correction for initially planned treatment-control comparisons as the '*conservative correction*' and the latter approach as the '*natural correction*'.

For the remainder of this paper, we will formally distinguish between the *safety selection step*, where unsafe treatments are dropped, and the *efficacy testing step*, where the remaining sufficiently safe treatments are each compared with a common control. The *overall two-step procedure* is then the combination of both steps. Here it is implied, that for each selection of treatments a multiple testing procedure is defined for this selection.

First focusing on such a multiple testing procedure in the efficacy testing step for treatments defined by the index set $S \subseteq \{1, 2, 3\}$, let $FWER(S, J)$ denote its *unconditional* (i.e. not conditioning on any safety data and thus on how $S$ was derived) probability for rejecting at least one of the hypotheses $H_i$, $i \in S \cap J$. Here $J$ again denotes the set of true null-hypotheses as defined above. We say that for the efficacy testing step, FWER control is maintained at level $\alpha_{\text{nom}}$ if $FWER(S) := \max_{J \subseteq \{1, 2, 3\}} FWER(S, J) \leq \alpha_{\text{nom}}$ for all possible subsets $S$. Given a predefined safety selection rule, $FWER_o$ denotes the $FWER$ of the overall two-step selection and testing procedure. We say that $FWER$ control for the overall two-step procedure is maintained at level $\alpha_{\text{nom}}$ if $FWER_o \leq \alpha_{\text{nom}}$.

In our example, testing each null hypothesis indexed by $S \subseteq \{1, 2, 3\}$ at significance level $\alpha_{\text{nom}}/|S|$ (here $|\cdot|$ denotes the cardinality of $S$) leads to $FWER(S) \leq \alpha_{\text{nom}}$, thus $FWER$ is controlled in the efficacy testing step at level $\alpha_{\text{nom}}$. However for particularly structured data this does not imply $FWER_o$ control. The intuition here is, that if the selection of a treatment for being tested for efficacy is associated with increased efficacy observations, $FWER_o$ can be increased. This becomes particularly obvious, if for example an (unreasonable) selection procedure would always select just the most promising treatment in terms of efficacy and therefore no correction of the significance level is performed. In contrast, $FWER$ control in the efficacy testing step implies $FWER_o$ control for the conservative correction (where null hypotheses for selected treatments are tested at level $\alpha_{\text{nom}}/3$) because rejection regions in this case do not depend on which and how many other treatments are selected. Therefore the inclusion of a selection procedure that precedes efficacy testing, can only reduce the number of false positive findings.

The main focus of this manuscript is on defining conditions under which the natural correction in the efficacy testing step leads to $FWER_o$ control, and on determining how to adjust this correction to ensure $FWER_o$ control when these conditions do not hold.

## 3 Model specification

As a general framework, we begin by assuming efficacy and toxicity data are normally distributed.

### 3.1 General framework

Let $I := \{1, \ldots, K\}$ be the set of all experimental treatments. Denoting the average efficacy outcome in group $i$ as $\mu_i$, $H_i$: $\mu_i - \mu_0 \leq 0$, $i \in I$ defines the elementary hypothesis stating that treatment $i$ cannot be regarded as more efficacious than the control (treatment 0).

We define the indicator $\tau_i$, where $\tau_i = 1$ when treatment $i$ is excluded from efficacy testing on the basis of the observed toxicity data, whereas $\tau_i = 0$ indicates that treatment $i$ is selected for efficacy testing. Then

$$\varsigma_S := \prod_{i \in S}(1 - \tau_i) \cdot \prod_{j \in I \setminus S} \tau_j$$

indicates the event that exactly the treatments in the (sub-)set $S \subseteq I$ are selected for efficacy testing, whereas the remaining treatments in $I \setminus S$ are dropped due to safety concerns.

At the end of the two-stage procedure, we represent the test decision on the rejection of an elementary null hypothesis $H_i$ by a binary indicator. As we have seen from our motivating example, the final decision rule for a treatment will depend on the safety selection step. For instance, the natural multiplicity correction depends on the actual number of treatments selected for efficacy testing, that is, $|S|$. For each selection set $S \in I$, let $\eta_i^S$ for $i \in S$, denote a binary indicator, where $\eta_i^S = 1$ denotes rejection of the $i$th null hypothesis, otherwise $\eta_i^S = 0$. Here the definition of $\eta_i^S$ depends on the selection subset, which is reflected by the superscript $S$. Since unsafe treatments are not subject to efficacy testing, we set $\eta_i^S = 0$ for $i \in I \setminus S$, without performing a statistical test of $H_i$.

Since $J \subseteq I$ denotes the set of all true null-hypotheses,

$$\varphi_J^S := 1 - \prod_{i \in S \cap J}(1 - \eta_i^S)$$

indicates a type I error has been committed for one of the selected but inefficacious treatments. The *FWER* is controlled in the efficacy testing step at level $\alpha_{\text{nom}}$ if, for each possible selection set $S \subseteq I$ and each configuration of true null-hypotheses $J$, it holds that $\mathbb{E}[\varphi_J^S] = FWER(S) \leq \alpha_{\text{nom}}$, where $FWER(S) := \max_{J \subseteq I} FWER(S, J)$. For the overall two-step procedure, $FWER_o$ is controlled at the nominal level if

$$\sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S \cdot \varphi_J^S] \leq \alpha_{\text{nom}}. \qquad (1)$$

Here $\mathcal{M}_J := \{S \subseteq I : S \cap J \neq \emptyset\}$ contains all possible selections with at least one true null hypothesis.

If $\varsigma_S$ and $\varphi_J^S$ are independent, then the addends in Eq (1) would decompose to $\mathbb{E}[\varsigma_S \cdot \varphi_J^S] = \mathbb{E}[\varsigma_S]\mathbb{E}[\varphi_J^S]$. In this case, *FWER* control in the efficacy testing step at level $\alpha_{\text{nom}}$ implies $FWER_o$ control at the same level, since the sum of the $\mathbb{E}[\varsigma_S]$ is less than or equal to one and therefore Eq (1) holds. In general we have

$$\begin{aligned}\mathbb{E}[\varsigma_S \cdot \varphi_J^S] &= P(\varsigma_S = 1, \varphi_J^S = 1)\\ &= P(\varsigma_S = 1|\varphi_J^S = 1) \cdot P(\varphi_J^S = 1)\\ &= \mathbb{E}[\varsigma_S|\varphi_J^S = 1]\mathbb{E}[\varphi_J^S],\end{aligned}$$

but the $\mathbb{E}[\varsigma_S|\varphi_J^S = 1]$ do not necessarily sum to a value less than or equal to one.

## 3.2 The bivariate normal model

For the control group ($i = 0$), all treatment groups $i \in I$, and all patients $j \in \{1, \ldots, n_i\}$, we consider a two dimensional vector of measurements ($x_{ij}$, $y_{ij}$) coming from a bivariate normal distribution $N((\mu_i, \theta_i), \Sigma)$ with $\Sigma = \begin{bmatrix} \sigma_i^2 & \rho_i\sigma_i\omega_i \\ \rho_i\sigma_i\omega_i & \omega_i^2 \end{bmatrix}$. Here $x_{ij}$ and $y_{ij}$ denote an efficacy and a toxicity measurement, respectively. Higher values for $x_{ij}$ and $y_{ij}$ shall be interpreted as higher efficacy and toxicity, respectively. Pairs of outcomes ($x_{ij}$, $y_{ij}$) are assumed to be independent across values of $i$ and $j$. In what follows, let $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the $n_i$-dimensional vectors of efficacy and toxicity outcomes measured in group $i$, respectively.

For the safety selection stage, $\tau_i := \mathbf{1}_{\{(b_i^y,\infty)\}}(T_i^y(\mathbf{y}_i))$ is defined before the start of the trial. Here $T_i^y$ is a pre-specified function of the group $i$ toxicity data which is compared with the predefined threshold $b_i^y$ (here $\mathbf{1}$ denotes the indicator function, so that $\tau_i = 1$ when $b_i^y < T_i^y(\mathbf{y}_i)$ and 0 otherwise).

For the purposes of efficacy testing, data are summarised by the test statistics $T_i^x$ and the decision for null hypothesis $H_i$ is given by $\eta_i^S := \mathbf{1}_{\{(b_{i,S}^x,\infty)\}}(T_i^x(S, D))$ with pre-defined thresholds $b_{i,S}^x$ and $D := \{\mathbf{x}_0, \ldots, \mathbf{x}_K\}$. In our framework these critical thresholds are chosen such that the $FWER(S)$ is controlled (and therefore depend on the number of treatments in the selected set $S$).

In what follows, when determining the efficacy success criteria we will focus on an important multiple testing procedure in the many-to-one comparisons framework of normally distributed variables, namely the Dunnett Test. We will consider the case of equal group variances (an assumption that will be discussed in Remark 2 in the Appendix), that is $\sigma_i^2 = \sigma^2 \, \forall i \subseteq I \cup \{0\}$, for both the Dunnett test for unknown (DT) and known variances (ZT), where in the latter case hypothesis test decisions are based on Z-statistics. For both DT and ZT, $FWER(S)$ is easily determined, since on the one hand the dependence on $S$ is only due to the number of treatments in $S$, on the other hand in these cases the simplification of above definition $FWER(S) = FWER(S, I) = FWER(S, S)$ holds. The straightforward definition of DTs or ZTs for every subset $S$ by using $|S|$ therefore ensures FWER control in the efficacy testing step. Similar arguments apply also for the Bonferroni correction.

## 3.3 Example (Cont.)

We continue the motivating example in Section 2, but assume a Dunnett test (with known variance) at one-sided level $\alpha_{\text{nom}} = 0.025$) instead of a Bonferroni correction to correct for multiple comparisons.

In our notation the selection of treatments 1 and 2 would be implied by observing $T_i^y(\mathbf{y}_i) \leq b_i^y$ for $i \in \{1, 2\}$ and $b_j^y < T_j^y(\mathbf{y}_j)$ for $j = 3$. Applying the natural correction in the efficacy testing step would now require the Dunnett test for two treatments. Assuming for example a variance of $\sigma_i = \sigma = 1$ and a group sample size of $n_i = 22$ for $i \in \{0, 1, 2\}$ requires boundaries of $b_1^x = b_2^x = 2.21$. In terms of p-values this would correspond to $\alpha$-levels

$\alpha_1 = \alpha_2 = 0.0135$. In comparison the conservative procedure would still use the boundaries $b_1^x = b_2^x = 2.35$ (corresponding to $\alpha_1 = \alpha_2 = 0.0094$), which are adequate for a Dunnett test with three treatment-control comparisons

## 4 FWER of the two-step procedure

In this section the relationship between the within patient correlation of safety and efficacy measurements and the $FWER_o$ is investigated. First, simulation results will be presented for the case of comparing two treatments groups with a single control and the maximum $FWER_o$ will be calculated for different correlations. Then we will present sufficient conditions which ensure $FWER_o$ control of the overall two-step procedure.

### 4.1 FWER maximizing correlation

We will show that $FWER_o$ control does not necessarily hold if the natural correction is applied to negatively correlated toxicity and efficacy data. In our scenario, the nominal one-sided significance level was fixed at $\alpha_{\text{nom}} = 0.025$. We will consider comparisons of $K$ treatment groups with a common control assuming patient responses follow the bivariate normal model defined in Section 3, setting $\sigma_i^2 = 1$, for $i = 0, 1, \ldots, K$, and $\omega_i^2 = 1$, for $i = 1, \ldots, K$, without loss of generality. For simplicity, we also assume equal correlations $\rho_1 = \ldots = \rho_K = \rho$. Basing efficacy decisions on one-sided two-sample z-tests, to find a clinically relevant effect of $\mu_d = 1$ with power $1 - \beta = 0.9$, the per-group sample size is set to $n = 2(z_{1-\alpha} + z_{1-\beta})^2/\mu_d^2$.

Similar to [12], we calculate the worst case $FWER_o$ of our test procedure for the situation when all expected responses are zero, that is, when $\mu_i = 0$, for $i = 0, \ldots, K$. This is the situation where for the classical Dunnett test the FWER is maximized. Furthermore we set $\theta_i = 0$, for $i = 1, \ldots, K$. We then search across values of the toxicity thresholds $b_1^y, \ldots, b_K^y$ for monitoring test statistics $T_i^y(\mathbf{y}_i) := \bar{\mathbf{y}}_i$ to find the configuration at which the maximum (worst case) $FWER_o$.

In Fig 1 (left), values of $FWER_o$ maximized for all toxicity parameter configurations and thresholds are shown for different numbers of treatments $K \in \{2, 3, 4\}$ and different correlations $\rho \in (-1, 1)$ in the ZT situation (where the variances are known). The worst case $FWER_o$ is maximized at $\rho = -1$ and decreases for increasing correlation until $\rho = 0$ is reached. For non-negative correlation, the worst case $FWER_o$ is constant at the nominal level $\alpha_{\text{nom}}$.

Fig 1 (left) suggests that the worst case $FWER_o$ inflation is maximized at $\rho = -1$. The following theorem provides a general statement about this observation. Here it is again assumed that $T_i^y = \bar{\mathbf{y}}_i$, but similar results are also straightforward to prove so long as after transformation of the data, a multivariate normal distribution is the resulting model.

**Theorem 1**. *Consider the bivariate normal model in Eq (3.1) and let $\emptyset \neq J \subseteq I$ be the set indexing inefficacious treatments. Let safety selection be based on the group mean toxicities $\bar{\mathbf{y}}_i$. Then the worst case $FWER_o$ for the ZT and the DT is reached at $\rho_i = -1, \forall i \in \{1, \ldots, K\}$.*

A proof of this result can be found in the Appendix.

### 4.2 FWER control for positive correlation

Fig 1 (left) indicates that a worst case $FWER_o$ inflation will not occur if the correlation between toxicity and efficacy responses is non-negative. Using the concept of association introduced in [21], we can generalize this observation with the following theorem:

**Theorem 2**. *Consider the bivariate normal model in Eq (3.1) and let $\emptyset \neq J \subseteq I$ be the set indexing inefficacious treatments. Let safety selection be based on the group mean toxicities $\bar{\mathbf{y}}_i$. If*

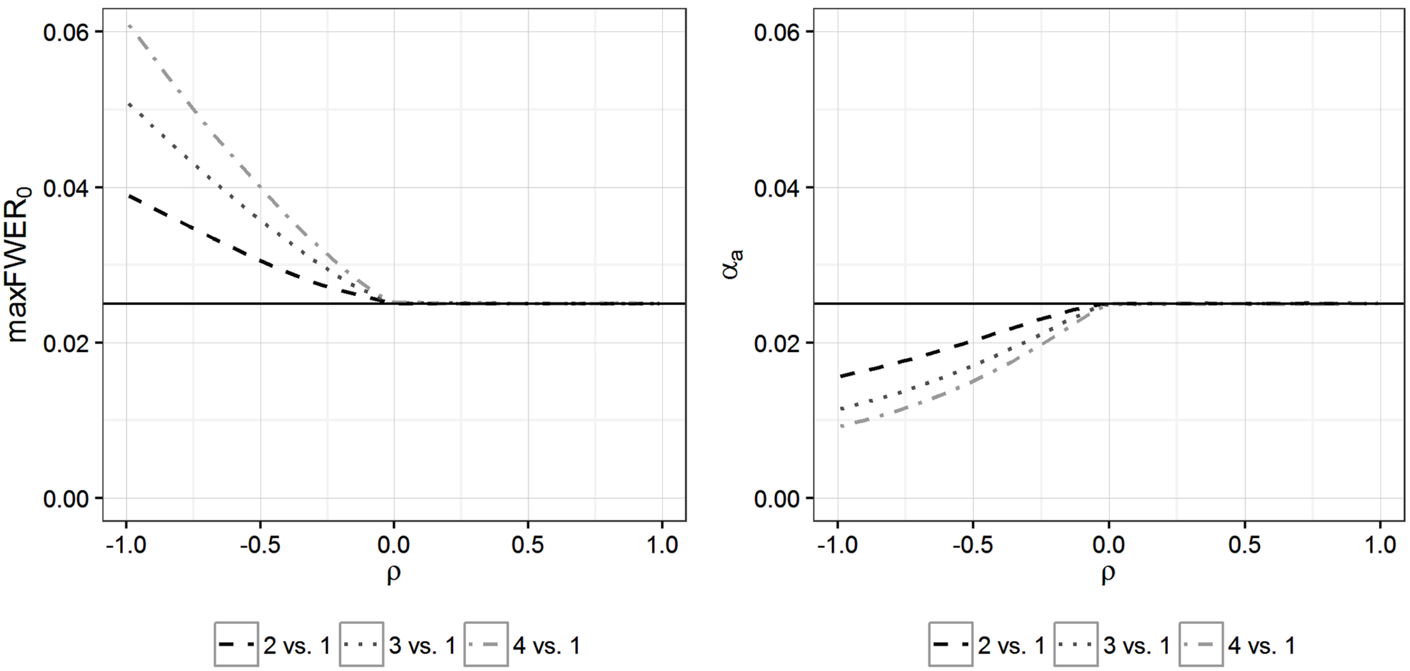

**Fig 1. Maximum *FWER_o* (left) and corresponding adjusted nominal levels (right).** In the left figure, worst case $FWER_o$ for varying correlations $\rho$ between toxicity and efficacy are drawn for the ZT for 2, 3, and 4 treatments. The maximum $FWER_o$ is constant for values of $\rho \geq 0$ and increases for decreasing $\rho < 0$. At $\rho = -1$ the maximum worst case $FWER_o$ is reached. In the right figure, the corresponding adjusted nominal levels $\alpha_a$ to be used in a ZT correcting for the number of selected treatments are presented as a function of the true group correlations $\rho$ (as described in Section 5.1). Positive correlation implies $FWER_o$ control, in which case $\alpha_a = \alpha_{\text{nom}}$. In case of negative correlation the $FWER$ in the efficacy testing step has to be controlled at a specific level $\alpha_a < \alpha_{\text{nom}}$ to ensure control of the worst case $FWER_o$ at level $\alpha_{\text{nom}}$.

https://doi.org/10.1371/journal.pone.0180131.g001

*$\rho_i \geq 0, \forall i \in I$, and $FWER(S) \leq \alpha_{nom}$ holds $\forall S \subseteq I$, then the FWER of the overall procedure is also controlled at level $\alpha_{nom}$ for the ZT and the DT.*

The proof of Theorem 2 is presented in the Appendix, together with a discussion of possibilities and limitations of extensions.

## 5 Correlation-based adjustments

We have seen that when applying the natural correction, the $FWER_o$ depends on the true correlation $\rho_i$. For known non-negative correlation it is reasonable to prefer the natural to the conservative correction, since the former approach achieves higher power, while the $FWER_o$ is still controlled. To compare $FWER_o$ for different procedures we write $FWER_o^{\text{NA}}(\rho)$ when using the natural correction and when the true correlation is $\rho$ (NA stands for natural correction). Here the assumption is equality of correlations across all treatment groups or, if equality cannot be assumed, we substitute the minimum correlation between toxicity and efficacy found across all groups. Using the monotonicity of the worst case $FWER_o$ as a function of the true correlation as discussed in the context of Theorem 1, this choice of $\rho$ then serves to derive an upper bound for the max$FWER_o^{NA}$. For the conservative procedure, we write $FWER_o^{\text{CO}}$, where CO stands for *conservative*. It is clear that max$FWER_o^{\text{NA}}(\rho) \geq$ max$FWER_o^{\text{CO}}$ for all correlations $\rho$.

In this section our goal is to define an adjustment to the natural multiplicity correction for *known negative* correlation which can be used as an alternative to the conservative test procedure. Then we apply this idea to case of *unknown* correlation.

## 5.1 Known correlation

We have seen that $\max FWER_o^{NA}(\rho) > \alpha_{nom}$ for $\rho < 0$. To avoid any inflation in the family-wise error rate, our proposed strategy is to apply the natural correction to a nominal significance level $\alpha_a \leq \alpha_{nom}$ which is chosen so that $FWER(S) \leq \alpha_a$, for all subsets $S \subseteq I$, implies $FWER_o \leq \alpha_{nom}$. This means that at the efficacy testing stage, each test statistic relating to an elementary null hypothesis for a selected treatment is compared with the Dunnett thresholds calculated at nominal level $\alpha_a$ adjusting for the number of selected treatments.

In Fig 1 (right), values of $\alpha_a$ used in the adjustment of the natural correction are drawn for different true and equal correlations in the ZT framework when comparing 2, 3, and 4 treatments with a common control. Of course these curves behave "inversely" to the curves shown in Fig 1 (left). Since for $\rho \geq 0$, no $FWER_o$ inflation occurs and therefore no adjustment of the nominal level is necessary we have $\alpha_a = \alpha_{nom}$. In other words, in case of non-negative correlation, the correlation-adjusted two-step procedure is identical to the natural procedure, because due to Theorem 2 the latter controls the $FWER_o$. For negative values of $\rho$ however, to decrease the maximum $FWER_o$ from $FWER_o^{NA}(\rho) > \alpha_{nom}$ to $FWER_o^{KC}(\rho) \leq \alpha_{nom}$, the natural correction has to be applied based on a lower significance level $\alpha_a < \alpha_{nom}$ (KC means *known correlation*). The lowest value of $\alpha_a$ has to be applied at $\rho = -1$, where the worst case $FWER_o^{NA}$ is maximized.

## 5.2 Unknown correlation

In this section we extend our designs to accommodate an unknown correlation. If, in the method described above, the adjusted nominal level, $\alpha_a$, is calculated assuming the common correlation is $\rho_a$ when in fact $\rho$ is the true correlation, we write $FWER_o^{KC}(\rho, \rho_a)$ for the $FWER_o$. In this more general notation, we can write $FWER_o^{KC}(\rho) = FWER_o^{KC}(\rho, \rho)$ and $FWER_o^{NA}(\rho) = FWER_o^{KC}(\rho, 0)$. Our results so far indicate that $\rho_a < \rho < 0$ implies conservatism, that is, $\max FWER_o^{KC}(\rho, \rho_a) < \alpha_{nom}$, which implies that our adjustment is more severe than necessary. The opposite is true when $\rho < \rho_a < 0$, in which case $\max FWER_o^{KC}(\rho, \rho_a) > \alpha_{nom}$.

We propose basing multiplicity adjustments on an estimate $\hat{\rho}$ of the common correlation. To account for the variability of the estimate we define $FWER_o^{PI} := \mathbb{E}[FWER_o^{KC}(\rho, \hat{\rho})]$, where the expectation is taken over realizations of $\hat{\rho}$ (PI stands for '*plug-in*').

The standard way to define $\hat{\rho}$ is as follows:

1. Estimate the empirical correlation coefficients $r_i$ between efficacy and toxicity in group $i$, $\forall i \in I$.
2. Apply Fishers's z-transformation (see e.g. [22]) to each estimate $r_i$, i.e. calculate $z_i := \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$. It follows from theory that approximately $z_i \sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n_i-3}\right)$.
3. The estimator for the correlation is then defined as $\hat{\rho} = \tanh\left(\frac{1}{|I|}\sum_{i=1}^{|I|} z_i\right)$, using the equal correlation assumption.

$FWER_o^{PI}$ control is no longer guaranteed. To investigate possible inflation, the worst case $FWER_o^{PI}$ of the overall procedure was simulated for the cases that two or three treatments are compared with a common control, setting the per-group sample size as $n$ = 5, 10, 22, 50 or 100. The testing scenario is as described in Section 4.1 and $5 \cdot 10^5$ simulations were run for each value of $\rho$ to estimate the maximum $FWER_o^{PI}$.

In Fig 2 an incremental inflation of the maximum $FWER_o^{PI}$ is visible for correlations around −0.7. On the other hand, for low sample sizes the procedure is slightly conservative for $\rho$ close to 0. This is reasonable since overestimating $\rho$ when in truth $\rho = 0$ does no "harm" in terms of

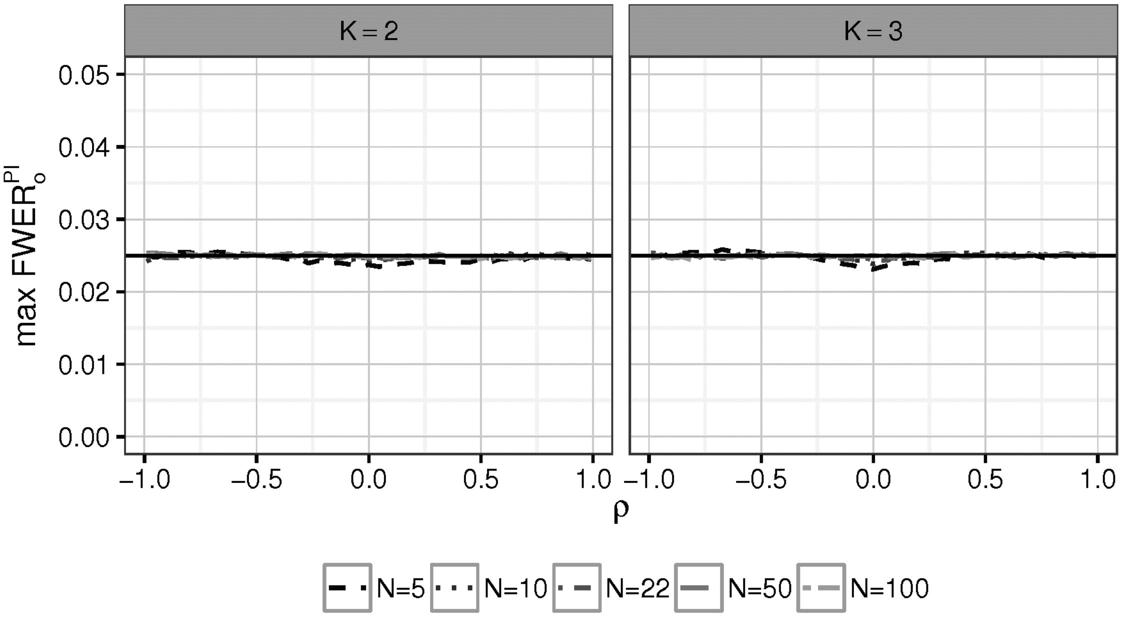

**Fig 2. Simulated values of worst case $FWER_o^{PI}$ as a function of $\rho$.** Two treatments were compared to one common control with equal per-group sample sizes *n*. A conservative behavior is visible around 0, when the per-group sample size is low.

https://doi.org/10.1371/journal.pone.0180131.g002

$FWER_o$, because $\alpha_a$ (interpreted as a function of $\hat{\rho}$) remains then constant. Only the error of underestimation takes effect by shifting the maximum $FWER_o$ slightly down. For increasing values of $n$, both effects diminish because the correlation estimate tends to be closer to the true value of $\rho$, so the impact of the conservatism due to underestimation of $\rho$ is reduced. Tables of the simulated values in Fig 2 are presented in S1 Tables.

## 6 Simulation study

We perform simulations of the comparison of 3 treatments with a single control group using the procedures discussed in the previous sections to adjust for multiple comparisons, namely the natural correction (NA); the conservative correction (CO); the adjusted method for known correlation (KC); and the adjusted method for unknown correlation with the plug-in correlation estimate (PI). The underlying statistical model as well as the formulation of the null hypotheses remains as described in Section 3. The group means of the efficacy data are said to have a linear relationship if they take the following form: $\mu_1 = \frac{1}{3}\mu_3$, $\mu_2 = \frac{2}{3}\mu_3$. On the other hand a constant relationship is defined as: $\mu_1 = \mu_2 = \mu_3$. For the control group, the mean is always set to $\mu_0 = 0$. The true toxicity means are either constant ($\theta_1 = \theta_2 = \theta_3 = 0$), linearly increasing ($\theta_1 = 0$, $\theta_2 = 0.5$, and $\theta_3 = 1$), or linearly decreasing ($\theta_1 = 1$, $\theta_2 = 0.5$, and $\theta_3 = 0$).

For **Scenario 1**, we define constant efficacy means and linearly increasing toxicity means. In **Scenario 2**, efficacy means are linearly increasing, while toxicity means are constant. In **Scenario 3**, efficacy and toxicity means both increase linearly. **Scenario 4** describes an inverse relationship between efficacy and toxicity such that efficacy means are linearly increasing while toxicity means are linearly decreasing. In all scenarios, the variances of the efficacy and toxicity outcomes are set to 1 and the correlation between toxicity and efficacy responses is constant across treatment groups.

We evaluate the power of the overall procedure when we apply the DT with a sample size of $n$ = 22 in each treatment group and on control. The measure of interest is disjunctive power, which is the probability of rejecting at least one false null hypothesis. The one-sided nominal significance level is set to $\alpha_{\text{nom}} = 0.025$.

The simulation process is split into two parts. First, the curve of values of $\alpha_a$ corresponding to the known correlation case as plotted in the right panel of Fig 2 (dotted line for 3 vs. 1) is simulated for different correlations $\rho$. Here the number of simulation runs is $5 \cdot 10^5$. Afterwards these simulated values are approximated by a cubic polynomial in the region of $\rho \in (-1, 0)$. In the second step of the simulation study, patient responses are sampled from the bivariate normal model and the fraction of simulated trials leading to the rejection of at least one false null hypothesis is counted. For the KC and PI procedures we use the simulated and smoothed $\alpha_a$ curve from the first step. The number of simulation runs in the second part is 20000.

## 6.1 Power

Fig 3 shows the results for Scenario 1. In the first column of figures, $\mu_3$ is set to 0.6, whereas in the second and third columns we have $\mu_3 = 0.8$ and $\mu_3 = 1$, respectively. Rows of figures are characterised by different values of the true correlation $\rho$, specifically −0.6, −0.3, 0, and 0.3. On the horizontal axis, the toxicity threshold (assumed to be equal across treatment groups) is drawn. On the vertical axis, we have the probability that the overall procedure rejects at least one null hypothesis.

Rejection probabilities always converge to 0 as the toxicity thresholds becomes sharper since eventually no treatment is selected and taken forwards for efficacy testing. However, rejection probabilities interpreted as a function of varying toxicity thresholds are not necessarily monotone. This can be explained by the fact that increasing the toxicity thresholds increases the expected number of treatments selected, but more treatments in the efficacy testing stage also implies sharper efficacy thresholds. Despite this, no large deviations from the rule-of-thumb that increasing the toxicity thresholds leads to increased disjunctive rejection probabilities are observed.

Regarding the curves for the two-step procedure using the natural correction and the procedure using the conservative correction, it is visible that at a certain point the former procedure starts to outperform the latter. This is the case when the safety selection rule begins to select treatments with non-negligible probability. At this point the efficacy testing boundaries for these procedures begin to differ and their power curves diverge. Increasing the toxicity thresholds further, eventually we find that all treatments will be selected with a high probability. The natural and conservative procedures then use the same boundaries in the efficacy testing step and both power curves converge towards each other again.

When we use the procedure for known correlation, following the steps set out in Section 5.1, then its power will never converge towards that of the natural procedure when $\rho < 0$. This is the case because a lower nominal significance level is assumed for the boundary calculation in the efficacy testing step (as can be seen in Fig 1, right). Meanwhile, the adjusted natural procedure outperforms the conservative procedure in terms of power in settings where all treatments are dropped with high probability. When $\rho = -0.6$, the gains in power made by the adjusted natural approach on the conservative procedure are small relative to the losses in power incurred when all treatments are regarded as safe. For increasing (negative) $\rho$, the power curve of the KC approach converges to that of the natural procedure, making the region of superiority over procedure CO broader. For non-negative correlation, KC is identical to NA as explained in Section 5, and therefore these procedures have the same power. Regarding the power of procedure PI, its power curve is very close to that of procedure KC.

For scenarios 2, 3 and 4 we set $\mu_3 = 0.8$ and $\rho \in \{-0.3, 0, 0.3\}$. In Scenario 2 of Fig 4 (left column), toxicity means are equal but efficacy means are linearly increasing. As the toxicity

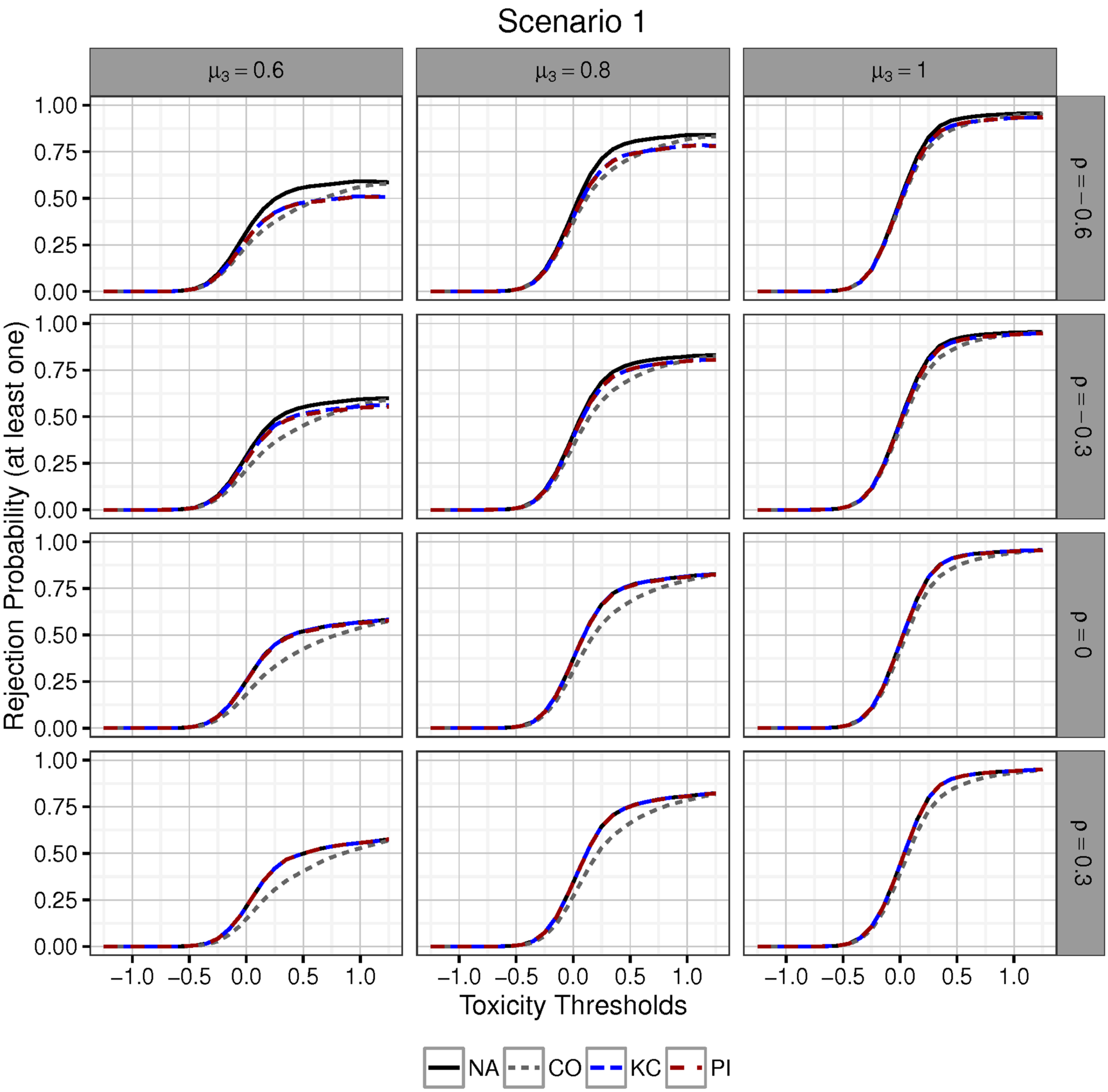

**Fig 3. Scenario 1; Power of different two-step procedures as a function of the toxicity thresholds.** High values of the toxicity threshold (defined to be the same in each group) imply more treatments are selected for efficacy testing. Rows correspond to different values of $\rho \in \{-0.6, -0.3, 0, 0.3\}$; columns correspond to values of $\alpha_3 \in \{0.6, 0.8, 1\}$ in the linear scenario (see also the main text). As $\rho$ increases, the region where the adjusted approaches dominate the conservative procedure in terms of power, increases in area. This behaviour accelerates for higher values of $\alpha_3$.

https://doi.org/10.1371/journal.pone.0180131.g003

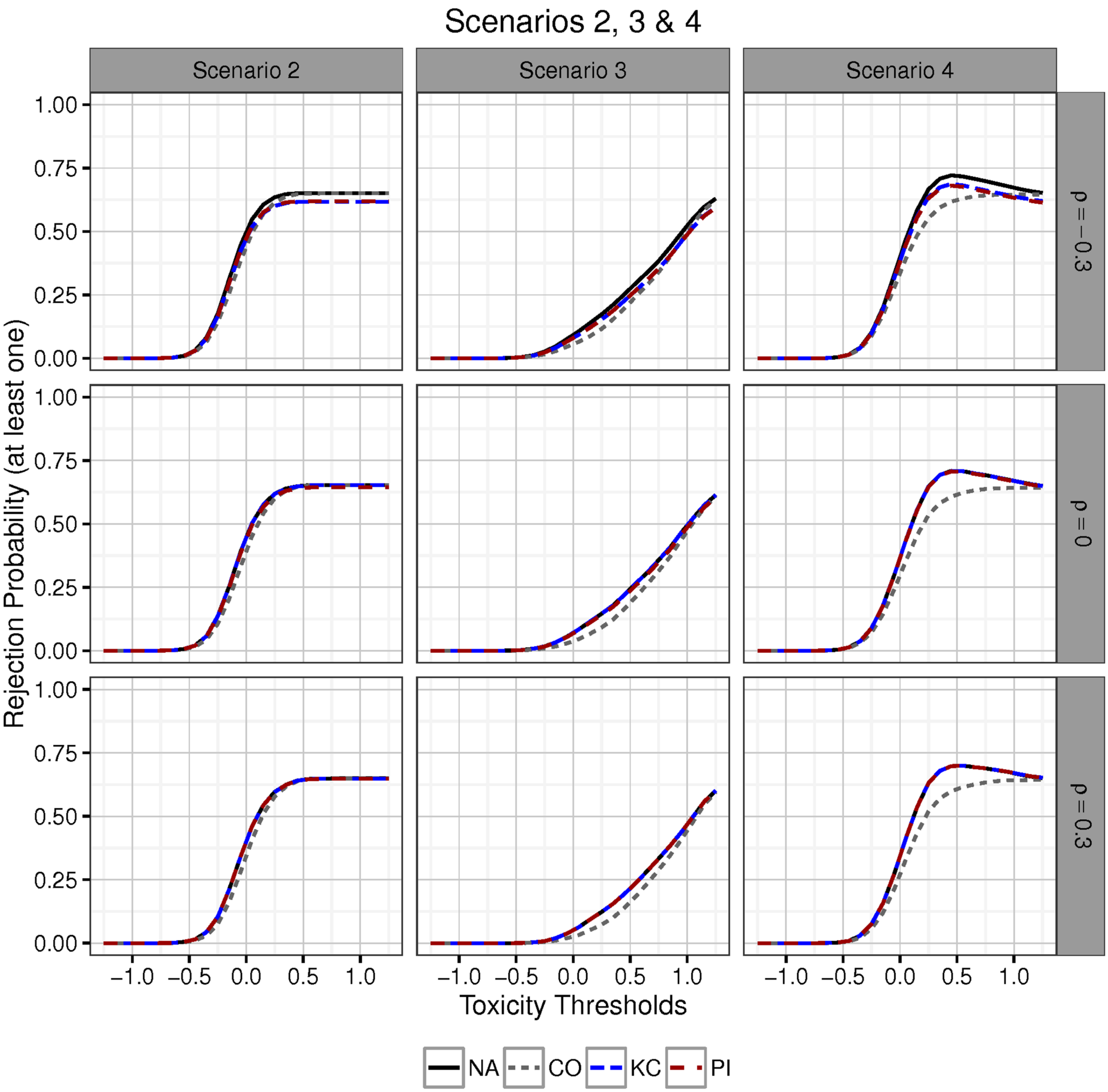

**Fig 4. Power in Scenario 2 (left column), Scenario 3 (middle column) and Scenario 4 (right column).** Different efficacy/toxicity patterns lead to varying gains in power over procedure CO. Rows of figures are generated setting $\rho = -0.3$, 0 or 0.3. In all scenarios, $\mu_3 = 0.8$. The smallest gains in power are made in Scenario 2, which represents an all-or-none selection situation. Deviating from this setting leads to gains in power in the area where dropping treatments is expected (Scenarios 3 and 4). The largest power gains are seen in Scenario 4 where the slopes of the linear configurations of the toxicity and efficacy means are of different signs.

https://doi.org/10.1371/journal.pone.0180131.g004

threshold increases, the probability that all three treatments are selected rapidly increases from 0 to 1. This is the case because toxicity means are all equal in this scenario and likely either all the treatments satisfy the selection criterion or none do. Differences in toxicity sample means are then due to random variation alone. Therefore in this scenario an all-or-none treatment selection situation is likely. As we have noted for Scenario 1, the other testing approaches dominate procedure CO in situations where dropping *some* (but not all) treatments is likely. Therefore the size of the region where procedures KC and PI dominate CO is small and the margin of dominance is negligible.

In Scenario 3 (Fig 4, middle column), toxicity means are linearly increasing. As the toxicity threshold is gradually relaxed, one-by-one the three treatments can be deemed safe in the sense that they will satisfy the safety selection criterion with high probability. Therefore, since the expected number of selected treatments increases incrementally with the safety threshold, the slope of the power curve is less steep compared to Scenario 2. On the other hand gains in power of NA, KC, and PI over procedure CO are larger than those seen in Scenario 2.

In Scenario 4, (Fig 4, right column), efficacy means are increasing across treatment groups 1, 2 and 3, while toxicity means are decreasing. This mimics the situation where toxicity might negatively impact on efficacy. In this scenario treatments with higher efficacy are more likely to be selected. Similar to scenario 3, since toxicity means for treatments are not all equal, we see a more gradual increase in power compared with Scenario. For intermediate values of the safety threshold, the treatments most likely to be selected are those with higher efficacy. Therefore, relaxing the boundaries of the DT results in a greater increase of power compared with other testing scenarios (where safer treatments are also less efficacious) and differences between the power curves of procedure CO and the other procedures are more marked.

In all scenarios, as the toxicity threshold is relaxed the expected number of null hypotheses that are tested increases. What can be gained by applying the proposed two-step procedures in comparison to the conservative procedure in terms of disjunctive power, depends on the exact configuration of toxicity and efficacy means. It can be seen from Figs 3 and 4 that changes in the safety selection rule have a similar impact on the operating characteristics of all MTPs, although clearly some procedures are more powerfull than others due to differences in the exact choice of efficacy boundaries applied to the selected null hypotheses. For example the possibilities for power gains compared to the conservative procedure increase, if treatments are selected, that also have the largest treatment effect. This situation is reflected in Scenario 4.

## 6.2 Impact of varying the sample size

Fig 5 plots the power of the two-step procedures under varying toxicity thresholds. Rows consider Scenario 1, Scenario 3 and Scenario 4, and different values of the correlation coefficient ($\rho \in \{-0.3, 0.3\}$). In the left column, the per-group sample size is $n = 50$ and $\mu_3$ is set equal to 0.52. In the middle ($n = 100$) and right ($n = 500$) column, we set $\mu_3 = 0.368$ and $\mu_3 = 0.164$, respectively. Values of $\mu_3$ are chosen such that a single treatment-control comparison has power 0.9 of detecting the effect $\mu_3$ for the given group size. As $n$ increases, the power curves of the two-step procedures essentially converge towards step functions. The reason for this is that with increasing sample size, the true toxicity means can be estimated with greater accuracy. As the sampling error of the group sample mean toxicities decreases, we can clearly distinguish between 'safe' and 'toxic' treatments; thus, varying the toxicity threshold on an interval between two true toxicity means has little impact on the treatment selection procedure in terms of the expected number selected and the identity of the treatments selected.

From Figs 3–5 it can be seen, that the prespecified sample sizes $n \in \{22, 50, 100, 500\}$ per group are large enough to achieve a high correspondence between the KC and the PI

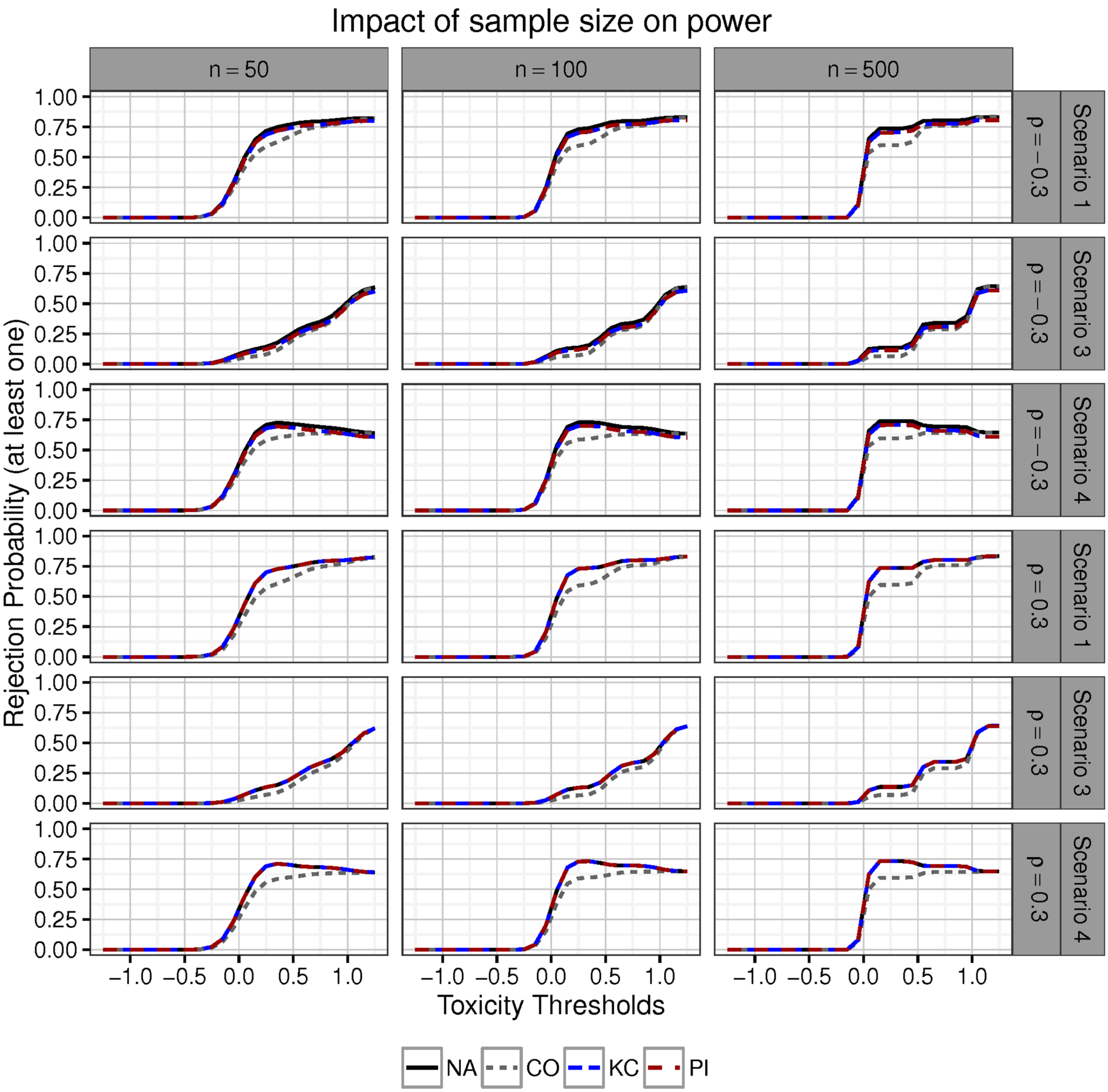

**Fig 5. Power of the four two-step procedures under varying toxicity thresholds and (per-group) sample sizes in Scenarios 1, 3, and 4.** For each group size, the effect size $\mu_3$ is chosen as the difference a single treatment comparison can detect with power 0.9. The correlation is set as $\rho \in \{-0.3, 0.3\}$.

https://doi.org/10.1371/journal.pone.0180131.g005

procedure. One intuitive explanation for this reason is that over- und underestimation of $\rho$ which occurs by using the empirical correlation coefficient, roughly balance out, as explained in Section 5.2. On the other hand the conservativeness of the PI procedure for small sample sizes around $\rho = 0$ as seen in Fig 2, suggests a small power loss compared to PI in this setting.

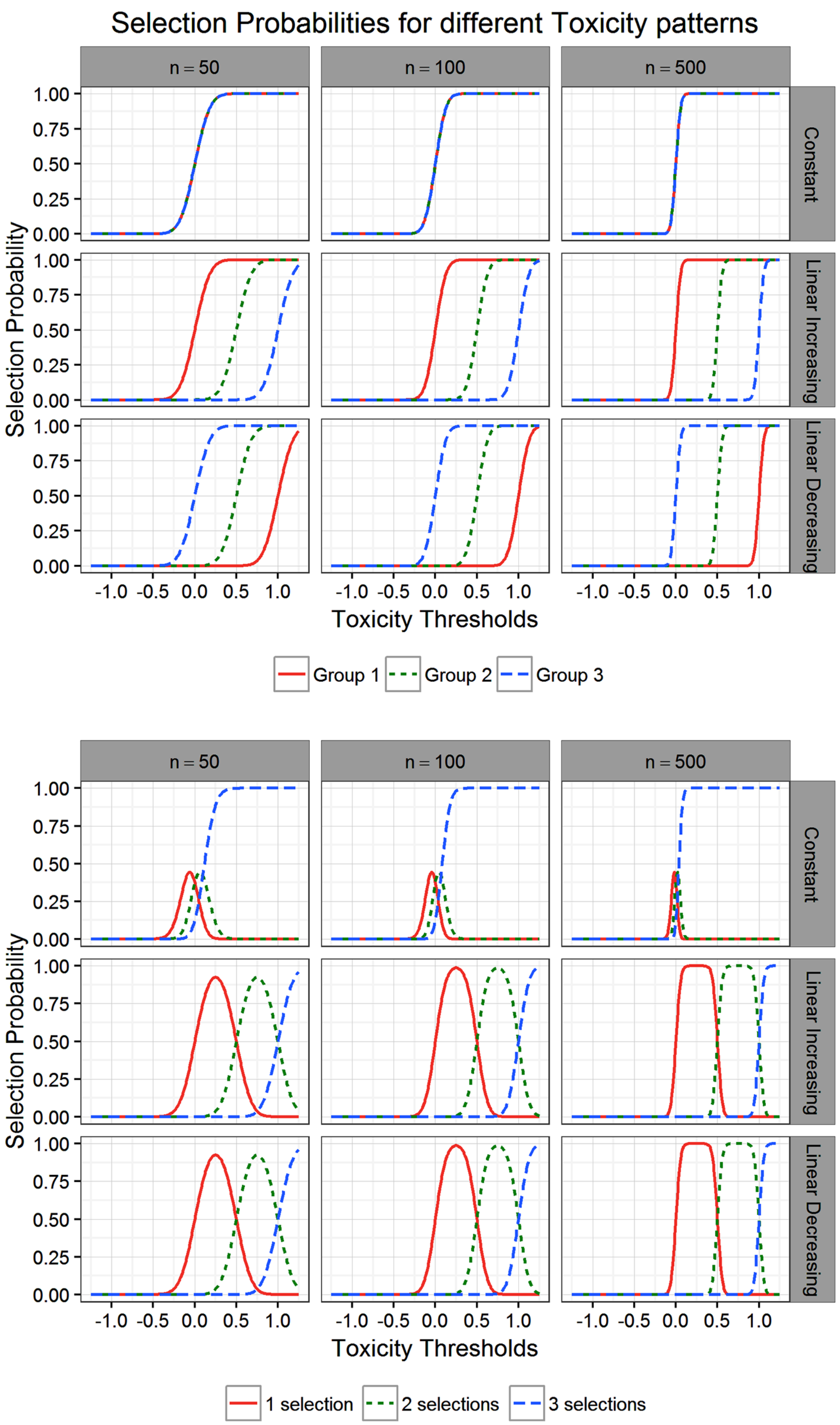

**Fig 6. Selection probabilities for different toxicity patterns.** In the upper plot the probabilities of selecting treatments 1, 2 or 3 are drawn. The probabilities of selecting exactly 1, 2, or 3 treatments are drawn in the lower plot.

https://doi.org/10.1371/journal.pone.0180131.g006

In S1 Fig a power curve is plotted for $\rho = 0$ and $n = 5$ in the situation of Scenario 1 with $\mu_3$ again chosen such, that the power for a single treatment-control comparison is 0.9. The graph shows a small power loss of the PI procedure due to estimation error.

### 6.3 Impact of the selection probability

The behaviour of the power curves seen in Fig 5 can be explained by taking a closer look at the safety selection procedure. The top plot of Fig 6 shows the probabilities of selecting different treatment groups. Rows correspond to constant (Scenario 2), linear increasing (Scenario 1 and 3), and linear decreasing (Scenario 4) toxicity means. The columns are characterised by different sample sizes. For increasing sample sizes, estimation of the group toxicity means becomes more precise, which leads to a decreased variability in the selection process and therefore to steeper power curves. At the boundaries 0, 0.5, and 1, the probabilities of being selected are sharply increasing for the corresponding group. In the lower plot of Fig 6, the probabilities of selecting at exactly 1, 2, or 3 treatments are drawn. For larger sample sizes, the area of overlap between these curves diminishes. This is the reason for the plateaus visible in Fig 5.

## 7 Discussion

We have considered two-step procedures, where safety selection precedes efficacy testing of multiple hypotheses using naive Dunnett boundaries and adjusting just for the number of selected treatments. For such a procedure we demonstrated that the *FWER* of the overall procedure may exceed the nominal level $\alpha$. The size of this deviation depends on the unknown distribution of the toxicity measurements and pre-defined toxicity thresholds. In addition, the dependency structure of the efficacy and toxicity outcomes determines the size of the maximum possible $FWER_o$. We assumed a bivariate normal model for efficacy and toxicity measurements, in which case the dependency structure is captured by correlations. Here toxicity was represented by a normally distributed variable. In practice, the adverse effects of a drug on the kidneys, liver or heart are often monitored by continuous parameters such as laboratory values or variables from an ECG, respectively.

A major finding of the present work is that selection of treatments based on observed safety data is indeed possible and multiple comparison procedures such as Dunnett or Bonferroni tests which adjust only for the number of selected treatments will control the $FWER_o$, so long as toxicity and efficacy measurements are (positively) associated in the sense of [21]. Intuitively this can be explained by the fact that in case of a positive association, treatments that show high efficacy are more likely to be dropped. Or, conversely, if a treatment is selected on the basis of lower toxicity, under the null hypothesis it will be more likely that lower efficacy is observed too. In the bivariate normal model, this is the case when the correlation between toxicity and efficacy outcomes are non-negative in all treatment groups. Clearly, if all correlations are zero, selection is independent of efficacy testing. Then a Dunnett adjustment for the remaining treatments controls the FWER and depending on the safety selection rule, it may become conservative.

In the case that the correlation between efficacy and toxicity is negative for at least one treatment group, dropping such a treatment group would imply dropping a treatment which likely has poor efficacy measurements relative to those on the other treatments. If then for the remaining treatments the efficacy testing thresholds are relaxed, the overall *FWER* will be inflated. In our particular setting, we have shown that this inflation will be maximized when all correlations between efficacy and toxicity approach the value −1. We proposed an adjusted approach, in which we change the nominal $\alpha$-level to be used in the Dunnett adjustment for efficacy testing depending on the correlation. This means that if one assumes no or positive correlations, then the resulting two-step procedure is exactly as described above. It is sufficient to take Dunnett boundaries adjusting for the number of selected treatments at nominal level $\alpha$. Assuming negative correlations will result in Dunnett tests adjusting for the number of selected treatments at a lower nominal $\alpha$-level. When the correlation is unknown, which will often be the case in clinical research, a reasonable lower boundary for the correlation may be assumed. For example in oncology it is occasionally assumed that the correlation between efficacy and toxicity is non-negative. On a trial level, this is shown by [23]. A reasonable lower boundary for the correlation is then 0. According to our results it is then possible to drop treatments due to toxicity and calculate the Dunnett boundaries at nominal level $\alpha_{nom}$ adjusting for the number of selected treatments only without inflating the *FWER*.

Alternatively it is suggested to estimate the correlation and use this estimate to adjust the nominal $\alpha$-level. This is especially useful, if correlations are assumed to be equal, which leads to a pooling of the different groups in terms of correlation estimation. Simulations have shown, that such a strategy is adequate even if results on exact overall *FWER* control or inflation are missing. Compared to the conservative approach (always adjusting for all a-priori defined comparisons) the proposed two-step testing strategies result in higher power for non-negative correlations. For negative correlations there is a gain in power if it is very likely that treatments have to be dropped due to toxicity. However, if all treatments are sufficiently safe (and likelihood of dropping is negligible), the proposed two step-testing procedures will have a loss of power as a lower nominal level alpha to be used in the Dunnett test. This is the price to be paid for adjusting just for the number of selected treatments. However, negative correlations may be implausible in many practical situations.

In the planning phase of a clinical trial, the first consideration is regarding the correlation between toxicity and efficacy. In accordance to the arguments presented before, if the assumption on positive or zero correlation is justified, then the use of the natural procedure is reasonable and simulation studies for different scenarios as performed in Section 6 support decisions regarding the planned per-group-sample sizes. Justification for non-negative correlation could be derived from a scientific understanding of the relationship between two endpoints, or on data from previous trials which measured the two endpoints. When an assumption on positive correlation is not plausible, considerations on reasonable values of the correlation may enable the use of the correlation-adjusted procedure. Another possibility is to estimate the correlation. In each of these cases, a comprehensive simulation study would typically be performed ahead of time to quantify the operating characteristics of the proposed testing procedure in scenarios consistent with any assumptions about the correlation coefficient, and to establish the robustness of properties to deviations from these assumptions. As a sidenote it should be mentioned, that at the end of the study prefixed and biologically motivated toxicity thresholds are used to perform the statistical analysis. Simulations for varying toxicity thresholds as performed in Section 6 are essential to determine the sample sizes required and to evaluate the impact of the proposed selection rule (and the underlying toxicity thresholds) on the operating characteristics for different efficacy and toxicity means.

In summary the proposed two-step procedures are a way to integrate both safety and efficacy aspects in a more formal way for decision making on the benefit and risk ratio of new treatments. All safety and efficacy evaluations are performed at the end of the study, when for all patients all safety and efficacy data are available. Treatments are first screened on the basis of safety, and only those deemed to be sufficiently safe are considered for the efficacy testing step. The arguments provided in this paper can be extended allowing to perform the safety selection step already earlier and more frequently in the trial, e.g., for clinical trials where a data and safety monitoring board will perform periodic safety monitoring when the trial is still on-going and more patients still have to be included. For example in more "traditional" adaptive seamless designs [24–26] treatment selection is suggested to be conducted at an adaptive interim analysis, where only a first cohort of patients have been included in the trial. If treatments are selected, a further cohort with new patients have to included and randomized to the selected treatment arm. This is in contrast to the proposed two-step procedure in this paper, where safety selection is conducted at the end. This prohibits a direct head-to-head comparison as selection rules would have to differ substantially using different amount of data. Therefore part of our future work is to extend our two-step procedures to allow interim safety selection as well as a comparison to adaptive seamless designs [24, 25] using the same amount of data for selection. Further investigations may focus on other types of safety data such as, binary or ordinally scaled, where results of positive association may be applicable as well, for example if the probability of having an adverse event increases with efficacy. A key point of the two-step procedure is that the toxicity boundary have to be fixed in advance and no efficacy data are involved when deciding which treatments shall be dropped.

# Appendix

## Proof of Theorem 1

We will prove Theorem 1 in Section 4.1. In this proof we will apply the Slepian theorem (see, for example, the Appendix of [27]) which states that for every $k$-dimensional random vector $\mathbf{z}$ following a multivariate normal distribution with zero mean and unit variance in each component and correlation matrix $\mathbf{R}$, the probability $P(z_1 \leq c_1, \ldots, z_k \leq c_k)$ is a strictly increasing function of every off-diagonal entry of $\mathbf{R}$.

*Proof.* According to Eq (1), the $FWER_o$ is a sum with addends of the following form:

$$P\left(\bigcap_{i\in S}\{\bar{\mathbf{y}}_i \leq b_i^y\} \cap \bigcap_{j\in I\setminus S}\{\bar{\mathbf{y}}_j > b_j^y\}\right) - C \cdot P\left(\bigcap_{j\in I\setminus S}\{\bar{\mathbf{y}}_j > b_j^y\}\right) \tag{2}$$

where $C := P(\bigcap_{i\in S\cap J}\{\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0 < b_{i,S}^x\} \cap \bigcap_{i\in S}\{\bar{\mathbf{y}}_i \leq b_i^y\})$ (here we made use of the independence between measurements in different groups). Following Slepian's result, -C is monotonically decreasing in $\rho$ and the statement is proofed for the ZT framework. For the DT, all probabilities in Eq (2) are conditional to an independent variance estimate (and also the thresholds $b_{i,S}^x$ now depend on these estimates). On all these probabilities the monotonicity statement remains to be true. Averaging over all possible values then shows, that our result is true in the DT framework as well.

## Proof of Theorem 2

In what follows we will provide and prove a sufficient condition for $FWER_o$ control of a procedure that controls the FWER in the efficacy testing step. To do this we first need the notion of association as defined in [21]:

*By definition the random variables* $(Q_1, \ldots, Q_N) =: \mathbf{Q}$ *are associated if and only if*

$$\mathbb{E}[f(\mathbf{Q}) \cdot g(\mathbf{Q})] \geq \mathbb{E}[f(\mathbf{Q})] \cdot \mathbb{E}[g(\mathbf{Q})] \tag{3}$$

*for every componentwise non-decreasing functions f and g.*

*Remark* 1. From the definition it follows immediately, that non-decreasing transformations of associated random variables are also associated. Another intuitively clear property of association is, that the set of a single variable is associated. Furthermore, if two sets of associated random variables are independent, then the union is also associated. All of these properties are proven in [21].

In preparation of proving Theorem 2, two lemmata are presented:

**Lemma 1**. *Let $\emptyset \neq J \subseteq I$ be the index set of inefficacious treatments. The maximum $FWER_o$ is controlled at level $\alpha_{nom}$ for a given multiple comparison procedure, if FWER(S) $\leq \alpha_{nom}$ (that means $\mathbb{E}[\varphi_J^S] \leq \alpha_{nom}$) and $\mathbb{E}[\varsigma_S \cdot \varphi_J^S] \leq \mathbb{E}[\varsigma_S] \cdot \mathbb{E}[\varphi_J^S]$ both hold for $\forall S \in \mathcal{M}_J := \{S \subseteq I : S \cap J \neq \emptyset\}$.*

*Proof*. We first note that only one of the binary indicators $\varsigma_S$ with $S \in \mathcal{M}_J$ can have the value 1 at once. Here $\mathcal{M}_J$ is the set of all possible selections, that contain at least one true null-hypothesis. The sets $S \notin \mathcal{M}_J$ are not relevant for type I error control. We have seen earlier in Eq (1) that the *FWER* of the overall procedure can be written as $FWER_o = \sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S \cdot \varphi_J^S]$. Because $\mathbb{E}[\varsigma_S \cdot \varphi_J^S] \leq \mathbb{E}[\varsigma_S] \cdot \mathbb{E}[\varphi_J^S]$ holds for each addend it holds $\sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S \cdot \varphi_J^S] \leq \sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S] \cdot \mathbb{E}[\varphi_J^S]$. The factor $\mathbb{E}[\varphi_J^S]$ is bounded from above by $\alpha_{\text{nom}}$. Therefore it holds

$$\begin{aligned} \sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S \cdot \varphi_J^S] &\leq \alpha_{\text{nom}} \cdot \left( \textstyle\sum_{S \in \mathcal{M}_J} \mathbb{E}[\varsigma_S] \right) \\ &\leq \alpha_{\text{nom}}. \end{aligned}$$

The last inequality follows from the fact that $\sum_{S \subseteq \mathcal{M}_J} \mathbb{E}[\varsigma_S]$ is just a sum of probabilities of disjoint events and therefore less or equal 1 (in general equality only holds, if $J = I$).

**Lemma 2**. *Let $\emptyset \neq J \subseteq I$ be the index set of inefficacious treatments. For all subsets $S \subseteq I$ let the families of indicators $(\tau_i)_{i \in I}$ and $(\eta_i^S)_{i \in I}$ have the following properties:*

1. *The vector $(\tau_i)_{i \in I \setminus S}$ is independent to the vector $((\tau_i)_{i \in S}, (\eta_i^S)_{i \in S})$.*
2. *The set $\{\tau_i : i \in I\} \cup \{\eta_i^S : i \in I\}$ is associated.*

*Then $\forall S \in \mathcal{M}_J$ it holds*

$$\mathbb{E}[\varsigma_S \cdot \varphi_J^S] \leq \mathbb{E}[\varsigma_S] \cdot \mathbb{E}[\varphi_J^S]. \tag{4}$$

*Proof*. Let $S \in \mathcal{M}_J$ be fixed. It holds $\mathbb{E}[\varsigma_S \cdot \varphi_J^S] = \mathbb{E}[\prod_{j \in I \setminus S} \tau_j] \cdot \mathbb{E}[(\prod_{i \in S}(1 - \tau_i)) \cdot \varphi_J^S]$ due to the definition of $\varsigma_S$ and property 1. We focus on the right-hand factor $\mathbb{E}[(\prod_{i \in S}(1 - \tau_i)) \cdot \varphi_J^S]$ and note that $f := (-1) \cdot \Pi_{i \in S}(1 - \tau_i)$ and $g := \varphi_J^S$ both interpreted as functions of the $\tau_i$ and $\eta_i^S$ are monotonically non-decreasing. Due to the association assumption Eq (3) it therefore holds $\mathbb{E}[(\prod_{i \in S}(1 - \tau_i)) \cdot \varphi_J^S] \leq \mathbb{E}[\prod_{i \in S}(1 - \tau_i)] \cdot \mathbb{E}[\varphi_J^S]$ (note that the minus sign in $f$ reverses the inequality sign). By again using the independence property 1, Eq (2) is proven.

We now have the tools to prove **Theorem 2**:

*Proof*. At first, we consider the ZT framework. We are going to prove that under the stated assumptions Lemma 2 holds. Then Lemma 1 can be applied. We set $\tau_i := \mathbf{1}_{(b_i^y, \infty)}(\bar{\mathbf{y}}_i)$ and $\eta_i^S := \mathbf{1}_{(b_{i,S}^x, \infty)}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0), \forall i \in I$ and for fixed $S \subseteq I$. The boundaries $b_{i,S}^x$ are chosen such, that the FWER is controlled at level $\alpha_{\text{nom}}$ in the efficacy testing step for every $S \subseteq I$. We know that by construction the vectors $(\bar{\mathbf{y}}_i : i \in I \setminus S)$ and $(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0, \bar{\mathbf{y}}_i : i \in S)$ are independent. This also holds for the corresponding vectors of the indicators $\tau_i$ and $\eta_i^S$, which verifies property 1 in Lemma 2.

It is easy to see that $\text{Cov}(\bar{\mathbf{y}}_i, \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0) = \frac{\rho_i \sigma \omega}{n_i} \geq 0$ and for $i \neq j$ it holds that $\text{Cov}\left(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_j - \bar{\mathbf{x}}_0\right) = \frac{\sigma^2}{n_0} \geq 0$, and therefore all pairwise correlations are non-negative. From [28] we know that positively correlated multivariate normal random variables are associated. This proves the association of the set of random variables $\{\bar{\mathbf{y}}_i : i \in I\} \cup \{\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0 : i \in I\}$. According to Remark 1, the same is then true for above defined indicators, which are monotonically non-decreasing functions. This concludes the proof of Lemma 2. After applying Lemma 1, this proofs $FWER_o$ control at level $\alpha_{\mathrm{nom}}$ in the *ZT* scenario, if all the correlations $\rho_i$ are non-negative.

For the proof of the DT framework by $V_i$ we denote the unbiased variance estimator in group *i*, that is $V_i := \frac{1}{n_i-1}\sum_{i=1}^{n_i}(x_{ij} - \bar{\mathbf{x}}_i)^2$. The test statistics of the DT use the overall variance estimator, which is function of the group variance estimates $V^S := \sum_{I\cup\{0\}} \frac{n_i-1}{\sum_{I\cup\{0\}}(n_i-1)} V_i$. The multivariate extension of a fundamental result in statistics states, that the sample covariance matrix (consisting of $V_i$) defined for the data matrix of both efficacy and toxicity is independent of the corresponding sample mean vector (see for example [22] p. 48). The same is then true for overall variance $W^S := -V^S$, which is just a function of the sample covariance matrix, and thererfore $(\bar{\mathbf{y}}_i)_{i\in I\setminus S}$ and $((\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0)_{i\in S}, (\bar{\mathbf{y}}_i)_{i\in S}, W^S)$ are independent. $W^S$ just influences the thresholds $b^x_{i,S}$ in $\eta^S_i$ and therefore by using the same arguments as in the ZT case, the fulfilment of property 1 in Lemma 2 is shown.

For property 2 we have to note that $W^S$ itself forms an associated set and due of independence the same is then true for the set $\{\bar{\mathbf{y}}_i : i \in I\} \cup \{\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_0 : i \in I\} \cup \{W^S\}$. For increasing $W^S$ (which means decreasing $V^S$), $b^x_{i,S}$ is decreasing and therefore $\eta^S_i$ is non-decreasing. For the remaining the argument from the ZT case apply. This proves property 2 in Lemma 2.

*Remark* 2. It can be seen from the proof, that in the DT setting the assumption of equal group variances as stated in Section 3.2 is not used, except that it implies the use of the group variance estimator $V^S$. This represents the classical situation of the Dunnett test. For the straightforward generalization of different variances, Theorem 2 remains true, if the variances are estimated by the $V_i$ (or functions of these). For the ZT situation, $\sigma$ can be substituted by $\sigma_i$ in the proof.

## Extensions

The bivariate normal model also serves as a basis for an extension of the ZT approach for binary toxicity indicators, e.g. adverse events. This can be done by defining indicators for patient *j* in group *i*: $z_{ij} := \mathbf{1}_{\{(b^z_{ij},\infty)\}}(y_{ij})$ with patient specific thresholds $b^z_{ij}$, where the normally distributed $y_{ij}$ are now interpreted as latent variables. Safety selection would then be based on the proportion of adverse events. Because the application of the indicator function on the latent variables $y_{ij}$ preserves association and independence, Theorem 2 can be easily extended for the present case.

Extending the DT setting to binary toxicity data is more complicated since the variance estimator and the proportion of adverse events are not independent any more. In this case the fulfilment of property 2 of Lemma 2 cannot be shown by considering the union of the independent associated sets.

Another possible extension for safety selection to consider the toxicity measurements of a treatment group relative to the control measurements, in contrast to examine safety based on the groups toxicity data only. If the control efficacy data is correlated with the group efficacy data, the arguments in Theorem 2 based on independence again cannot be applied.

For the latter two cases, for the possibility of $FWER_o$ control formal proofs cannot be derived. These scenarios remain an open topic for future research.

## Supporting information

**S1 Tables. Simulated maximum FWERs, that are presented in Fig 2.**
(PDF)

**S1 Fig. Power for $\rho = 0$ and $n = 5$.**
(PDF)

**S1 Example. Hypothetical example of the two-step procedure.**
(PDF)

## Acknowledgments

The research leading to these results has received funding from the EU Seventh Framework Programme [FP7 2007-2013] under grant agreement no. 602552. We are grateful to the editor and three referees for their helpful comments, which improved the presentation of the article.

## Author Contributions

**Conceptualization:** GH LVH FK.

**Formal analysis:** GH FK.

**Funding acquisition:** FK.

**Investigation:** GH FK.

**Methodology:** GH FK.

**Software:** GH.

**Visualization:** GH.

**Writing – original draft:** GH LVH FK.

**Writing – review & editing:** GH LVH FK.

## References

1. Parmar MK, Carpenter J, Sydes MR. More multiarm randomised trials of superiority are needed. The Lancet. 2014; 384(9940):283–284. https://doi.org/10.1016/S0140-6736(14)61122-3
2. Bogaerts J, Sydes MR, Keat N, McConnell A, Benson A, Ho A, et al. Clinical trial designs for rare diseases: studies developed and discussed by the International Rare Cancers Initiative. European journal of cancer. 2015; 51(3):271–281. https://doi.org/10.1016/j.ejca.2014.10.027 PMID: 25542058
3. Billingham L, Malottki K, Steven N. Research methods to change clinical practice for patients with rare cancers. The Lancet Oncology. 2016; 17(2):e70–e80. https://doi.org/10.1016/S1470-2045(15)00396-4 PMID: 26868356
4. ICH. Topic E 9: Statistical Principles for Clinical Trials. http://wwwemaeuropaeu. 1998;
5. Wason J, Stecher L, Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done. Trials. 2014; 15(1):364. https://doi.org/10.1186/1745-6215-15-364 PMID: 25230772
6. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association. 1955; 50(272):1096–1121. https://doi.org/10.1080/01621459.1955.10501294
7. Herberich E, Sikorski J, Hothorn T. A robust procedure for comparing multiple means under heteroscedasticity in unbalanced designs. PloS one. 2010; 5(3):e9788. https://doi.org/10.1371/journal.pone.0009788 PMID: 20360960
8. Bretz F, Hothorn T, Westfall P. Multiple comparisons using R. CRC Press; 2010.
9. Dmitrienko A, D'Agostino R. Traditional multiplicity adjustment methods in clinical trials. Statistics in Medicine. 2013; 32(29):5172–5218. https://doi.org/10.1002/sim.5990 PMID: 24114861
10. Alosh M, Bretz F, Huque M. Advanced multiplicity adjustment methods in clinical trials. Statistics in Medicine. 2014; 33(4):693–713. https://doi.org/10.1002/sim.5974 PMID: 24105821

11. Jennison C, Turnbull BW. Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. Biometrics. 1993; p. 741–752. https://doi.org/10.2307/2532195 PMID: 8241370

12. König F, Bauer P, Brannath W. An adaptive hierarchical test procedure for selecting safe and efficient treatments. Biometrical journal. 2006; 48(4):663–678. https://doi.org/10.1002/bimj.200510235 PMID: 16972719

13. Jaki T, Hampson LV. Designing multi-arm multi-stage clinical trials using a risk–benefit criterion for treatment selection. Statistics in medicine. 2016; 35(4):522–533. https://doi.org/10.1002/sim.6760 PMID: 26456537

14. Juurlink DN, Mamdani MM, Lee DS, Kopp A, Austin PC, Laupacis A, et al. Rates of Hyperkalemia after Publication of the Randomized Aldactone Evaluation Study. New England Journal of Medicine. 2004; 351(6):543–551. https://doi.org/10.1056/NEJMoa040135 PMID: 15295047

15. European Medicines Agency. Valdoxan: European Public Assessment Report—ANNEX I SUMMARY OF PRODUCT CHARACTERISTICS; 2016. EMEA/H/C/000915 -IAIN/0034.

16. Shah RR, Morganroth J. Evaluating the QT-Liability of a Drug during its Development. Pharmaceutical Medicine. 2008; 22(3):151–164. https://doi.org/10.1007/BF03256697

17. Alvarez PA, Pahissa J. QT alterations in psychopharmacology: proven candidates and suspects. Current drug safety. 2010; 5(1):97–104. https://doi.org/10.2174/157488610789869265 PMID: 20210726

18. Baron G, Perrodeau E, Boutron I, Ravaud P. Reporting of analyses from randomized controlled trials with multiple arms: a systematic review. BMC medicine. 2013; 11(1):84. https://doi.org/10.1186/1741-7015-11-84 PMID: 23531230

19. Weinberger MH, Roniker B, Krause SL, Weiss RJ. Eplerenone, a selective aldosterone blocker, in mild-to-moderate hypertension. American Journal of Hypertension. 2002; 15(8):709—716. https://doi.org/10.1016/S0895-7061(02)02957-6 PMID: 12160194

20. Weisler RH, Biederman J, Spencer TJ, Wilens TE, Faraone SV, Chrisman AK, et al. Mixed Amphetamine Salts Extended-Release in the Treatment of Adult ADHD: A Randomized, Controlled Trial. CNS Spectrums. 2006; 11(8):625–639. https://doi.org/10.1017/S1092852900013687 PMID: 16871129

21. Esary JD, Proschan F, Walkup DW, et al. Association of random variables, with applications. The Annals of Mathematical Statistics. 1967; 38(5):1466–1474. https://doi.org/10.1214/aoms/1177698701

22. Tong YL. The Multivariate Normal Distribution. Springer-Verlag New York; 1990.

23. Abola M, Prasad V, Jena A. Association between treatment toxicity and outcomes in oncology clinical trials. Annals of oncology. 2014; 25(11):2284–2289. https://doi.org/10.1093/annonc/mdu444 PMID: 25193993

24. Bauer P, Bretz F, Dragalin V, Koenig F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. Statistics in Medicine. 2016; 35(3):325–347. https://doi.org/10.1002/sim.6472 PMID: 25778935

25. Bretz F, Koenig F, Brannath W, Glimm E, Posch M. Adaptive designs for confirmatory clinical trials. Statistics in Medicine. 2009; 28(8):1181–1217. https://doi.org/10.1002/sim.3538 PMID: 19206095

26. Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in Medicine. 2008; 27(10):1612–1625. https://doi.org/10.1002/sim.3048 PMID: 17876763

27. Hochberg Y, Tamhane AC. Multiple Comparison Procedures. New York, NY, USA: John Wiley & Sons, Inc.; 1987.

28. Pitt LD. Positively correlated normal variables are associated. The Annals of Probability. 1982; p. 496–499. https://doi.org/10.1214/aop/1176993872