universität wien

MAGISTERARBEIT / MASTER'S THESIS

Titel der Magisterarbeit / Title of the Master's Thesis

„Optimal exact tests for multiple binary endpoints“

verfasst von / submitted by

Dipl.-Ing. Dr. Robin Ristl, Bakk.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag. rer. soc. oec.)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 951
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Magisterstudium Statistik UG2002
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Walter Gutjahr

Contents

Acknowledgements

The work presented in this thesis is part of my contribution to the research project ASTERIX (Advances in Small Trials dEsign for Regulatory Innovation and eXcellence), which is funded by the 7th Framework Program of the European Union (grant agreement number 603160). The development of this work was supported by Ekkehard Glimm and Dong Xi, both working at Novartis, and it was constantly supervised by Martin Posch, the head of the Section of Medical Statistics of the Medical University of Vienna, at which I am working. A joint publication including the main results of this work is in preparation. The key algorithms presented here were developed based on the contents of the lecture on Decision Support held by the supervisor of this thesis at the University of Vienna, Walter Gutjahr.

I further want to express my gratitude for the permanent support from my family and first of all my partner Christina in all my efforts of pursuing an academic career and adding to our research field.

Summary

There is increasing interest in studying cures for rare diseases and investigating the effect of medical treatments in specific small populations. In this setting, the number of patients that can be recruited to an individual trial is inherently small. The major challenges in the statistical analysis of data from such small trials are the overall low amount of information and the potential lack of accuracy of asymptotic approximations that are, however, the fundament of many standard inference methods. Further, in many trials more than one outcome variable is observed to assess the effect of a treatment. This can provide an advantageous increase of information, at the same time the family wise type I error rate needs to be controlled, that means the statistical analysis must account for the increased probability of observing a seemingly strong effect in at least one outcome variable in the absence of a true treatment effect.

The aim of this work is to provide in some sense optimal exact hypothesis tests for the comparison of two treatment groups with respect to the success rates of multiple binary endpoints with strict control of the family wise type I error rate.

In a first approach, the discrete permutation joint distribution of marginal Fisher's exact test statistics is used to define multivariate rejection regions such that either the exhaustion of the nominal significance level, the number of elements in the rejection region or the power under a specific alternative is maximized. In contrast to previous similar proposals in the literature, an additional constraint is introduced that ensures that the rejection regions are contiguous sets. This constraint is required for unambiguous interpretation of the test results and to provide powerful multiple testing procedures by virtue of the closed testing principle. The underlying optimization problem is solved by a branch and bound algorithm or, alternatively, by linear integer programming.

In a second approach, the marginal distributions of discrete statistics are used to define optimally weighted Bonferroni tests, meeting optimization criteria analogous to those for the multivariate rejection regions. In addition, greedy algorithms that result in tests with close to optimal exhaustion of the nominal level and robust power properties are derived, using either multivariate joint distributions or marginal distributions of discrete test statistics.

A numeric study of the power of the proposed methods under a wide range of scenarios

with two and three binary endpoints shows that the proposed methods outperform existing procedures such as the Bonferroni-Holm adjustment and the minP test. The application of optimal exact tests is recommended to make best use of the information contained in multiple binary endpoints in the small sample setting.

Kurzfassung

In der jüngeren Vergangenheit hat sich das Interesse an der Erforschung seltener Krankheiten verstärkt. Ebenso besteht großes Interesse daran, die Wirksamkeit von Behandlungen in speziellen kleinen Patientenpopulationen zu untersuchen. Dabei ist jedoch die Anzahl der Patienten, die in eine einzelne Studie eingeschlossen werden können, naturgemäß beschränkt. Die geringen Fallzahlen solcher Studien erschweren die statistische Analyse durch hohe Variabilität von Schätzern und durch möglicherweise unzureichende Genauigkeit asymptotischer Verteilungsapproximationen, die jedoch die Grundlage vieler inferenzstatistischer Verfahren sind. Zusätzlich wird in vielen klinischen Studien der Behandlungseffekt anhand mehrerer Zielvariablen beurteilt. Das kann zu einer vorteilhaften Erhöhung der verfügbaren Gesamtinformation führen. Allerdings muss bedacht werden, dass bei fehlendem Behandlungseffekt die Wahrscheinlichkeit für einen falsch positiven Effekt in zumindest einer Zielgröße mit der Zahl der beobachteten Zielgrößen steigt. Eine konfirmatorische statistische Analyse muss daher so ausgeführt werden, dass diese studienbezogene Wahrscheinlichkeit für eine Fehler erster Art das vorgegebene Signifikanzniveau nicht überschreitet.

Das Ziel dieser Arbeit ist die Entwicklung von exakten Hypothesentests für den Vergleich von zwei Behandlungsgruppen anhand multipler binärer Zielgrößen, die gewisse Optimalitätskriterien erfüllen und gleichzeitig die studienbezogene Wahrscheinlichkeit für eine Fehler erster Art kontrollieren.

Dazu wird in einem ersten Ansatz die gemeinsame diskrete Permutationsverteilung der Teststatistiken von marginalen exakten Tests nach Fisher herangezogen. Für diese Verteilung werden multivariate Ablehnbereiche definiert, sodass entweder die Ausschöpfung des vorgegebenen Signifikanzniveaus, die Zahl der Elemente das Ablehnbereichs oder die Güte des Tests unter einer vorgegebenen Alternative maximiert werden. Die zugrundeliegenden Optimierungsprobleme werden mit einem Branch and Bound Algorithmus

gelöst. Alternativ ist auch eine Formulierung als binäres lineares Programm und dessen Lösung möglich. Im Gegensatz zu vergleichbaren, in der Literatur beschriebenen Methoden wird dabei eine zusätzliche Bedingung eingeführt, die sicherstellt, dass die resultierenden Ablehnbereiche zusammenhängende Regionen sind. Diese Bedingung ist für eine eindeutige Interpretation der Testergebnisse notwendig. Sie bewirkt darüberhinaus, dass die grundsätzlich hohe Güte der vorgeschlagenen Tests auch in multiple Testprozeduren, die aus diesen durch Anwendung des Abschlusstestprinzips gebildet werden, erhalten bleibt.

In einem zweiten Ansatz werden die marginalen diskreten Verteilungen der Teststatistiken benutzt, um optimal gewichtete Bonferroni Tests zu definieren, die Optimalitätskriterien in Analogie zu jenen der multivariaten Ablehnbereiche erfüllen.

Zusätzlich werden sogenannte Greedy Algorithmen erstellt, die auf Basis diskreter gemeinsamer oder marginaler Verteilungen der Teststatistiken Tests mit nahezu optimaler Ausschöpfung des Signifikanzniveaus und robusten Güteeigenschaften ergeben.

Die numerische Untersuchung der Güte der vorgeschlagenen Methoden für zahlreiche Szenarien mit zwei oder drei binären Zielgrößen zeigt, dass für das gestellte Testproblem die hier vorgeschlagenen Methoden existierenden Verfahren wie der Bonferroni-Holm Prozedur oder dem minP Test überlegen sind. Die Anwendung der hier entwickelten optimalen exakten Tests wird daher empfohlen, um die Information aus multiplen binären Zielgrößen in klinischen Studien mit geringer Fallzahl bestmöglich zu nutzen.

1 Introduction

1.1 Confirmative inference in clinical trials

Modern medical research is subject to the principle of evidence based medicine [1], with the gold standard for a single clinicial trial being the randomized controlled trial. In the most simple case of such a trial, patients are randomly assigned to either of two treatment groups, such as placebo and active treatment. The effect of a treatment is assessed by observing one or several outcome variables, and comparing the distribution or some appropriate summary measure of these variables between the two groups. Randomization ensures that any systematic difference between the two groups can be attributed to application of different treatments in a causal manner. Most generally speaking, a treatment is considered efficacious with respect to some outcome measure, if the distribution of the outcome variable is more favourable for the hypothetic population of patients receiving the treatment than for the population of patients under placebo. The observed patients are usually seen as a random sample from these populations, and the observed data is used to test hypotheses about the underlying true distributions. In the majority of cases, the primary hypotheses of a clincial trial are formulated for location parameters.

False test decisions from clinical trials can have wide-ranging implications. False positive conclusions will result in the erroneous believe of having found an efficacious treatment. In this case, a potentially large number of patients will receive a treatment that has no effect at all, or that may have no beneficial effect but still results in unwanted side effects. Also, research for other medication may be put at halt as a consequence of one seemingly efficacious treatment option having become available. A false negative result, on the other hand, will delay or even preclude access of patients to an efficacious treatment.

Therefore, strict control of the type I error rate, i.e. bounding the probability for a false positive test decision, is mandatory in clinical trials. At the same time it is required that hypothesis tests in confirmatory clinical trials have sufficient power to detect a relevant treatment effect. In practice, these prerequisites are controlled through ethics committees and guidance from regulatory agencies, in particular the European Medicines Agency (EMA), the U.S. Food and Drug Administration (FDA) and the Japanese Phar-

maceuticals and Medical Devices Agency (PMDA). Cooperation between these regulatory agencies and also industry parties is in part organized through the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH), which has issued a set of guidelines on the design, conduct and analysis of clinical trials and related aspects. With respect to the application of statistical methods, the ICH guideline E9 "Statistical Principles for Clinical Trials" is relevant [2].

Note that in clinical trials the observable outcome variables related to treatment success are often referred to as endpoints. The focus of this work will be on binary endpoints, that is on Bernoulli distributed outcome variables. These are of particular importance, because in many trials treatment success is measured by the presence or absence of a defined event within a certain period of time after the treatment intervention.

1.2 Research on rare diseases

From a public health perspective, the incentive to invest in rare disease research comes from the consideration that overall a large number of patients is affected by rare diseases, even though there are few patients suffering from each single rare disease. A frequently cited estimation says that in Europe approximately 30 million people are affected by one out of more than 5000 rare diseases, however precise prevalences are difficult to obtain [3]. Investment of pharmaceutical companies in the development of drugs for rare diseases is supported in the European Union by the orphan drug act [4]. This legislation grants a company a certain time frame of market exclusivity for a product that is intended for the cure or diagnosis of a rare disease or another condition, for which it is considered unlikely that enough returns are generated to justify the research investment in economic terms without incentives. In this legislation, a rare disease is defined as affecting not more than five in 10,000 people. In the United States of America, a similar legislation defines as rare a disease that does not affect more than 200,000 persons in the United States [5], which at the time of writing this work approximately corresponds to six in 10,000 people. These definitions are widely used. Note, though, that the foundation for these definitions are economic considerations on the number of required patients to make drug development profitable, which justifies the use of an absolute prevalence threshold as in the US legislation, in contrast to a relative prevalence threshold that would be expected

for a medical based definition.

From a research viewpoint, regardless of economic considerations, the major challenge with rare diseases is the limited number of patients that can be recruited for a single trial, which directly translates to a limited amount of information that can be obtained from a study. In the statistical analysis this is reflected by larger variability of estimates, lower power of hypothesis tests and wider confidence intervals as compared to large studies. Further, inferential methods that are based on asymptotic results may not have the desired properties when applied to small samples. In particular the approximation of the distribution of a test statistic by its limiting distribution may not be accurate enough to provide valid probability statements. Further, in many cases knowledge of nuisance parameters, such as variances or covariances, is required to specify the distribution of the parameter of main interest. While the unknown values of nuisance parameters can be replaced by sample estimates from a large sample, in a small study additional considerations on the uncertainty of these estimates need to be taken into account.

The European Medicines Agency has issued a guideline on the design and statistical analysis of clinical trials in small populations [6]. Despite the inherent difficulties with the analysis of data from small trials, from this guideline it is clear that overall the same standards as in large trials should apply. Additional effort is required, though, to design a trial such that the small number of subjects reveals a large amount of information and, further, to use the information in the data as efficiently as possible. This is not easily accomplished and so methodological research is required to provide methods in pursuit of the spirit of the above statement. In the 7th Framework Program, the European Commission has funded three international research projects that aim to provide improved statistical methodology for the design and analysis of clinical trials for rare diseases or small populations. One of these projects is ASTERIX (Advances in Small Trials dEsign for Regulatory Innovation and eXcellence), and the work of this master's thesis is part of my contribution to the part of the project located at the Medical University of Vienna.

1.3 Remarks on notation and abbreviations

Standard notation is used throughout the work. For clarity, some details are noted here. When dealing with random variables and realizations, random variables are denoted with

upper case letters and their realizations with lower case letters. When describing scalars, vectors, matrices or sets, lower case letters are used for scalars and vectors, upper case letters are used for matrices and sets. Elements of vectors, matrices or sets are indexed by subscripts. If vectors, matrices or sets are themselves subsets of some larger set and indexing within this larger set is required, superscripts in brackets will be used. This also applies when indexing sequences of sets within an iterative algorithm. The superscript T for vectors and matrices means transposed. The operator $|\cdot|$ is used to denote the cardinality (the number of elements) of a set. The notation $x \in \{0,1\}^m$ is used to indicate that x is a vector of length m with entries equal to 0 or 1. For sets A and B, $A\backslash B$ means A without elements contained in B. The symbol P is used to denote probability measures. Subscripts to P are used to further specify the probability measure if required. In particular, P_H is the probability measure under a null hypothesis H. E denotes the expected value of a random variable. exp denotes the exponential function.

The abbreviation s.t. for 'subject to' is used to refer to constraints in an optimization problem. FWER is the abbreviation for family wise type I error rate and HKT for the Hommel and Krummenauer improvement of Tarone's test.

1.4 Testing statistical hypotheses

In this work we will be mainly concerned with testing parametric hypotheses. The basic definitions of a hypothesis test, type I error rate and power are given here. Details and a more theoretic background can be found, e.g., in [7] or [8].

Let X be a sample of observations taken from a sample space $\mathcal{X}$, let further be $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ a family of distributions that are defined on that sample space and that are characterized by a parameter vector $\theta \in \Theta$, where Θ is the parameter space. A hypothesis H is a subset of $\mathcal{F}$. Typically, a pair of hypotheses, the null hypothesis $H_0 \in \mathcal{F}$ and the alternative hypothesis $H_A = \mathcal{F}\backslash H_0$ are defined and we are interested in deciding whether the true data generating distribution of X is in H_0 or in H_A. If H_0 is defined through a single member of $\mathcal{F}$ it is called a simple null hypothesis, else H_0 is a composite null hypothesis. In terms of notation, once the statistical model is specified, it will in most cases suffice to state the hypotheses in terms of the parameter vector solely, e.g. $H_0 : \theta \in \Theta_0$ for some $\Theta_0 \subseteq \Theta$.

A statistical test φ is a measurable function $\varphi : \mathcal{X} \to \{0, 1\}$. If $\varphi(X) = 1$, the null hypothesis is rejected, else it is retained. A test is performed by calculating the value of a suitable test statistic T, which is a measurable function of the observed data, $T : \mathcal{X} \to V$. Typically, the target set V of a single hypothesis test is unidimensional, however when testing multiple hypotheses simultaneously as will be done in this work, tests with multidimensional V are found. The rejection region of the test is a set $R \subseteq V$, and $T(X) \in R$ directly means $\varphi(X) = 1$. The choice of R determines the probability to reject H_0, both, under the null and alternative hypothesis. These probabilities are formally characterized through the type I error rate and the power as defined below.

The rejection of H_0 even though H_0 is true, is called type I error. For a specific $F_\theta \in H_0$ we call $P_{F_\theta}(\varphi = 1)$ the type I error rate of the test φ under $F_\theta \in H_0$. If H_0 is a composite null hypothesis, the maximal type I error rate under H_0, $\sup_{F_\theta \in H_0} P_{F_\theta}(\varphi = 1)$ is of interest. A parameter configuration under which the maximal type I error rate is attained is referred to as the least favorable configuration under H_0. The rejection region R is typically chosen in such a way, that the maximal type I error rate is bounded from above by a pre-specified value. We call this value the nominal significance level of the test and it is commonly denoted by α. For brevity, α will often be referred to as significance level or just as level. We will sometimes call a test for which the maximal type I error rate is bounded by α a level α test. If the maximal type I error rate is below the nominal significance level, the test is called conservative. If a null hypothesis is rejected by a level α test we say the test (for the particular data) is significant at level α.

For specific $F_\theta \in H_A$, we call $P_{F_\theta}(\varphi = 1)$ the power of the test φ under F_θ. Non-rejection of H_0 even though the alternative holds is called type II error, and so $P_{F_\theta}(\varphi = 0)$ for $F_\theta \in H_A$ may be referred to as type II error rate, though this term will not be used here.

If it is possible to find an order on V with respect to how strongly different realizations of T contradict the null hypothesis, it is possible to define a p-value as $p(t) = \sup_{F_\theta \in H_0} P_{F_\theta}(T \geq t), t \in V$. When using the p-value for the test decision, $p \leq \alpha$ leads to rejection of H_0 at the significance level α. This rule implies a certain rejection region.

In the medical setting outlined above, a simple null hypothesis may correspond to a model in which the observations from both treatment groups come from the same distribution. Note however that in many applications the interest is not in the full parameter

vector θ but rather in a single parameter that is, e.g., connected to the mean of a distribution. The further elements of θ then are nuisance parameters. The term point (null) hypothesis will be used if the hypothesis specifies a single value for the parameter of interest but the model possibly contains further nuisance parameters which are not determined at single values under that hypothesis.

Constructing a level α hypothesis test for a pair (H_0, H_A) thus requires to define a test statistic T, to identify its distribution under the least favorable configuration in H_0, and to find a rejection region R, such that $\sup_{F_\theta} P_{F_\theta}(T \in R) \leq \alpha$. Further, without yet specifying any optimality criteria, T and R should be chosen such that the test has large power under possibly all relevant members of H_A. In the presence of nuisance parameters with unknown values, there is typically some extra challenge involved in finding the least favorable configuration under H_0 and the distribution of T, as will be exemplified below.

1.5 Two group comparison of one binary endpoint

Throughout this work, we will consider tests to compare two treatment groups, e.g. active treatment versus placebo, with respect to one ore more binary endpoints. In this section, the statistical set-up and tests for two-group comparison with respect to a single endpoint are described. The hypotheses of interest will always be one-sided, because in a typical confirmative clinical trial there is clearly one direction of the treatment effect that needs to be shown. Denote with $i \in \{1, 2\}$ the treatment group, and with n_i the number of observations in group i. In a trial with random treatment allocation, the n_i are subject to some randomness, depending on the randomization strategy. In the analysis strategies discussed here, the n_i will be considered fixed, that is, inference is conditional on the observed group sizes. Let y_{ij} be the observed outcome of the j-th observational unit in group i and Y_{ij} the corresponding random variable with possible values in $\{0, 1\}$. As data generating model, we assume that the Y_{ij} are independently and within each group identically Bernoulli distributed with parameter $p_i = P(Y_{ij} = 1)$ taking values in $(0, 1)$. We are interested in the null hypothesis $H_0 : p_1 \leq p_2$. Equivalently the null hypothesis can be stated in terms of the odds ratio $OR = (p_1/(1 - p_1))/(p_2/(1 - p_2))$, $H_0 : OR \leq 1$. This has some particular importance, as the log-odds is the natural parameter of the exponential family form of the Bernoulli distribution.

The data can be aggregated in a 2×2 frequency table without loss of information, as is shown in Table 1.

Table 1: Outline of a 2×2 frequency table.

	Group 1	Group 2	Total
Success	$\sum_{j=1}^{n_1} y_{1j}$	$\sum_{j=1}^{n_2} y_{2j}$	$m_1 = \sum_{j=1}^{n_1} y_{1j} + \sum_{j=1}^{n_2} y_{2j}$
Failure	$n_1 - \sum_{j=1}^{n_1} y_{1j}$	$n_2 - \sum_{j=1}^{n_2} y_{2j}$	$m_2 = N - m_1$
	n_1	n_2	$N = n_1 + n_2$

Hence, tests for comparing the proportions p_1 and p_2 are closely related to tests of independence in 2×2 tables. Note that our model amounts to the binomial sampling scheme for 2×2 tables [9]. For most tests of this problem the case $p_1 = p_2$ will be the least favorable configuration, so that it is often enough to consider the point null hypothesis $H_0' : p_1 = p_2$. This is still a composite null hypothesis as the common value of p_1 and p_2 can be any value $p_0 \in (0, 1)$. Thus, the sampling distribution of the data and hence any test statistic are not fully determined by stating H_0', but further depend on the unknown value of the nuisance parameter p_0.

The problem of testing H_0' in the presence of the unknown p_0 has been studied extensively for more than 100 years and numerous possible tests have been derived and studied, see e.g. [10–12] for a comprehensive overview. There are three general types of solutions in dealing with the nuisance parameter: Asymptotic tests, that rely on consistently estimating p_0; exact conditional tests, in which the influence of the nuisance parameter is removed by conditioning on a sufficient statistic; and exact unconditional tests, in which the worst case over all possible values of p_0 is considered.

In the following, the asymptotic score test is described in detail, to demonstrate the role of p_0 in the asymptotic approach. Then the conditional exact test originally proposed by Fisher [9, 13] is derived. This test is most relevant for the remainder of the work, because our aim is to provide exact inference for multiple binary endpoints based on a conditional joint distribution of multiple Fisher's exact test statistics or on the marginal exact distributions of these test statistics. Nonetheless, a brief overview on the principle of unconditional exact tests is included in this section as well as some reference on the discussion on the proper choice between the testing approaches.

1.5.1 Asymptotic score test

The maximum likelihood estimate for p_i is the sample mean $\hat{p}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$. The restricted maximum likelihood estimate under the restriction $p_1 = p_2$ is the grand mean $\hat{p}_0 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}/(n_1 + n_2)$. For testing H_0 in large samples, the asymptotic score test (see [7] chapters 3 and 8) is often used. The test statistic is $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ and the test rejects $H_0 : p_1 \leq p_2$ at level α if $Z \geq z_{1-\alpha}$, where $z_{1-\alpha}$ denotes the $1-\alpha$ quantile of the standard normal distribution.

The test statistic can be rewritten as

$$Z = \frac{\sqrt{n_1}(\hat{p}_1 - p_1)}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(1+\frac{n_1}{n_2}\right)}} - \frac{\sqrt{n_2}(\hat{p}_2 - p_2)}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(1+\frac{n_2}{n_1}\right)}} + \frac{p_1 - p_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

Under the point null hypothesis $p_1 = p_2$, it is easily shown by standard results of asymptotics [14] that Z converges in distribution to a standard normal random variable as $n_1 \to \infty$ and $n_2 \to \infty$. (Note that unless $\frac{n_1}{n_2}$ converges to a finite constant, the sample size of the group with slower growth will determine the asymptotics, still the result holds.) If $p_1 \neq p_2$, the two leading terms in Z converge to a normal distribution with mean zero and some variance not necessarily equal to 1. The final term, however, goes to $+\infty$, if $p_1 > p_2$, and to $-\infty$, if $p_1 < p_2$. Therefore, the test is consistent for each alternative $p_1 > p_2$, which means the probability to reject H_0 goes to 1 with increasing n_1 and n_2. Similarly, the probability to reject H_0 goes to zero for $p_1 < p_2$. So $p_1 = p_2$ is indeed the least favorable configuration, and in this case the probability to reject goes to α.

Note that this test is equivalent to Pearson's Chi-squared test of independence for 2×2 tables. Further note that it is asymptotically locally most powerful for alternatives in the immediate neighbourhood of $p_1 = p_2$ [7]. For other alternatives and $n_1 \neq n_2$ the corresponding Wald test can be more powerful. (For $n_1 = n_2$ they are identical.)

In any case, the true finite sample distribution of the test statistic is discrete and for small sample sizes the normal approximation may result in a notable inflation of the type I error rate. A common rule of thumb is that the expected number of events of either type (0 or 1) in each group should be at least 5 [15]. A continuity correction can be applied, largely removing the problem, however at the price of an overly conservative behaviour of the test [15].

1.5.2 Conditional exact test (Fisher's test)

In Fisher's exact test, inference is conditional on both margins of the 2×2 table representing the data (Table 1). The test statistic is the entry of either of the four cells in the table, e.g. $T = \sum_{j=1}^{n_1} Y_{ij}$. It is easy to see that under H_0' and conditional on the margins this statistic has a hypergeometric distribution. That is

$$P_{H_0'}(T = t) = \frac{\binom{m_1}{t}\binom{m_2}{n_1 - t}}{\binom{N}{n_1}}$$

for values of $t \in V = \{s \in \mathbb{N} : \max(0, n_1 - m_2) \leq s \leq \min(m_1, n_1)\}$. H_0 is rejected at level α if the observed t is such that $P_{H_0'}(T \geq t) \leq \alpha$. It will be shown below, that H_0' is the least favorable configuration under H_0 for this test.

The test controls the conditional type I error rate at level α, that means $E(\varphi|\text{observed margins}) \leq \alpha$ for all possible values of the table margins. Therefore the test also controls the unconditional type I error rate, since by monotonicity of the expectation $E(\varphi) = E(E(\varphi|\text{observed margins})) \leq \alpha$.

In line with the rejection rule, the one-sided p-value for Fisher's exact test is $p(t) = P_{H_0'}(T \geq t)$. A possible p-value for a two-sided test can be defined, by using the order induced by the probability under H_0', as $p(t) = P_{H_0'}\{s \in V : P_{H_0'}(T = s) \leq P_{H_0'}(T = t)\}$. Other definitions are possible, see [16]

There are different ways to derive Fisher's exact test. Here, the approach via a logistic regression model is pursued, because it highlights the importance of the log-odds and the odds-ratio as natural parametrization and further makes obvious the connection of Fisher's exact test with other models of conditional inference in the class of exponential families and in regression modelling (compare [7, 9]). In the following, the general term density will be used with the understanding that it can be applied to discrete distributions, too, and there it means the same as probability mass function. We will require the following definition for a sufficient statistic, adapted from the book by Shao [17].

Definition 1. (Sufficient statistic) Let X be a sample from a distribution $F \in \{F_\theta : \theta \in \Theta\}$. A statistic $T(X)$ is sufficient for θ, if the conditional distribution of X given $T(X)$ does not depend on θ.

Thus, a conditional statistic contains all the information on θ that is in X. Further, to identify a sufficient statistic the factorization theorem, adapted from [17] and with a proof therein, is useful.

Theorem (Factorization theorem): Let X be a sample from a distribution $F \in \{F_\theta : \theta \in \Theta\}$. A statistic $T(X)$ is sufficient for θ if and only if there are non-negative measurable functions g and h such that the density of X can be written as $f(x) = g(T(X), \theta)h(x)$.

This means, the density of X can be factorized in two terms, one depending on θ and on the data only through $T(X)$, the other depending on the data, but not on θ.

The density of a Bernoulli random variable Y with parameter $p = P(Y = 1)$ is

$$f(y) = p^y(1-p)^{(1-y)}, y \in \{0, 1\}$$

and $f(y) = 0$ for $y \notin \{0, 1\}$. The Bernoulli distribution is a member of the one-parameter exponential family. Using the natural parameter $\theta = \log\left(\frac{p}{1-p}\right)$, the density is written in the canonical form as

$$f(y) = \exp\left(y\theta - \log(1 + \exp(\theta))\right)$$

where exp denotes the exponential function. In our model we have Bernoulli distributed observations from two groups with parameters p_i or $\theta_i = \log\left(\frac{p_i}{1-p_i}\right), i = 1, 2$. To jointly model the distribution of the observations from both groups, the log-odds for the j-th observation in group i are written as $\theta_{ij} = \beta_0 + x_{ij}\beta_1$, where $x_{ij} = 1$ for $i = 1$ and $x_{ij} = 0$ for $i = 2$. In this parametrization, the odds ratio is

$$OR = (p_1/(1-p_1))/(p_2/(1-p_2)) = \exp(\beta_1)$$

So the null hypothesis $H_0 : p_1 \leq p_2$ is equivalently stated as $H_0 : \beta_1 \leq 0$. β_0 is the nuisance parameter in this parametrization.

Since the Y_{ij} are independent, the joint density is

$$f(y; x, \beta) = \prod_{i=1}^{2}\prod_{j=1}^{n_i} f(y_{ij}; x_{ij}, \beta) =$$

$$= \exp\left(\beta_0 \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}\right) \exp\left(\beta_1 \sum_{i=1}^{2}\sum_{j=1}^{n_i} x_{ij}y_{ij}\right) \prod_{i=1}^{2}\prod_{j=1}^{n_i} (1 + \exp(\beta_0 + x_{ij}\beta_1))^{-1} \quad (1)$$

By the factorization theorem, the total number of events, $M_1 = \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}$ is a sufficient statistic for β_0. The number of events in group 2, $T = \sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij}$, is a sufficient statistic for the log-odds ratio β_1. Conditionally on the observed total number of events m_1, the distribution of the data does not depend on the nuisance parameter β_0. Thus, by performing inference conditionally on m_1 the nuisance parameter is removed. The conditional distribution still depends on β_1, the parameter of interest, and the information we have on β_1 is captured by its sufficient statistic T. Therefore T is used as test statistic.

The role of the two groups and event versus non-event can be exchanged by accordingly changing the definitions of x and p. That is, in a 2×2 table of the number of events by group, the observed frequency in any arbitrarily chosen cell can be used as test statistic with identical results.

We already know that T is hypergeometrically distributed under $H_0' : OR = 1$. For an arbitrary value λ of the odds-ratio, and conditional on the table margins, T has the following non-central hypergeometric distribution

$$P_{OR=\lambda}(T = t) = \frac{\binom{m_1}{t}\binom{m_2}{n_1-t}\lambda^t}{\sum_{s\in V}\binom{m_1}{s}\binom{m_2}{n_1-s}\lambda^s}$$

for values of $t \in V = \{s \in \mathbb{N} : \max(0, n_1 - m_2) \leq s \leq \min(m_1, n_1)\}$. For a derivation see the multivariate case in sections 4.1 and 4.2, of which this result is a special case.

Now knowing the general distribution of T, the following lemma shows that the probability to reject $H_0 : OR \leq 1$ with Fisher's exact test is monotone in the odds ratio. As a consequence H_0' is the least favorable configuration in H_0.

Lemma 1. Let T be a discrete random variable with $n < \infty$ distinct possible values $v_1 < ... < v_n$ and a probability distribution given by $P_\lambda(T = v_i) = \frac{w_i\lambda^{v_i}}{\sum_{j=1}^{n} w_j\lambda^{v_j}}$, with $w_i > 0, \forall i = 1, ..., n$ and $\lambda \in (0, \infty)$. Then the probability $P_\lambda(T \geq v_k)$ is strictly increasing in λ for all $1 < k \leq n$, i.e. $P_{\lambda+\epsilon}(T \geq v_k) > P_\lambda(T \geq v_k)$ for all $\epsilon > 0$ and $1 < k \leq n$.

Proof. $P_\lambda(T \geq v_k) = \frac{\sum_{i=k}^{n} w_i\lambda^{v_i}}{\sum_{j=1}^{n} w_j\lambda^{v_j}}$ increasing in λ is equivalent to the odds $\frac{P_\lambda(T\geq v_k)}{1-P_\lambda(T\geq v_k)} = \frac{\sum_{i=k}^{n} w_i\lambda^{v_i}}{\sum_{j=1}^{k-1} w_j\lambda^{v_j}}$ increasing in λ. Let $\epsilon > 0$, and $f_i = \left(\frac{(\lambda+\epsilon)}{\lambda}\right)^{v_i}$. $\frac{\lambda+\epsilon}{\lambda} > 1$ and so $f_{i-1} < f_i, \forall i = 2, ..., n$. Then

$$\frac{\sum_{i=k}^{n} w_i(\lambda+\epsilon)^{v_i}}{\sum_{j=1}^{k-1} w_j(\lambda+\epsilon)^{v_j}} = \frac{\sum_{i=k}^{n} w_i\lambda^{v_i}f_i}{\sum_{j=1}^{k-1} w_j\lambda^{v_j}f_j} > \frac{\sum_{i=k}^{n} w_i\lambda^{v_i}f_k}{\sum_{j=1}^{k-1} w_j\lambda^{v_j}f_{k-1}} = \frac{f_k}{f_{k-1}}\frac{\sum_{i=k}^{n} w_i\lambda^{v_i}}{\sum_{j=1}^{k-1} w_j\lambda^{v_j}} > \frac{\sum_{i=k}^{n} w_i\lambda^{v_i}}{\sum_{j=1}^{k-1} w_j\lambda^{v_j}}$$

□

1.5.3 Consequences of discreteness

Example 1. Assume $n_1 = n_2 = 15$ and a marginal number of successes $m_1 = 20$. Table 2 shows the exact conditional distribution under H_0' of the test statistic $T = \sum_{j=1}^{n_1} Y_{ij}$. In the last column of the table the rejection region R for the test of H_0 at the nominal level $\alpha = 0.025$ is indicated. Here, T can take values between 5 and 15, and the rejection region is $R = \{14, 15\}$. The actual conditional type I error rate is $P_{H_0'}(R) = 0.0026$, which is far below the nominal level. However, the next largest region $\{13, 14, 15\}$ would result in a conditional type I error rate of 0.0251, exceeding the nominal level.

Table 2: Null-distribution of T in the example. Probabilities are rounded to four decimal places. $\mathbb{1}_{\{t \in R(0.025)\}}$ indicates if t is in the level $\alpha = 0.025$ rejection region.

t	$P_{H_0'}(T = t)$	$\mathbb{1}_{\{t \in R(0.025)\}}$
5	0.0001	0
6	0.0025	0
7	0.0225	0
8	0.0975	0
9	0.2274	0
10	0.3001	0
11	0.2274	0
12	0.0975	0
13	0.0225	0
14	0.0025	1
15	0.0001	1

The example illustrates a general property of discrete tests. The rejection region can only grow incrementally, therefore its probability mass under the null hypothesis is in most cases below the nominal level. As a consequence, the resulting rejection region may contain only a small number of quite extreme events, such that its probability mass under an alternative may also be small. The number of possible values for T has a huge impact on how strongly discreteness affects the characteristics of a test. The test statistic of Fisher's exact test has for small sample sizes a very limited number of possible realizations and as

such the test can become quite conservative.

A standard way of performing a test with a discrete statistic such that the nominal level is exhausted (and the power is increased) is of course to randomize the test decision if that value of the test statistic is observed, which is just outside the strict level α rejection region R, such that the $E(\varphi) = \alpha$ [8]. That means, for this particular outcome t^*, $\varphi = 1$ with probability under the null hypothesis $(\alpha - P(R))/P(t^*)$, and otherwise $\varphi = 0$. In the example, the randomized test would always reject H_0 if $t \geq 14$, and it would reject H_0 with a probability of 0.996 if $t = 13$. The randomized version of Fisher's exact test is known to be the uniformly most powerful unbiased (UMPU) test for H_0 [8, 16, 18]. A test is unbiased if its rejection probability is larger under every alternative than under any configuration in the null hypothesis.

There is, however, the convincing principle that applying the same test to the same data should give the same result [7]. This principle may be weakened to some extent to allow for small variations in the results of resampling methods. It clearly opposes the application of randomized tests for the analysis of an actual data set, though. And indeed, randomized tests are not used in the clinical trial setting.

However, the idea of randomization has led to a suggestion that in some way approximates the properties of the randomized test and meets the above principle, that is the use of so called mid-p-values [19]. The one-sided p-value for Fisher's exact test is $p(t) = P_{H_0'}(T \geq t)$. The one-sided mid p-value is the average between $p(t)$ and $p(t+1)$, i.e. $\textit{mid-}p(t) = 0.5 * P_{H_0'}(T = t) + P_{H_0'}(T > t)$. With this approach, H_0 is rejected if $\textit{mid-}p(t) \leq \alpha$. The mid-p-value has the appealing properties that its expectation under the point null hypothesis is 0.5, and that the sum the of two one-sided mid-p-values $0.5 * P_{H_0'}(T = t) + P_{H_0'}(T > t)$ and $0.5 * P_{H_0'}(T = t) + P_{H_0'}(T < t)$ is 1, similar to p-values for continuous test statistics. In contrast, for the usual p-value $P_{H_0'}(T \geq t)$ the respective values are above 0.5 and 1, indicating the conservativeness of decisions based thereon. In the example, the use of the mid-p-value would result in the rejection region $\{13, 14, 15\}$.

As a drawback, the mid-p-value approach does not strictly control the unconditional type I error rate at the nominal level. However, the actual level of the test will mostly be close to the nominal level without severe exceedance [9, 19, 20]. Several authors argue in favor of using Fisher's exact test with the mid-p-value approach [11, 12, 19–21]. Fisher's exact test is used frequently in clinical trials, and in the majority of cases it seems the

exact conservative p-value is used. There are however many examples of clinical trials in which the mid-p-value was used for the test decision, e.g. in studying medication for Huntinton's disease [22], in neurology [23] and in breast cancer research [24].

A less well known consequence of discreteness is that in some cases the unconditional power of a discrete test may in fact increase if the sample size is slightly decreased or the imbalance between group sample sizes is increased [25]. This property is not restricted to Fisher's exact test, though, and can even be observed for the asymptotic score test for certain parameter configurations, see Figure 2 in [12]. This property has little impact on the actual choice of a test, as usually the amount of non-monotonous change in power is small and in a clincial trial the sample size in each group is often not pre-determined precisely due to dependence on random allocation and the possibility of patients dropping out of the study.

1.5.4 Unconditional exact tests by maximization

Consider a test for $H_0 : p_1 \leq p_2$ with least favorable configuration $p_1 = p_2 = p_0$. Under this configuration, for a given value of p_0 the data are exactly Bernoulli distributed with success probability p_0 and the number of successes in each group follows exactly a corresponding binomial distribution. Thus, the exact distribution of the test statistic can be calculated. Importantly, this distribution is based on a considerable number of possible events and so the test would be affected by discreteness to a much lesser extent than the conditional exact test. In the example with $n_1 = n_2 = 15$, $(n_1 + 1) * (n_2 + 1) = 265$ different outcomes are possible. Even though not all of these will necessarily lead to different values of a test statistic, the number of possible values for the test statistic will usually be larger than in the conditional approach.

Denote with φ_{p_0} such an exact test for the specific choice of p_0. The common principle of unconditional exact tests is to reject H_0 if $\inf_{p_0 \in (0,1)} \varphi_{p_0} = 1$. Equivalently, the p-value for an unconditional exact test with observed test statistic t is

$$p(t) = \sup_{p_0 \in (0,1)} P_{p_1 = p_2 = p_0}(T \geq t) \tag{2}$$

This approach was first proposed by Barnard [26, 27]. Few accounts on this test fail to notice that he later favored the use of Fisher's exact test instead of this approach

on behalf of philosophical arguments over the proper choice of the reference set for an experimental outcome [16, 21, 28].

Starting from a slightly different viewpoint, Boschloo [29] studied the maximal unconditional level of Fisher's exact test over the range $p_0 \in [0, 1]$. He suggested to use a significance level increased as much as possible for the test decision, such that the maximal unconditional type I error rate over $p_0 \in [0, 1]$ was still controlled at the original nominal level. This procedure is equivalent to using the p-value from Fisher's exact test as test statistic in (2). By construction, this test is never less powerful than Fisher's exact test, and in many scenarios has considerably larger power. Some other choices of a test statistic for use in (2) perform poorly though, see [30] for a numeric investigation.

In a further refinement, the confidence set method by Berger and Boos is used [31]. There, H_0 is rejected if $\sup_{p_0 \in K(\gamma)} P_{p_1=p_2=p_0}(T \geq t) + \gamma \leq \alpha$, where $K(\gamma)$ is an exact $1 - \gamma$ confidence interval for p_0. With this approach, maximization is performed over a reasonably restricted range of p_0 given by a confidence interval, and this restriction is accounted for by adding the probability of non-coverage to the p-value. The performance of this method depends on the choice of γ, typically a small value such as 0.0001 is chosen.

The major difficulty with unconditional exact tests is the actual calculation of the maximum over the continuous interval for values of p_0. In general, there seems to be no closed form solution to (2) available, therfore the solution is found by numeric search over a fine enough grid. Equation (2) can have several local maxima [12, 27]. Hence, a considerable number of points for p_0 needs to be evaluated and at each point the exact distribution, too, needs to calculated numerically. This is feasible for a single nuisance parameter, but it can be restrictive in multidimensional settings.

1.5.5 Choosing a test

In the literature there is a long standing discussion on the proper choice between the approaches outlined above and between tests within each approach. In particular, arguments in favor as well as against the use of Fisher's exact test are laid out in numerous works, with some authors prominently changing opinion over the years [21]. Hirji et al. [20] give a comprehensive overview on the different positions. Briefly, the proponents of strict conditional inference argue that inference should be based on the observed data only, such that conditional inference is the only meaningful choice. The argument does

not withhold closer scrutiny on the extent of aspects to condition on, as some randomness in the data needs to be acknowledged to generalize inferential findings to some population beyond the actual sample. An argument along similar lines in favor of unconditional exact inference is that no conditioning should be done on parameters that are not fixed by design.

Looking at the discussion from some distance, most authors, including myself, regard the unconditional frequentist properties of a test as essential criteria. Both, conditioning and unconditional maximization are valid inferential tools providing tests with control of the unconditional type I error rate for the testing problem laid down at the beginning of this section. Only with some particular sampling schemes, such as response adaptive randomization, extra care in defining the sample space and the exact distributions is required [32]. For actual applications, unconditional type I error rate control, unconditional power and feasibility of computation are the relevant aspects guiding the choice for a test.

In this work, an exact conditional approach will be pursued to construct tests for hypotheses on multiple binary endpoints. The approach will be exact, to provide strict type I error rate control in the setting of small sample sizes, and among the exact tests conditional methods will be preferred for several reasons. Firstly, a general framework of exact tests for multiple hypotheses will be proposed, which is applicable to multivariate test statistics with discrete distributions. This framework is most easily applied to the setting of binary endpoints when well tractable conditional exact distributions are used. Secondly, the influence of discreteness, if properly taken into account, can be strongly reduced as soon as a global hypothesis test (see section 1.7) for more than one endpoint is considered. This is true for, both, approaches regarding the multivariate joint distribution of test statistics and the joint treatment of marginal discrete distributions. Therefore the main argument that can be brought forward against the use of an exact conditional approach has little strength here. Decisions on individual hypothesis in the proposed closed testing approach (see sections 1.6.4 and 3.3) could be improved, though, by application of maximization tests for individual hypotheses. This method will not be further studied, but it is straight forward to use in actual applications. Third, for the multivariate case, computation of an unconditional maximization can be computationally extremely demanding. This is illustrated by an example with two binary endpoints and two groups with sample size $n_1 = n_2 = 15$. There are 5456 possible outcomes and at least three

nuisance parameters require consideration, the marginal success rates for each endpoint and the joint probability for success in both endpoints. If 100 values are to be evaluated for each nuisance parameter, the total number of scenarios is in the order of 10^9. Thus, a conditional permutation approach is preferred here.

1.6 Multiple testing procedures

In many clinical trials multiple endpoints are evaluated to assess the treatment effect. This can be necessary, because the condition under study cannot be fully characterized by only one endpoint. In other cases, multiple endpoints may be investigated if it is sufficient to show overall efficacy for a set of related endpoints. For such a claim of efficacy it is sufficient to reject at least one elementary null hypothesis for an individual endpoint, or even to just reject the global intersection null hypothesis (see below) of no effect in any endpoint. In this section, important concepts of testing multiple hypotheses will be introduced.

1.6.1 Intersection hypotheses

Denote with $\mathcal{H} = \{H_1, ..., H_k\}$ a family of hypotheses. E.g. in a clinical trial, $\mathcal{H}$ can contain null hypotheses of no treatment effect with respect to k different endpoints. Another example are hypotheses regarding different pair-wise comparisons of treatment groups with respect to one endpoint. It is often useful, both, for the interpretation of scientific results as well as for the formal study and derivation of multiple testing procedures to regard intersection hypotheses.

Definition 2. Intersection hypothesis. Let $\theta \in \Theta$ be a vector of parameters in a statistical model. Let $H_i : \theta \in \Theta_0^{(i)}$ for $\Theta_0^{(i)} \subseteq \Theta, i = 1, ..., k$, be hypotheses in this model. Let $I \subseteq \{1, ..., k\}$ be an index set. The intersection hypothesis $\cap_{i \in I} H_i$ is $\theta \in \cap_{i \in I} \Theta_0^{(i)}$. Simply put, the intersection hypothesis states that all hypotheses $H_i, i \in I$ are true simultaneously. The intersection of all hypotheses in a family $\mathcal{H}$ is often called global intersection hypothesis.

Remark: When considering hypotheses on multiple endpoints, the statistical model in definition 2 is in fact a model for a multivariate outcome. However, the hypotheses are typically formulated for models of smaller dimension, in most cases with a univariate

outcome variable. With respect to the multivariate model, these hypotheses are composite in the sense that the parameters unrelated to the specific lower dimensional (or univariate) model are not specified.

An example illustrates the concept of intersection hypotheses in the setting of binary endpoints.

Example 2. Assume a clinical trial with two treatment groups and two binary endpoints. Denote with θ_i the log odds-ratio for the between group comparison with respect to endpoint $i = 1, 2$. We want to test the two null hypotheses $H_1 : \theta_1 \leq 0$ and $H_2 : \theta_2 \leq 0$. In terms of the multivariate model, the null hypotheses for endpoint 1 would be written $H_1 : \theta = (\theta_1, \theta_2) \in \Theta_0^{(1)} = \{x \in \Theta = \mathcal{R}^2 : x_1 \leq 0\}$, and the null hypothesis for endpoint 2 is defined analogously. The intersection hypothesis $H_1 \cap H_2$ then is $\{(\theta_1, \theta_2) \in \mathcal{R}^2 : x_1 \leq 0, x_2 \leq 0\}$. Figure 1 shows a graphical representation of this example.

Note that in both, the univariate and the multivariate way of expressing the null hypotheses, nuisance parameters are omitted. For the univariate models these are the actual success probabilities in either group. For the multivariate model, there are two additional nuisance parameters that capture the association between the two endpoints in either group. This association can be parametrized by the probability for a simultaneous success in both endpoints or by the product moment correlation between the endpoints.

It will become clear in the next subsection that tests for intersection hypotheses are an important tool in the construction of multiple testing procedures. In many applications there also is an interest in intersection hypotheses in their own right. When testing hypotheses about a treatment effect on multiple endpoints, rejecting an intersection hypothesis allows for the conclusion of an effect in at least one endpoint, although it remains to be established for which endpoints specifically the treatment is effective.

1.6.2 Family wise type I error rate

The high importance of type I error rate control in clinical trials comes from the requirement to limit the risk for false positive conclusions (compare section 1.1). In many clinical trials, more than one statistical hypothesis is tested to assess the potential benefit of a treatment. In this case care must be taken to limit the risk for false positive results, because the probability to observe an extreme event with respect to any one of these

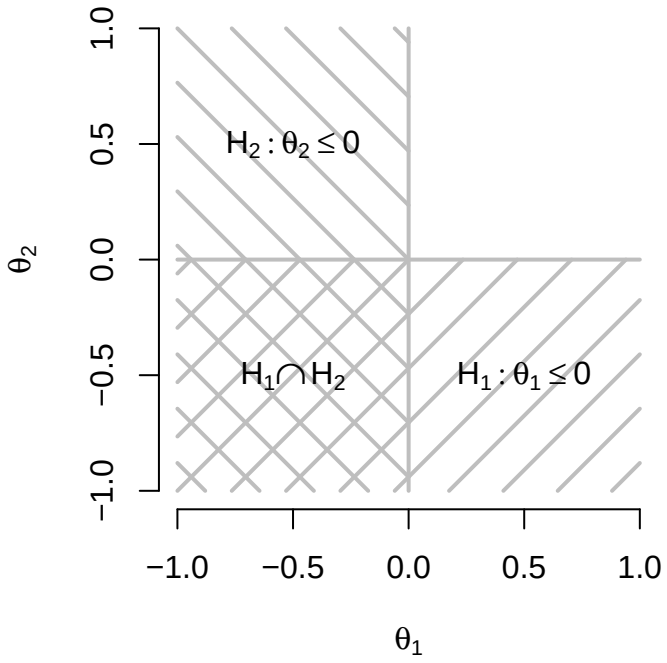

Figure 1: Intersection of the hypotheses $H_1 : \theta_1 \leq 0$ and $H_2 : \theta_2 \leq 0$

hypotheses is larger than the probability to observe a similarly extreme event for just one single hypothesis. Thus, if several null hypotheses are tested by individual level α tests, and given there is more than one true null hypothesis, the probability for a false rejection will be larger than α. This is often referred to as inflation of the type I error rate. Usually, control of the type I error rate is warranted for the overall claim of a study. Therefore, control of the family wise type I error rate (FWER) as defined below is required. The importance of FWER control is highlighted by the respective guideline of the European Medicines Agency [33].

Definition 3. (Family wise type I error rate (FWER)) Consider a family of null hypotheses $\mathcal{H} = \{H_1, ..., H_k\}$ and a testing procedure that results in a decision on each hypothesis in $\mathcal{H}$. The family wise type I error rate of this procedure is the probability for at least one rejection of a true null hypothesis in $\mathcal{H}$, regardless of the configuration of true and false hypotheses in $\mathcal{H}$. Formally, let the index set $I \subseteq \{1, ..., k\}$ indicate the

true null hypotheses and let $H_I = \cap_{i \in I} H_i$. Then,

$$\text{FWER} = \sup_{I \subseteq \{1,...,k\}} \sup_{F_\theta \in H_I} P_{F_\theta}(\text{reject at least one } H_i : i \in I)$$

A testing procedure is said to control the FWER at level α if FWER $\leq \alpha$.

The following example gives some illustration on the extent of the possible inflation of the type I error rate, when a naive testing approach is pursued.

Example 3. Assume that k null hypotheses $H_i, i = 1, ..., k$ are tested by respective level α tests $\varphi_i, i = 1, ..., n$. Further assume all these null hypotheses are true, such that the marginal probabilities for rejection are $P(\varphi_i = 1) = \alpha$ for all $i = 1, ..., k$, and that the tests are stochastically independent, that means $P(\varphi_i = 1, \varphi_j = 1) = P(\varphi_i = 1)P(\varphi_j = 1), \forall i \neq j$. Then the probability to falsely reject at least one null hypothesis is $1-(1-\alpha)^k$. A first order Taylor expansion of this expression around 0 gives $k\alpha(1-\alpha)^k$, which is approximately $k\alpha$ for k small and α close to 0. For instance, if $k = 3$ hypotheses are tested at the usual one-sided level $\alpha = 0.025$, this probability is $1 - (1 - 0.025)^3 = 0.073 \approx 3 * 0.025$. For $k = 5$ and $\alpha = 0.05$, $1 - (1 - 0.05)^5 = 0.223 \approx 5 * 0.05$. For growing k the probability for a rejection approaches 1.

Thus, when FWER control is warranted, decisions on individual hypotheses must be based on a multiple testing procedure that takes into account the joint probability for rejections across the whole family of hypotheses and involved tests. Important concepts for such multiple testing procedures are covered in the following subsections.

1.6.3 Bonferroni test

The Bonferroni test, or Bonferroni adjustment, is a widely used method to test a family of hypotheses $\mathcal{H} = \{H_1, ..., H_k\}$ with FWER control. In a Bonferroni testing procedure, each hypothesis $H_i, i = 1, ..., k$ is tested by a corresponding test φ_i at a local level $\alpha_i \in [0, 1]$. H_i is rejected if $\varphi_i = 1$. As key property of a Bonferroni procedure, the local levels are chosen such that $\sum_{i=1}^k \alpha_i \leq \alpha$. Then the FWER of this procedure is bounded by α.

Proof. Use the same notation as in definition 3. For the Bonferroni procedure, the FWER can be written as

$$\text{FWER} = \sup_{I \subseteq \{1,...,k\}} \sup_{F_\theta \in H_I} P_{F_\theta}\big(\cup_{i \in I} \{\varphi_i = 1\}\big)$$

By the σ-subadditivity of probability measures

$$P_{F_\theta}\big(\cup_{i\in I}\{\varphi_i = 1\}\big) \leq \sum_{i\in I} P_{F_\theta}(\varphi_i = 1) \tag{3}$$

and so

$$\text{FWER} \leq \sup_{I\subseteq\{1,...,k\}} \sup_{F_\theta\in H_I} \sum_{i\in I} P_{F_\theta}(\varphi_i = 1) \leq \sup_{I\subseteq\{1,...,k\}} \sum_{i\in I} \sup_{F_\theta\in H_I} P_{F_\theta}(\varphi_i = 1)$$

Further,

$$\sup_{F_\theta\in H_I} P_{F_\theta}(\varphi_i = 1) \leq \sup_{F_\theta\in H_i} P_{F_\theta}(\varphi_i = 1) \leq \alpha_i$$

where the first inequality holds because $H_I \subseteq H_i$ and the second inequality is the definition of φ_i as level α_i test. Therefore,

$$\text{FWER} \leq \sup_{I\subseteq\{1,...,k\}} \sum_{i\in I} \alpha_i = \sum_{i=1}^{k} \alpha_i$$

□

A common choice is $\alpha_i = \alpha/k, \forall i = 1, ..., k$. In most circumstances and in absence of further explanations, the term Bonferroni test is understood to imply this choice of equal local levels. For distinction, the term unweighted Bonferroni test may be used when $\alpha_i = \alpha/k, \forall i = 1, ..., k$ and the terms weighted Bonferroni test or Bonferroni-type may be used to indicate arbitrarily chosen local levels $\alpha_i \in [0, \alpha],$ s.t. $\sum_{i=1}^{k} \alpha_i \leq \alpha$.

In the context of multiple hypothesis testing, (3) is often referred to as Bonferroni inequality, after the Italian mathematician Emilio Bonferroni, giving the test its name.

The Bonferroni test requires no information on the joint distribution of the test statistics or, more generally, on the stochastic dependence between the individual tests. In fact, FWER control is guaranteed for any dependence structure. This makes the procedure applicable to basically all multiple testing problems. On the other hand, as a consequence of its generality the Bonferroni procedure will for most actual problems be conservative, i.e. the FWER of the procedure will be strictly below α. This is illustrated by an, admittedly extreme, example.

Example 4. Assume that k true null hypotheses $H_i, i = 1, ..., k$ are tested by respective tests $\varphi_i, i = 1, ..., n$ at local levels $\alpha_i = \alpha/k, \forall i = 1, ..., k$. However, the dependence of the underlying data is such that tests are perfectly positively correlated when performed at the same local level, that means $\varphi_1 = ... = \varphi_k$. Then with probability α/k, all H_i are rejected and with probability $1 - \alpha/k$ no hypothesis is rejected. The FWER is $\alpha/k < \alpha$.

Since we want to avoid a type I error, a FWER below α is no problem in itself. However, for a reasonably constructed local test φ_i the power will be an increasing function of the nominal significance level α_i. Thus, an unnecessarily small α_i will reduce the power to detect a real effect.

If the the joint distribution of multiple test statistics is known, this information can be used in constructing more powerful multiple testing procedures with FWER control. Even if the type of the joint distribution of the test statistics is known, though, its full specification typically involves additional nuisance parameters, compare example 2. These can be dealt with in principle by the same methods as discussed for the single nuisance parameter in section 1.5. However, asymptotic methods will require large enough sample sizes and the maximization approach may be computationally demanding even when the number of multiple hypotheses is small, as discussed in section 1.5.5. Methods of conditional inference are generally applicable if sufficient statistics for the nuisance parameters can be found and the conditional distribution of the data can be derived, which can be restrictive, too.

1.6.4 Closed test

The closed testing approach is a general way to construct and formalize multiple testing procedures with FWER control [34]. As before, let $\mathcal{H} = \{H_1, ..., H_k\}$ be a family of null hypotheses and denote an intersection hypothesis by $H_I = \cap_{i \in I} H_i$ for some $I \subseteq \{1, ..., k\}$. Now let $\mathcal{H}' = \{H_I : I \subseteq \{1, ..., k\}\}$ be the set of all intersection hypothesis that can be defined for $\mathcal{H}$. (The term closed test is motivated from $\mathcal{H}'$ being closed under intersection.) In a closed testing procedure, for each $H_I \in \mathcal{H}'$ a local level α test φ_I is specified. The closed testing procedure then rejects a hypothesis $H_I \in \mathcal{H}'$ if and only if $\varphi_J = 1$ for all $J \supseteq I$. Thus, the closed testing procedure rejects an elementary (or intersection) null hypothesis at the global level α if all intersection hypotheses it is contained in can be rejected locally at level α. Figure 2 illustrates the decision in a closed testing procedure for three hypotheses.

The closed testing procedure controls the FWER, because at some point the intersection of all true null hypotheses is tested with a local level α test. Only if this test becomes significant locally, any true null hypothesis can be rejected. But this happens with probability less or equal α by definition of the local test.

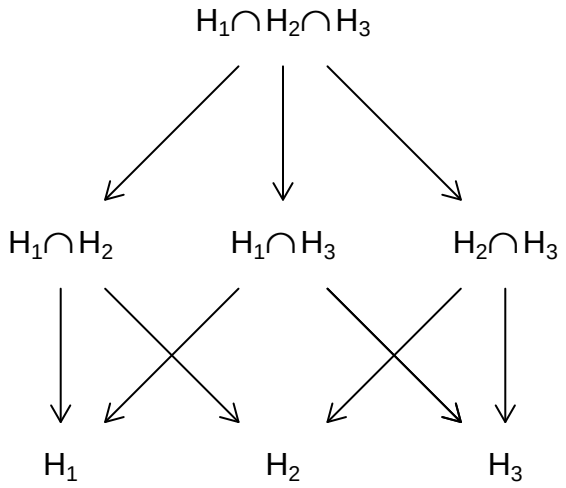

Figure 2: Scheme of a closed testing procedure for three hypotheses. An (intersection) hypothesis is rejected by the closed test, if all intersection hypotheses it is contained in are rejected by local level α tests.

As an important example, Holm's improvement of the Bonferroni test [35] can be derived as closed test. In Holm's procedure, a local test for each hypothesis $H_i \in \mathcal{H}$ is performed and the corresponding p-value p_i is calculated. Let $p_{(1)} \leq ... \leq p_{(k)}$ denote the ordered p-values and $H_{(1)}, ..., H_{(k)}$ the matching null hypotheses. A hypothesis $H_{(i)}$ is rejected by the Bonferroni-Holm procedure if for all $j \leq i$ it holds that $p_{(j)} \leq \alpha/(k-j+1)$. The same procedure results, if in the closed test each $H_I \in \mathcal{H}'$ is tested locally by an unweighted Bonferroni test. This means, H_I is rejected locally if for any $i \in I$, $p_i \leq \alpha/|I|$, where the cardinality $|I|$ is the number of hypothesis subsumed in H_I. Note that also the original Bonferroni test can be written as closed test, in which an H_I is rejected locally if for any $i \in I$, $p_i \leq \alpha/k$. Clearly, Holm's procedure is valid whenever the Bonferroni test ist. But the rejection region of the Bonferroni procedure is contained in that of Holm's procedure and so Holm's test is uniformly not less powerful than the Bonferroni procedure.

Sonnemann [36] showed that all coherent multiple testing procedures with FWER control can be written as closed tests. A test is coherent if non-rejection of an intersection hypothesis H_I implies that no hypothesis that is a subset of H_I is rejected [37]. Furthermore, Sonnemann and Finner [38] showed that for each non-coherent multiple testing

procedure with FWER control there is a non-inferior coherent procedure with FWER control. Consequently, it is sufficient to treat multiple testing procedures in terms of the closed test principle [39]. (However, for some problems application of the partitioning principle may provide more powerful procedures than a formal application of the closed test principle [40].)

An important property in the context of closed multiple testing procedures is consonance [37].

Definition 4. (Consonance) A test multiple testing procedure is said to be consonant, if whenever an intersection hypothesis H_I is rejected by the procedure at least one elementary hypothesis H_i such that $i \in I$ is rejected, too.

We will later discuss the role of consonance in the context of constructing optimal exact tests for multiple binary endpoints.

It is often of interest to calculate multiplicity adjusted p-values for a multiple testing procedure. For a closed test, this can be done by adopting a slightly different definition for a p-value. There, a p-value is defined as the smallest possible level α for which the test rejects the null hypothesis. For a single hypothesis test, this definition is identical to the usual one given in section 1.4. Consequently, multiplicity adjusted p-values for intersection or elementary hypotheses H_I in a closed test are found as $p^*(I) = \max\{p(J) : I \subseteq J\}$, where $J \subseteq \{1, ..., k\}$ and $p(J)$ is the p-value of the local test for the intersection hypothesis H_J. The calculation of p-values will not be studied further in this work, though.

Further details on multiple testing procedures and their applications, in particular for clinical trials, can be found in the books by Dmitrienko, Tamhane and Bretz [41] or Bretz, Hothorn and Westfall [42].

1.7 Multiple testing adjustemtens for discrete tests

We have already seen that discrete tests are in general conservative. When applied to a set of discrete tests, multiplicity adjustments that do not take into account the inherent discreteness and conservatism of the individual tests are typically too strong, reducing the probability for a true positive finding under the alternative.

Approaches to account for discreteness when adjusting for multiplicity are found in the literature. The methods allow for direct decisions on elementary hypotheses. At the

same time they can be seen as tests for the global intersection hypothesis. The repeated application of such tests to subset intersection hypotheses in a closed testing procedure usually leads to improved power to reject elementary hypotheses, similar to the Holm improvement for the Bonferroni test. As this modification will be obvious, we will not go into further detail on this possibility. Note though that it has been covered in the cited references.

1.7.1 Tarone's test and improvements

We will first discuss improvements of the Bonferroni test for the application to a set of discrete tests for the k hypotheses $H_1, ..., H_k$. The key aspect that is utilized in all these approaches is the fact that a discrete test with a finite set of possible values for the test statistic has a distinct most extreme event. The probability under the null hypothesis for this event is the smallest possible p-value for this test or, equivalently, the largest possible nominal significance level for which the test can become significant at all. Denote this quantity with α_i^* for the test of H_i. If for some hypothesis test $\alpha_i^* > \alpha$ it is impossible to reject the according null hypothesis with this test. Thus, as a first improvement of the Bonferroni test due to Gart and others, it is sufficient to apply the local level α/b to each test, where b is the number of tests for which $\alpha_i^* \leq \alpha$ [43].

The idea was refined by Tarone [44], basically aiming for the correction factor b being as small as possible by taking into account that tests with $\alpha_i^* > \alpha/b$ do not contribute to the FWER. To perform Tarone's test at a global level α, define $m(j) = |\{i \in \{1, ..., k\} : \alpha_i^* \leq \alpha/j\}|$, for $j = 1, ..., k$. Then find $b = \min\{j \in \{1, ..., k\} : m(j) \leq j\}$, and reject all H_i for which the local p-value is $p_i \leq \alpha/b$.

Hommel and Krummenauer [45] and, independently, Roth [46] noted that the decisions of Tarone's procedure are not monotone in the choice of the global level α. That means, it is possible that a hypothesis cannot be rejected when the procedure is performed at some level α, while with the same data the hypothesis could be rejected if the test was performed at some smaller, seemingly more stringent nominal level. Monotonicity of the test decision in α is often called alpha-consistency and is formally defined below. This property will be discussed further in section 3.4.

Definition 5. (Alpha-consistency) In refinement of the definition in section 1.4, let

$\varphi : \mathcal{X} \times [0,1] \to \{0,1\}$ be a hypothesis test as (measurable) function of the data $X \in \mathcal{X}$ and the nominal significance level $\alpha \in [0,1]$. The test is said to be alpha-consistent if for all $X \in \mathcal{X}$ and for all $\alpha \in [0,1]$, $\varphi(X,\alpha) = 1$ implies that $\varphi(X,\alpha') = 1$ for all $\alpha' > \alpha$.

Both, Hommel and Krummenauer, and Roth showed that a multiple testing procedure that has FWER control, that is alpha-consistent and that is more powerful than Tarone's test is obtained, if the rejection of a null hypothesis H_i is decided if there is at least one level $\alpha' \leq \alpha$, such that Tarone's procedure rejects H_i when it is performed at that level α' [45, 46]. Both publications contain algorithms to actually calculate such a test without considering all possible values $\alpha' \in [0,\alpha]$. The algorithm due to Hommel and Krummenauer appears easier to implement and works as follows: Find an integer $r, 1 \leq r \leq k$, such that either $r < n$ AND $(r-1)\alpha_r^* \leq \alpha < r\alpha_{r+1}^*$, or $r = n$ AND $(n-1)\alpha_k^* \leq \alpha \leq 1$. Obviously, there is always a unique such r. Then reject $H_i, i = 1, ..., k$ if and only if $p_i \leq \alpha/r$ OR $p_i < \alpha_r^*$ [45]. The abbreviation HKT for the Hommel and Krummenauer improvement of Tarone's test will be used in later sections.

Type I error rate control of all these procedures holds by the Bonferroni inequality. As such they can be seen as special cases of weighted Bonferroni tests, with information being required only on the smallest achievable p-value of each test and the number of tests. We will later discuss a simple though general approach to find in some sense optimal Bonferroni-type tests, which use information from the complete discrete distributions.

1.7.2 The minP permutation test

Further improvements can be achieved by studying the joint distribution of the multiple test statistics. Typically, this distribution is unknown. However, in the important case of between-group comparisons, an exact joint distribution of statistics can be found by permutation of the group labels. The resulting distribution is conditional on the observed data. Further, it is the null distribution first of all under the null hypothesis of exchangeability of observations between groups. See section 4.1 and [47–50] for further discussion. The permutation approach is used in the well known minP test due to Westfall and Young [51], that is also applicable to discrete tests [52].

The test statistic of the minP test for an intersection hypothesis $H_I, I \subseteq \{1, ..., k\}$, is the minimum of the marginal p-values $p_{I,min} = \min_{i \in I} p_i$. Let $\{\tilde{x}^{(1)}, ..., \tilde{x}^{(r)}\}$ be the

set of all unique copies of the observed data set x with permuted group labels. Let $\tilde{p}_{I,min}(\tilde{x}^{(j)})$ be the minimum p-value calculated for the permuted data set $\tilde{x}^{(j)}$. Then the reference null distribution for the test statistic $p_{I,min}$ is found as the empirical distribution of $\{\tilde{p}_{I,min}(\tilde{x}^{(1)}), ..., \tilde{p}_{I,min}(\tilde{x}^{(r)})$. The p-value for the minP test is $p = |\{\tilde{p}_{I,min}(\tilde{x}^{(j)}), j = 1, ..., r : \tilde{p}_{I,min}(\tilde{x}^{(j)}) \leq p_{I,min}\}|/r$. If the number of all possible permutations is too large to calculate $\tilde{p}_{I,min}$ for all of them, the procedure can be performed for a smaller, though still large enough, number of random permutations.

As a further refinement, a combination of the approach due to Tarone and the minP test was suggested in [53] in that the minP test is applied to the set of tests that can become significant at the adjusted level in Tarone's procedure.

1.7.3 Discrete rejection regions and integer programming

In the multiple testing procedures discussed so far, i.e. the Bonferroni-type tests and the minP test, the local test decision on each intersection hypothesis is determined by the largest marginal test statistic or the smallest marginal p-value of the involved tests. With respect to the joint distribution of the individual test statistics, this dependence on a maximum-type statistic results in a rectangular cone shaped rejection region. As main result of this work, rejection regions of almost arbitrary shape will be derived for multivariate test statistics with discrete joint (permutation) distributions. These rejection regions will be optimized against directed, marginally one-sided, alternatives to either maximize exhaustion of the type I error rate, maximize the power under a specific alternative point hypothesis or simply maximize the number of elements in the region. This will be achieved by methods of numeric optimization.

Some work has been done in this direction. Paroush [54] studied the general setting of testing a simple null hypotheses H_0 versus a simple alternative H_A using a univariate discrete test statistic (with a finite number of possible realizations). Let $V = \{v_1, ..., v_m\}$ be the set of values the test statistic T can take. Let the rejection region $R \subseteq V$ be defined through a vector $x \in \{0, 1\}^m$ such that $x_i = 1$ means $t_i \in R$ and $x_i = 0$ means $t_i \notin R$. Then, as Paroush noted, the optimal rejection region for a test with type I error rate control at level α is found as solution to

$$\sum_{i=1}^{m} x_i P_{H_A}(v_i) \to \max, \text{s.t. } \sum_{i=1}^{m} x_i P_{H_0}(v_i) \leq \alpha \tag{4}$$

Gutman and Hochberg [53] extended the approach to the case of testing multiple hypotheses against composite alternatives in the same way as is used in this work and that is described in detail in section 3.1. Basically, the vector of marginal test statistics $(T_i : i \in J)$ is used as multivariate test statistic for the intersection null hypothesis H_J and an according multivariate discrete rejection region is defined. In the absence of a model with known joint distribution under the alternative, Gutman and Hochberg defined a closed testing procedure, in which the rejection region for each intersection hypothesis test is found by optimizing the exhaustion of the type I error rate. This means for each local test to solve a problem of the type

$$\sum_{i=1}^{m} x_i P_{H_0}(v_i) \to \max, \text{s.t. } \sum_{i=1}^{m} x_i P_{H_0}(v_i) \leq \alpha \tag{5}$$

where, in contrast to (4), the v_i are multidimensional.

Their work included a simulation study for the comparison of multiple binary endpoints between two groups. The closed test with optimal exhaustion of the local level was applied to test the family of null hypotheses $\{H_i : OR_i = 1, i = 1, ..., k\}$ against two-sided alternatives, using the joint distribution of the test statistics from marginal Fisher's exact tests. The paper is scarce on details of how the joint distribution of the test statistics was found in the case of non-independent tests, but it can be assumed that a permutation approach was used similar to the one described in this work. In the simulation, the procedure showed low power for the rejection of elementary hypotheses, though, when compared to other multiple testing procedures. The authors attributed this to the lack of consonance of the procedure. However, the actual cause of the weak performance is a more serious flaw, in that there is no guarantee that the regions resulting from (5) are in some sense contiguous sets and as a consequence the test decisions are not necessarily monotone in a suitable norm of the test statistic. This can be illustrated even in the univariate case, see Example 5.

Example 5. Consider the same setting as in Example 1 where the test statistic of Fisher's exact test has values in $\{5, ..., 15\}$. There are four possible rejection regions that maximize the exhaustion of the nominal level α subject to $\alpha \leq 0.025$, these are $\{6, 7\}$, $\{6, 13\}$, $\{7, 14\}$ and $\{13, 14\}$. As we are interested in a one-sided alternative it may be reasonable to restrict the optimization to the right tail of the distribution and choose $R = \{13, 14\}$,

see Table 3. We now have a test with an actual conditional type I error rate virtually identical to the nominal level. (When rounding to four digits there is no notable difference, but the actual level is closely below 0.025.) There are configurations under the alternative, for which $R = \{13, 14\}$ is the best choice in terms of power. However, the case $T = 15$, which is the most extreme event and which would provide the strongest evidence against the null hypothesis, is not included in the rejection region. From Lemma 1 it is clear that with an increasing odds ratio the probability for this event is monotonically increasing and converges to 1. Consequently, starting from a large enough odds ratio, the power of our seemingly well chosen test decreases when the effect size increases.

Table 3: Null-distribution of T in the example. Probabilities are rounded to four decimal places. $\mathbb{1}_{\{t \in R(0.025)\}}$ indicates if t is in the level $\alpha = 0.025$ rejection region defined by a naive optimal alpha exhaustion.

t	$P_{H_0'}(T = t)$	$\mathbb{1}_{\{t \in R(0.025)\}}$
5	0.0001	0
6	0.0025	0
7	0.0225	0
8	0.0975	0
9	0.2274	0
10	0.3001	0
11	0.2274	0
12	0.0975	0
13	0.0225	1
14	0.0025	1
15	0.0001	0

Thus, non-monotonicity in the test statistic (or some suitable norm of it) can affect the power of an individual test. In a closed test, it may further cause the rejection regions of nested intersection hypotheses to share only few points. This reduces the power to reject elementary hypotheses and causes non-consonant decisions. As such, lack of consonance is as consequence of the way this test is constructed, rather than a problem in itself. Further, the interpretation of the test is problematic. There is no good reason not to

reject the null hypothesis in case of a large observed effect, while a smaller effect would have allowed for a rejection.

For these reasons, in the optimization framework of this work, an additional monotonicity constraint is included. As will be seen, the theoretical optimality of the resulting tests indeed translates to large power. It will further be shown how to even use this constraint to efficiently solve the optimization problem numerically with a branch and bound algorithm.

2 Aim and outline

The main aim of this work is to construct exact testing procedures for multiple hypotheses on binary endpoints that meet certain optimality criteria. Three main optimization criteria will be considered: Maximal exhaustion of the nominal type I error rate, maximizing the number of elements in the multivariate rejection region and maximizing the power of the resulting test for a specific alternative. To construct such optimal tests, the general setting of multiple discrete tests will be studied in section 3. The first approach will be the optimization of discrete multivariate rejection regions for tests with a known joint distribution of the test statistics. A second approach will be the optimal selection of critical thresholds in Bonferroni-type tests, which only requires knowledge of the exact marginal distributions of the test statistics. In both cases, the optimal rejection regions will be found through methods of numeric optimization. Multiple testing procedures will be constructed by virtue of the closed testing principle. Further, an algorithm providing an alpha-consistent multiple testing procedure for discrete tests will be derived. In section 4, the developed procedures will be applied to the setting of testing multiple binary endpoints with exact conditional tests. The unconditional power of the resulting testing procedures will be studied numerically for scenarios with two and three endpoints in section 5. The main work will be concluded with a discussion on the proposed methods. Selected R functions used in this work and exemplary analysis commands are contained in the appendix. Also, detailed results of the numeric power study are found in the appendix.

3 Optimal rejection regions for discrete tests

Throughout this section, consider a setting with k elementary hypotheses $H_i, i = 1, ..., k$. The test statistics of the corresponding marginal tests are $T_i, i = 1, ..., k$. The alternative for each H_i is one sided and it is assumed without loss of generality that for all $T_i, i = 1, ..., k$, larger values are in favor of the alternative. The global null hypothesis is $H_0 = \cap_{i=1}^{k} H_i$, and this is the null hypothesis we are primarily interested in.

3.1 Multivariate rejection regions for testing the global null hypothesis

Assume for now that the joint distribution of $(T_1, ..., T_k)$ under H_0 is known. (Details on finding a permutation joint distribution for the particular problem of multiple binary endpoints are given in section 4.1.) Let $V \subseteq \mathbb{N}^k$ be the set of possible values for $T = (T_1, ..., T_k)$. Our aim is to use the joint information from all elements of T to test $H_0 = \cap_{i=1}^{k} H_i$ versus the alternative of at least one H_i being false at a pre-specified level α. In the most general form of this approach, a multivariate rejection region $R \subseteq V$ is defined and H_0 is rejected if $T \in R$. No constraints due to any global test statistic are imposed on R. Only the following two conditions are required to define $R \subseteq V$ as a valid rejection region.

(i) $P_{H_0}(T \in R) \leq \alpha$

(ii) If $(t_1, ..., t_k) \in R$ then $\{(s_1, ..., s_k) \in V : s_1 \geq t_1, ..., s_k \geq t_k\} \subseteq R$

where P_{H_0} is the probability under the global null hypothesis.

Condition (i) establishes type I error rate control. Condition (ii) ensures that R is a contiguous set. That is, whenever a point $t = (t_1, ..., t_k)$ is contained in the rejection region, all points that are equally or more extreme with respect to the null hypothesis are contained in the rejection region too. This condition is necessary to avoid implausible regions that would lead to rejection of the null hypothesis for one vector of test statistics but not for another vector in which some elements are larger than in the first one, as was discussed in example 5. Without this condition, application of the test to actual problems is not recommended.

Among all regions R satisfying (i) and (ii) we search for a region that is optimal in some sense. As a first optimization criterion, the actual type I error rate is considered. Discrete tests are conservative, and often overly so, because the nominal type I error rate can in general not be exhausted. This problem can be alleviated to some extent by flexibly choosing the elements of the rejection region. The first optimization problem therefore is

$$R : P_{H_0}(T \in R) \rightarrow \max, \text{ s.t. conditions } (i) \text{ and } (ii) \tag{6}$$

However, better exhaustion of the nominal level does not necessarily translate to a great increase in power for relevant assumptions about the alternative. An alternative optimization criterion is the number of elements in the rejection region, $|R|$, arguing that increasing the volume of the region will increase the probability of the event $\{T \in R\}$ under the alternative (or under any equivalent probability measure). The optimization problem then is

$$R : |R| \rightarrow \max, \text{ s.t. conditions } (i) \text{ and } (ii) \tag{7}$$

Finally, for a specific simple alternative H_A, the rejection region is optimal if its probability mass under this specific alternative is maximal. Such a region may also be optimal for other alternatives representing stronger treatment effects in all endpoints than the specified one, however this is not easily verified. Finding the region with optimal power presumes knowledge of the joint distribution of T under H_A. The optimization problem is

$$R : P_{H_A}(T \in R) \rightarrow \max, \text{ s.t. conditions } (i) \text{ and } (ii) \tag{8}$$

3.1.1 Optimization using a branch and bound algorithm

The optimization problems (6)-(8) can be formulated as binary linear programs. For notation, let $x = (x_1, ..., x_m) \in \{0, 1\}^m$ be a solution vector with $m = |V|$ the number of points in the search space V. Each element x_i of x corresponds to one distinct point $t^{(i)} \in V$, with $x_i = 1$ indicating that $t^{(i)}$ is contained in the rejection region represented by the solution x, and $x_i = 0$ indicating that $t^{(i)}$ is not contained in the rejection region.

To achieve the type I error rate control, let $a_i = P_{H_0}(t^{(i)})$ be the contribution of $t^{(i)}$ to the type I error rate and let $a = (a_1, \ldots, a_m)^T$. Then condition (i) can be written as $a^T x \leq \alpha$. The objective function for any of the optimization problems (6)-(8) can be

written as $\sum_{i=1}^{m} w_i x_i$, where w_i denotes the contribution of $t^{(i)}$ to the objective when it is included in the rejection region. For the different objective functions (6) to (8), choose $w_i = a_i$ to optimize the type I error rate, $w_i = 1$ to optimize $|R|$, and $w_i = P_{H_A}\left(t^{(i)}\right)$ to optimize the power under H_A. Let $w = (w_1, \ldots, w_m)^T$. Optimization means maximizing $w^T x$ subject to the constraints imposed by conditions (i) and (ii).

We propose a branch and bound algorithm [55] to solve this type of problem. The algorithm takes into account the constraint due to condition (ii) and in fact makes efficient use of it. In the algorithm, a node is the set of a current solution vector x, and an upper bound and a lower bound of the objective function value resulting from this solution. In the current solution vectors, entries not yet determined (i.e. x_i's for which the algorithm has not yet decided whether they belong to the rejection or not) are assigned a value of -1.

1. Initiate the set S containing one node with the solution $x = (-1, -1, ..., -1)$, a lower bound of 0, an upper bound of 0 and initiate the current best lower bound for the objective to $L = 0$

2. Of all nodes in S with not fully determined solution, select the one with the largest lower bound, denote it with N. In the current solution x of N, identify the first entry that is -1. Denote its index with i.

3. Remove N from S and add two copies of N to S. In the first one set $x_i = 0$, in the second one set $x_i = 1$. Do steps 4 to 6 for each of the two new nodes:

4. Check if the chosen value for x_i requires any other entries of x that are -1 to take a specific value due to condition (ii). If so, set the required values in x.

5. Set the lower bound of this node to $\sum_{i:x_i=1} w_i$ and the upper bound to $\sum_{i:x_i \neq 0} w_i$

6. If $\sum_{i:x_i=1} P_{H_0}\left(t^{(i)}\right) > \alpha$ remove the node from S.

7. Update L as the maximum over the lower bounds of all nodes in S.

8. Remove from S all nodes with an upper bound $< L$.

9. Repeat steps 2 to 8 until S contains only nodes with fully determined solutions. These are optimal solutions.

Step 4 is the the only one not included in a standard branch and bound algorithm. It has a double function of controlling for condition (ii) and reducing the number of solutions that need to be considered. Computationally, this step is best implemented by preparing a look up matrix D of dimension $m \times m$. An element d_{ij} of this matrix is equal to 1, if $t^{(j)}$ is equal to or more extreme than $t^{(i)}$, i.e. $t_l^{(j)} \geq t_l^{(i)}$ for all $l = 1, \ldots, m$, and 0 otherwise. Then, in step 4, if $x_i = 1$, x_j is set to 1 for all $j = 1, ..., m$ that meet $d_{ij} = 1$. Similarly, if $x_i = 0$, x_j is set to 0 for all $j = 1, ..., m$ that meet $d_{ji} = 1$. An R code of the implementation of this algorithm is given in the appendix section 8.1.

3.1.2 Optimization using linear integer programming

The constraint due to condition (ii) can be formulated as $Cx \leq 0$. Here, C is a matrix with m columns, which captures the requirement that if $x_i = 1$ and $t^{(j)}$ is more extreme than $t^{(i)}$, then x_j must be equal to 1. Using that formulation the constrained optimization problem can be written as linear integer program. Numeric solutions can be found using standard LP solvers, e.g. lpsolve [56], which can also be accessed through R [57] using the package lpSolve [58].

A simple way to define C is to use $m(m-1)/2$ rows $c^{(i,j)}, i = 1, ..., m, j = 1, ..., m, i \neq j$. In this representation, $c_i^{(i,j)} = 1$ and $c_j^{(i,j)} = -1$ if $t^{(j)}$ is more extreme than $t^{(i)}$. All other entries in C are equal to 0. Thus ${c^{(i,j)}}^T x = 1 > 0$ if and only if $x_i = 1$ and $x_j = 0$, which violates condition (ii) if $t^{(j)}$ is more extreme than $t^{(i)}$. The following more efficient way to define a constraint matrix C was suggested by my collaboration partner Dong Xi (personal communication). It requires at most m rows. Denote by M_i the set of points that are equal to or more extreme than $t^{(i)}$, i.e., $M_i = \{t^{(j)} \in V : t_l^{(j)} \geq t_l^{(i)} \text{ for all } l = 1, \ldots, m\}$. Then a matrix C can be constructed with elements $c_{ii} = |M_i|$ and $c_{ij} = -1$ if $i \neq j$ and $t^{(j)} \in M_i$. With $c_{i.}$ the i-th row of C, $c_{i.}^T x > 0$ if and only if $t^{(i)}$ is part of the solution while there are points more extreme than $t^{(i)}$ that are not part of the solution, which is a violation of condition (ii). Rows corresponding to points for which there are no more extreme points, i.e. $|M_i| = 1$, can be removed from the constraint matrix.

Both conditions, $a^T x \leq \alpha$ and $Cx \leq 0$, are combined using a final constraint matrix $B = (a, C^T)^T$ and a vector $b = (\alpha, 0, ...0)$, such that the linear optimization problem reads

$$w^T x \to \max, \text{ s.t. } Bx \leq b, x \in \{0, 1\}^m \tag{9}$$

However, when using the lpSolve package in R for different numeric examples for optimizing rejection regions according to (9), occasionally numeric issues were observed that resulted in non-optimal solutions or extremely long run times. According to personal communication of Dong Xi with the maintainers of lpsolve, these issues may result from the involved probability values ranging across several orders of magnitude, hence proper scaling in the underlying simplex algorithm may be difficult. Therefore, for the numeric calculations presented in this work, the branch and bound algorithm was used.

3.1.3 Pre-processing

Simple considerations on whether a point in V can at all meet condition (i) or (ii) may be used to reduce the search space for the actual optimization algorithm in two pre-processing steps.

In the first pre-processing step, condition (i) is used to remove all points from the search space V that would inevitably lead to a rejection region with a probability mass greater than α. If a point $t = (t_1, ..., t_k) \in V$ was selected to be part of the rejection region, condition (ii) implies that the set $\{(s_1, ..., s_k) \in V : s_1 \geq t_1, ..., s_k \geq t_k\}$ is part of the rejection region. Thus the probability mass of a rejection region containing t is at least $P(\{(s_1, ..., s_k) \in V : s_1 \geq t_1, ..., s_k \geq t_k\})$. Therefore all points $t \in V$ for which $P(\{(s_1, ..., s_k) \in V : s_1 \geq t_1, ..., s_k \geq t_k\}) > \alpha$ are removed from the search space. Denote the set of the remaining points by $V^{(1)}$.

In the second pre-processing step we identify points that are definitely contained in an optimal level α rejection region, regardless of the optimality criterion. For each point $t \in V^{(1)}$, we calculate an upper bound for the probability mass under H_0 of all possible rejection regions that do not contain t. If $t \notin R$, the set $A(t) = \{(s_1, ..., s_k) \in V^{(1)} : s_1 \leq t_1, ..., s_k \leq t_k\} \notin R$, because otherwise condition (ii) would be violated. The upper bound for the level when point t is not included in the rejection region is $\alpha_{max,-t} = P_{H_0}(V \backslash A(t))$. If $\alpha_{max,-t} < \alpha$, even the largest rejection region not containing t could possibly be made larger, and the only way to do so is adding t, because of condition (ii). So if $\alpha_{max,-t} + P_{H_0}(t) \leq \alpha$, t must be included in each optimal rejection region and therefore can be removed from the search space. Denote the set of the points still remaining after this step by $V^{(2)}$. It is then sufficient to perform the optimization on the remaining search space $V^{(2)}$ for a significance level of $\alpha - P_{H_0}(V^{(1)} \backslash V^{(2)})$.

The two pre-processing steps are illustrated in Figure 3.

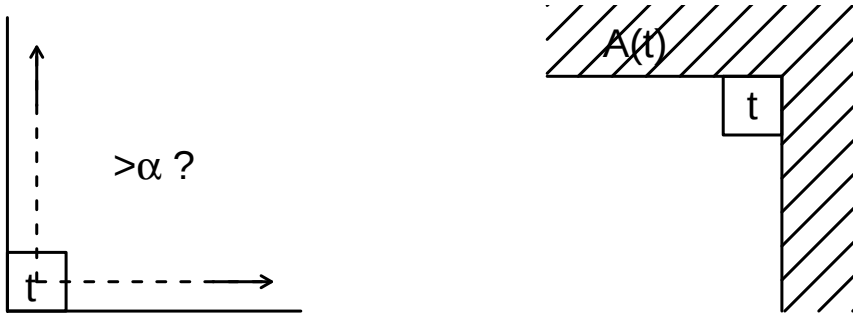

Figure 3: The two pre-processing steps. Left panel: Exclude points $t \in V$ for which $P(\{(s_1, ..., s_k) \in V : s_1 \geq t_1, ..., s_k \geq t_k\}) > \alpha$ from the search space as including these points will result in a level $> \alpha$. Right panel: For all points $t \in V$, get the largest region $A(t)$ which satisfies condition (ii) and does not include t. If $P_{H_0}(A(t)) + P_{H_0}(t) \leq \alpha$, t must be contained in any optimal rejection region.

3.1.4 Further computational aspects

Regardless of the optimization objective and the applied algorithm, there can be more than one optimal solution. This may happen, e.g., if the joint distribution is highly symmetric. In this case an a-priori order of the endpoints, based on clinical importance, may be invoked to define a unique solution that slightly favors the more important endpoint. In other cases it is possible to combine optimization criteria. When optimizing $|R|$, a way to reduce the number of solutions, in most cases to a unique one, is to select of all solutions with maximal $|R|$ that one with best alpha exhaustion or the largest power under a specified alternative. This is achieved computationally by using as contribution to the objective function $w_i = 1 + P_{H_0}(t^{(i)})$ or $w_i = 1 + P_{H_A}(t^{(i)})$, respectively. When optimizing exhaustion of the nominal level, adding a small constant to the contribution of each point as suggested in [53] can be used to select the largest region with the maximal achievable actual level.

When the search space V is large, there are typically many points with very small probability mass leading to a large number of close to optimal solutions and as a consequence to long computation times. A feasible close to optimal solution can be found with reduced computational effort by separate treatment of these points. First, a threshold c considerably smaller than the nominal level of significance is set, e.g. $c = 10^{-4}$.

Following the pre-processing, the largest possible set of points $C \subseteq V^{(2)}$, such that $\max_{t \in C} P_{H_0}(t) < \min_{t \in V^{(2)} \setminus C} P_{H_0}(t)$ and $P_{H_0}(C) \leq c$, is identified. The set C is removed from the search space for the optimization, and the subsequent optimization is performed on $V^{(2)} \setminus C$ for a significance level of $\alpha - P_{H_0}(V^{(1)} \setminus V^{(2)}) - P_{H_0}(C)$. After the rejection region in $V^{(2)} \setminus C$ has been found, all points in C can be added subject to condition (ii).

3.2 Optimal tests using marginal distributions

In some cases it may not be feasible to use a joint distribution of the test statistics, e.g. when it is difficult to derive an exact joint distribution or when testing composite null hypotheses that do not immediately imply a joint null distribution. In these cases, Bonferroni-type optimal tests can be derived using the exact marginal distribution for each null hypothesis.

To start, let V_i denote the set of values that the test statistic T_i for the i-th hypothesis test can take. It is assumed that these values are finite and their number is finite. Denote with $S_i(t) = P_{H_i}(T_i \geq t)$ the probability that the test statistic for the i-th endpoint is greater or equal some value t under the marginal null hypothesis H_i. For a Bonferroni-type test, critical boundaries $c_1, ..., c_k$ are defined such that $\sum_{i=1}^{k} S_i(c_i) \leq \alpha$. The test rejects H_i if $T_i \geq c_i$ with FWER control at level α. Recall that for an unweighted Bonferroni test, c_i is chosen such that $S_i(c_i) \leq \alpha/k$, which does not take into account the discreteness of the marginal distribution. However, within the class of weighted Bonferroni tests, the boundaries can be chosen according to several optimization criteria, by utilizing information from the marginal distributions, resulting in less conservative and more powerful procedures. Optimal exhaustion of the Bonferroni bound on the type I error rate is achieved by choosing $c = (c_1, ..., c_k)$ as a solution to

$$c : \sum_{i=1}^{k} S_i(c_i) \to \max, \text{ s.t. } \sum_{i=1}^{k} S_i(c_i) \leq \alpha \tag{10}$$

When aiming for a rejection region with maximal number of elements, a simple way is to solve

$$c : \prod_{i=1}^{k} |\{t \in V_i : t < c_i\}| \to \min, \text{ s.t. } \sum_{i=1}^{k} S_i(c_i) \leq \alpha \tag{11}$$

If it is known which combinations of the test statistic cannot be observed (i.e. if there is

information on the multivariate support of the joint distribution), the number of elements as a function of c is known exactly, and c can be chosen to maximize this number.

When marginal alternatives $H_i^{(1)}, i = 1, ..., k$, are specified, the solution to

$$c : 1 - \prod_{i=1}^{k} P_{H_i^{(1)}}(T_i < c_i) \rightarrow \max, \text{ s.t. } \sum_{i=1}^{k} S_i(c_i) \leq \alpha \tag{12}$$

results in a test with optimal power to reject at least one elementary hypothesis under the assumption of independent marginal tests. The optimization problems 11 and 12 can be equivalently stated using sums of logarithmic terms and it is straight forward to formulate them in terms of a linear program (see below). The search space for this problem is growing polynomially in sample size and in the number of endpoints. Further, only values for c_i above the $1 - \alpha$ quantile of the respective marginal distribution need to be considered. Therefore the search space is far smaller than for the optimization problem using joint distributions and it will often be possible to even find the solution by complete enumeration.

3.2.1 Formulation as linear program

When solving the optimization problem for multivariate rejection regions using the joint distribution of the test statistics, a solution vector was used that indicated which point belonged to the rejection region. Here a slightly different approach is used. We will in principle consider as search space for the critical boundaries a stacked vector $v^T = ({v^{(1)}}^T, ..., {v^{(k)}}^T)$ of the possible values the k test statistics can take. However, for each critical boundary a possible value of ∞ will be considered, too, to allow for an allocation of zero weight to the respective test. We will define a solution vector x of same length as v, that has exactly k entries equal to 1, and all other entries equal to 0. Entries with value 1 will indicate the positions in v at which we find the values for the thresholds $c_i, i = 1, ..., k$. As mentioned above we need to consider only values in V_i that can lead to a rejection at the full local level α. Denote the according set $V_i^{(1)} = \{t \in V_i : S_i(t) \leq \alpha\}$. Reducing the search space to $V_i^{(1)}, i = 1, ...k$ is similar to the first pre-processing step for the tests using joint distributions. Now $v^{(i)}$ is defined as the vector of the elements of $V_i^{(1)} \cup \infty$.

By this definition we allow for $c_i = \infty$, which would result in $S_i(c_i) = 0$. In this case zero weight is attributed to the corresponding test, and as such it is removed from the set

of tests between which the nominal significance level is distributed. Therefore, approaches as Tarone's test [44] are covered as a special case in this optimal framework. If the test cannot become significant even if the full level α would be allocated to it, ∞ is the only element of $v^{(i)}$, so this special case of removing a test is covered immediately, too.

The solution vector is $x \in \{0,1\}^{\sum_{i=1}^{k}(|v^{(i)}|)}$. The constraint

$$\begin{pmatrix} J_1 & 0 & \dots & 0 \\ 0 & J_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & J_k \end{pmatrix} x = \begin{pmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{pmatrix}, \text{ with } J_i = (1, \dots, 1)_{1 \times |v^{(i)}|} \tag{13}$$

ensures that x contains one entry equal to 1 for each search vector $v^{(i)}$. Further, let $a^{(i)}$ be the vector of contributions to the type I error rate with elements $a_j^{(i)} = S_i(v_j^{(i)})$. Let $a^T = ({a^{(1)}}^T, ..., {a^{(k)}}^T)$. Then the constraint

$$a^T x \leq \alpha \tag{14}$$

guarantees type I error rate control by the Bonferroni inequality.

The contributions of the k tests to the objective function that result from different choices for $c_i, i = 1, ...,$ are formalized similarly in terms of a stacked vector $w^T = ({w^{(1)}}^T, ..., {w^{(k)}}^T)$. Here $w_j^{(i)}$ is the contribution of the i-th test to the objective function for the particular choice $c_i = v_j^{(i)}$. The optimization problems (10)-(12) then have the common form $w^T x \to \max$, subject to constraints (13) and (14). For the optimization problem (10), $w = a$. For problem (11) taking the logarithm in (11) and changing signs leads to the equivalent problem of maximizing $\sum_{i=1}^{k} -\log|\{t \in V_i : t < c_i\}|$ subject to the constraints. Therefore, $w_j^{(i)} = -\log|\{t \in V_i : t < c_i\}|$. The optimization problem (12) is identical to minimizing $\prod_{i=1}^{k} P_{H_i^{(1)}}(T_i < c_i)$ subject to the constraints. Again taking the logarithm and changing signs results in $w_j^{(i)} = -\log\left(P_{H_i^{(1)}}(T < v_j^{(i)})\right)$.

Thus, all problems considered here to optimize discrete Bonferroni-type tests can be formulated as linear integer programs with the further constraint that the elements of x are in $\{0,1\}$ and solved by existing software. No problems were encountered when using the lpSolve package in R for these problems and the package was used for the actual calculations in sections 4 and 5.

3.3 Closed testing procedure and consonance

So far we have considered tests for the global null hypothesis $H_0 = \cap_{i=1}^{k} H_i$. In the small sample setting this may indeed be the most important hypothesis, because the power to support more detailed claims may be inherently small. Nonetheless, in many applications the question for a treatment effect in certain individual endpoints needs to be addressed. In this case a multiple testing procedure with FWER control is required.

Such a procedure can be derived by constructing optimal local level α tests for all intersection null hypotheses $H_I = \cap_{i \in I} H_i, I \subseteq \{1, ..., k\}$ using the same optimzation approach and then applying the closed testing principle as outlined in section 1.6.4.

Note, however, that even if each of the intersection hypothesis tests satisfies the selected optimality criterion, this does not imply that some optimality holds for the overall closed testing procedure. In principle, the rejection regions of local tests may be optimized simultaneously by considering a stacked solution vector x of all local tests. An objective function defined for the set of all local rejection regions may then be used in a branch and bound algorithm and also further constraints, e.g. demanding consonance, may be included when checking if a current solution is feasible. The definition of appropriate objective functions and formulation of computationally efficient algorithms for this type of problem may be a matter of future research and is not included in this work.

The particular concern of Gutman and Hochberg [53] that closed tests with locally optimal discrete rejection regions need not be consonant was discussed in section 1.7.3. They attributed to the lack of consonance the low power to reject elementary hypotheses, which was observed for these procedures in their simulation study. However, the actual flaw in their testing procedure was rather the lack of a monotonicity constraint like condition (ii) of this work.

Nonetheless, a consonant test may indeed provide larger power to reject elementary hypotheses than a non-consonant one. In the special case of only two endpoints, a consonant optimal closed test can be constructed by excluding from the search space for the optimal rejection region for $H_1 \cap H_2$ all points that are not included in the union of the elementary test rejection regions for H_1 and H_2. With more endpoints, the rejection regions of local tests may be optimized simultaneously, with an additional constraint ruling out all solutions that lead to a non-consonant procedure. This constraint is not linear in

the solution vector. However, it can be included in the branch and bound algorithm when checking if a current solution is feasible.

For the weighted Bonferroni tests, consonance is met if, starting from the global intersection hypothesis test, the critical boundaries for each endpoint are non-increasing [59]. Formally, for all $J' \subset J \subseteq \{1, ..., k\}$ and $i \in J'$, $c_i^{(H_{J'})} \leq c_i^{(H_J)}$ needs to hold, where $c_i^{(H_J)}$ is the critical boundary for the test statistic of the i-th endpoint in the test for H_J. This additional constraint can be easily implemented when optimizing the critical boundaries for the marginal tests, simply by reducing the search space accordingly, and it is in fact advantageous to include it. It does not affect the power of the global test. As a consequence, the power to reject at least one elementary hypothesis cannot be decreased by including the consonance constraint. In addition, the search space for the local critical boundaries is reduced. The consonance constrained is therefore used in all actual calculations of weighted Bonferroni tests in sections 4 and 5.

Note that for the optimal tests using joint distributions, the consonance constraint may reduce the power for rejecting the global null hypothesis, at the same time this requirement increases the computational burden. It is therefore not clear from the onset whether consonance is an advantage with these procedures. Both variants are included in the numeric calculations in sections 4 and 5.

3.4 Alpha-consistency and a greedy algorithm

The optimal rejection regions defined so far depend on the pre-specified level α. Without further constraints the rejection regions corresponding to increasing values of α are not necessarily nested subsets and so the tests are not alpha-consistent. Numeric examples showing the lack of alpha-consistency can be easily found, one is given below.

Example 6. Consider a bivariate test statistic with possible values in $\{(0,0),(0,1),(1,0),(1,1)\}$ and probabilities for these values (in the same order) under the null hypothesis of 0.94, 0.02, 0.01, 0.03. At nominal level $\alpha = 0.05$, the rejection region which best exhausts α is $(0,1),(1,1)$, so for an observed realization of $(1,0)$ the null hypothesis is not rejected. However, for a lower level $\alpha = 0.04$, the rejection region with best α exhaustion is $(1,0),(1,1)$, leading to rejection if $(1,0)$ is observed.

Also, examples can be found to show that the optimally weighted Bonferroni tests are

not necessarily alpha-consistent (in contrast to the unweighted Bonferroni test).

In general, when a test is not alpha-consistent it may happen that a test statistic is observed for which the null hypothesis cannot be rejected at the pre-specified level α, while the researcher may identify a smaller level such that the corresponding rejection would allow for rejection of the null hypothesis. This situation seems unsatisfactory. However, it should be kept in mind, that both the nominal level of the test and the optimization criterion are chosen before the start of the trial, knowing that a different rejection region may be optimal as the nominal level of the test is decreased. This non-consistency neither affects the type I error rate controlling property of the test, nor does it invalidate its optimality under the chosen criterion.

One drawback of tests lacking alpha-consistency is that it is not possible to define multiplicity adjusted p-values as defined in section 1.6.4.

An optimal rejection region for a test for H_0 with alpha-consistency can be found through the following argument. Let $R(\alpha)$ denote a rejection region for the alpha-consistent test at level α. Alpha-consistency demands

$$\alpha' > \alpha \Rightarrow R(\alpha') \supseteq R(\alpha) \tag{15}$$

This must hold for all choices of $\alpha', \alpha \in [0, 1]$, especially for $\alpha' = \alpha + \delta(\alpha)$ where $\delta(\alpha)$ is the smallest possible increment in the actual level that can occur by adding a point in the discrete space to $R(\alpha)$, i.e. $\delta(\alpha) = \min\{P_{H_0}(t^{(i)}) : t^{(i)} \notin R(\alpha) \text{ and } R(\alpha) \cup t^{(i)} \text{ meets } (ii)\}$. (Only points $t^{(i)}$ in the actual search space are considered, which could be reduced by some further constraints, e.g. to impose consonance.) Therefore, for this choice of α', any $R(\alpha') \supseteq R(\alpha)$ must be $R(\alpha') = R(\alpha)$ or $R(\alpha') = R(\alpha) \cup t^{(i)}$ with i such that $P_{H_0}(t^{(i)}) = \delta(\alpha)$. Otherwise there would be a contradiction to (15).

If the test is to be optimal regardless of the choice of the nominal significance level, $R(\alpha') = R(\alpha) \cup t^{(i)}$ must be chosen at every instance. In contrast, selecting $R(\alpha') = R(\alpha)$ at some particular level α' would mean not to increase the rejection region at some point in the algorithm and in turn have more points to choose from at a later step. But since the idea of alpha-consistency is to construct a test without pre-specification of the level of significance there is no obvious way how to decide at which point we should abstain from increasing R.

If $\delta(\alpha)$ is not unique, some pre-specified selection criterion should be applied. If

the distribution of T under a specific alternative is known, the point with the largest contribution to power may be selected. Otherwise, an a-priori order of the endpoints may be invoked, and the point which facilitates rejection of the highest valued endpoint can be chosen. As starting condition select the point $t^{(i)}$ meeting $P_{H_0}(t^{(i)}) = \delta(0)$. The basic algorithm to find an alpha-consistent optimal rejection region is given below. It is a greedy algorithm, which is fast to compute.

Greedy algorithm for discrete multivariate rejection regions

1. Set $j = 1, \alpha^{(j)} = 0$ and $R^{(j)} = R(0) = \{\}$
2. Calculate $\alpha^{(j+1)} = \alpha^{(j)} + \delta(\alpha^{(j)})$ and identify a unique $t^{(j)}$ such that $P_{H_0}(t^{(j)}) = \delta(\alpha^{(j)})$
3. If $\alpha^{(j+1)}$ exceeds the pre-specified nominal level, stop, else set $R^{(j+1)} = R^{(j)} \cup t^{(j)}$
4. Repeat steps 2 and 3 until the stopping criterion is met. The rejection region is $R = R^{(j)}$.

Besides resulting in an alpha-consistent test, the greedy algorithm has no guarantee to meet any of the optimality criteria (6)-(8), however it can be seen as a heuristic attempt to maximally exhaust the type I error rate.

The above discussion on alpha-consistency applies euqally to the optimal Bonferroni-type tests. Also, a greedy algorithm with the same principle can be applied to the Bonferroni-type tests. There, in each iteration one element is added to the rejection region of that marginal test, for which the resulting increase in the local level is smallest.

Greedy algorithm for discrete Bonferroni-type tests

1. Set $j = 1, \alpha^{(j)} = 0$ and the marginal rejection regions $R_i^{(j)} = \{\}, i = 1, ..., k$. Denote with the complement ${R_i^{(j)}}^C$ the set of values that T_i can take and that are not yet contained in the rejection region.
2. Find eligible thresholds $e = \left(\max\left({R_1^{(j)}}^C\right), ..., \max\left({R_k^{(j)}}^C\right)\right)$
3. Calculate $\delta' = \min\left(P_{H_1}(e_1), ..., P_{H_k}(e_k)\right)$
4. Set $\alpha^{(j+1)} = \alpha^{(j)} + \delta'$ and find an index $l \in (1, ..., k)$, such that $P_{H_l}(e_l) = \delta'$

5. If $\alpha^{(j+1)}$ exceeds the pre-specified nominal level, stop, else set $R_l^{(j+1)} = R_l^{(j)} \cup e_l$ and $R_i^{(j+1)} = R_i^{(j)}$ for all $i \neq l$.

6. Repeat steps 2 to 5 until the stopping criterion is met. The marginal rejection regions are $R_i = R_i^{(j)}, i = 1, ..., k$.

When applied to each intersection hypothesis in a closed testing scheme, the greedy algorithm for Bonferroni-type tests by construction provides non-increasing critical boundaries for each endpoint (compare section 3.3) and therefore leads to a consonant closed test.

From the algorithms it is obvious that demanding alpha-consistency greatly reduces the possibilities to optimize the test in terms of power and that it leaves no scope on the number of elements in R. In an actual application with a pre-specified level of significance, lack of alpha-consistency may be accepted in exchange for a larger power. However, the numeric power calculations presented in section 5 show that in many cases the tests based on the greedy algorithms may be a good choice in terms of power.

4 Application to multiple binary endpoints

The theory of section 3 will now be applied to find optimal exact tests for the comparison of two groups with respect to k binary endpoints.

Consider a treatment (Trt) and control (Ctr) group with n_g independent subjects in group $g \in \{Trt, Ctr\}$. For each binary endpoint we are interested in the null hypothesis $H_i : p_i^{(Ttr)} - p_i^{(Ctr)} \leq 0, i = 1, ..., k$. Here, $p_{i,g}$ is the marginal success rate in the i-th endpoint in group g. Equivalently, the null hypothesis can be stated in terms of marginal odds ratios $OR_i = (p_i^{(Ttr)}/(1 - p_i^{(Ttr)}))/(p_i^{(Ctr)}/(1 - p_i^{(Ctr)}))$ as $H_i : OR_i \leq 1$. Recall that we are interested in one sided alternatives $p_i^{(Ttr)} > p_i^{(Ctr)}$ or $OR_i > 1$, because the aim in a clinical trial typically is to establish higher efficacy of a new treatment compared to placebo or to a reference treatment.

Fisher's exact test [9, 13] provides condititional exact inference about an individual H_i (see section 1.5.2). Calculating optimal Bonferroni-type tests as described in section 3.2 for multiple Fisher's exact tests is straight forward using the known marginal hypergeometric distribution of the individual test statistics $T_i, i = 1, ..., k$.

The vector of test statistics resulting from k marginal Fisher's exact tests $T = (T_1, ..., T_k)$ will be used as multivariate test statistic to perform tests according to the methods described in section 3.1. The conditional joint distribution of these test statistics will be derived and related to the permutation distribution from the observed data in the following section.

4.1 Conditional joint distribution of the test statistic

Regarding the joint observations in one patient with respect to all k endpoints, there are $d = 2^k$ possible outcome categories. We can formally define these categories by a set of index vectors $\mathcal{S} = \{s_1...s_k : s_i \in \{0, 1\}, i = 1, ..., k\}$ such that $s_i = 1$ if the particular patient had a success in endpoint i. E.g. with $k = 2$ endpoints, the set would be $\mathcal{S} = \{11, 10, 01, 00\}$ indicating a success in both endpoints, endpoint 1 only, endpoint 2 only or neither endpoint. To simplify notation, these d different outcome categories will be indexed by $s = 1, ..., d$ in the following equations whenever possible.

The observation on the j-th patient, $j = 1, ..., n_g$, in group $g \in \{Trt, Ctr\}$ can be written as vector $Y^{(j,g)} \in \{z \in \{0, 1\}^d : \sum_{s=1}^d z_s = 1\}$. The single non-zero entry $Y_s^{(j,g)} = 1$

indicates that the patient is in the s-th outcome category. Let $q_s^{(g)} = P(Y_s^{(j,g)} = 1)$ be the probability for this event. The distribution of $Y^{(j,g)}$ is characterized by the vector $q^{(g)} = (q_1^{(g)}, ..., q_d^{(g)})$ with $0 < q_s^{(g)} < 1$ and $\sum_{s=1}^{d} q_s^{(g)} = 1$. The observations from different patients are assumed to be independent.

The data resulting from this model can be aggregated without loss of information in a $d \times 2$ contingency table with columns $Y^{(Trt)} = \sum_{j=1}^{n_{Trt}} Y^{(j,Trt)}$ and $Y^{(Ctr)} = \sum_{j=1}^{n_{Ctr}} Y^{(j,Ctr)}$ (see Table 4 in section 4.3 for an example), and row margins $M = Y^{(Trt)} + Y^{(Ctr)}$. The vector of observed row margins is denoted by $\tilde{m}$.

As for marginal Fisher's exact tests, define $T_i = \sum_{s_1...s_k:s_i=1} Y_{s_1...s_k}^{(Trt)}$ as the number of subjects in the treatment group with a sucess in endpoint i, and let $T = (T_1, ..., T_k)$. Thus, T is a linear function h of $Y^{(Trt)}$. The distribution of T conditional on $M = \tilde{m}$ will be used for exact inference about $H_0 = \cap_{i=1}^{k} H_i$. This distribution is a function of the conditional distribution of $Y^{(Trt)}$ given $M = \tilde{m}$

$$P(T_1 = t_1, ..., T_k = t_k | M = \tilde{m}) = \sum_{y^{(Trt)} \in W : h(y^{(Trt)}) = t} P(Y^{(Trt)} = y^{(Trt)} | M = \tilde{m}) \qquad (16)$$

where $W = \{y \in \mathbb{N}^d : \sum_{s=1}^{d} y_s = n_{Trt} \text{ and } \exists z \in \mathbb{N}^d : y + z = \tilde{m}\}$ is the set values of $y^{(Trt)}$ that are possible given the table margins. The conditional distribution of $Y^{(Trt)}$ given $M = \tilde{m}$ is found as multivariate (non-central) hypergeometric distribution, using that $Y^{(Trt)}$ and $Y^{(Ctr)}$ are independent and each is multinomially distributed with respective parameters $(n_g, q^{(g)}), g \in \{Trt, Ctr\}$.

$$P(Y^{(Trt)} = y^{(Trt)} | Y^{(Trt)} + Y^{(Ctr)} = \tilde{m}) = \frac{P(Y^{(Trt)} = y^{(Trt)}, Y^{(Trt)} + Y^{(Ctr)} = \tilde{m})}{P(Y^{(Trt)} + Y^{(Ctr)} = \tilde{m})} =$$

$$= \frac{P(Y^{(Trt)} = y^{(Trt)}, Y^{(Ctr)} = \tilde{m} - y^{(Trt)})}{P(Y^{(Trt)} + Y^{(Ctr)} = \tilde{m})} = \frac{P(Y^{(Trt)} = y^{(Trt)}) P(Y^{(Ctr)} = \tilde{m} - y^{(Trt)})}{P(Y^{(Trt)} + Y^{(Ctr)} = \tilde{m})} =$$

$$= \frac{1}{P(Y^{(Trt)} + Y^{(Ctr)} = \tilde{m})} n_{Trt}! \prod_{s=1}^{d} \frac{\left(q_s^{(Trt)}\right)^{y_s^{(Trt)}}}{y_s^{(Trt)}!} n_{Ctr}! \prod_{s=1}^{d} \frac{\left(q_s^{(Ctr)}\right)^{\tilde{m}_s - y_s^{(Trt)}}}{(\tilde{m}_s - y_s^{(Trt)})!} =$$

$$P(Y^{(Trt)} = y^{(Trt)} | M = \tilde{m}) = \frac{1}{N} \prod_{s=1}^{d} \frac{1}{y_s^{(Trt)}! (\tilde{m}_s - y_s^{(Trt)})!} \left(\frac{q_s^{(Trt)}}{q_s^{(Ctr)}} \right)^{y_s^{(Trt)}} \qquad (17)$$

Here, $N = \sum_{y^{(Trt)} \in W} \prod_{s=1}^{d} \frac{1}{y_s^{(Trt)}!(\tilde{m}_s - y_s^{(Trt)})!} \left(\frac{q_s^{(Trt)}}{q_s^{(Ctr)}} \right)^{y_s^{(Trt)}}$ is a normalizing constant.

When $q_s^{(Trt)} = q_s^{(Ctr)}$ for all $s = 1, ..., d$, (17) is equivalent to the permutation distribution of $Y^{(Trt)}$ given $\tilde{m}$ that results from performing all possible permutations of the group labels and (16) is equivalent to the corresponding permutation distribution of T. Similar to the minP approach [51, 52], this permutation distribution of T is now used as null distribution under $H_0 : \cap_{i=1}^{k} H_i$. Optimal multivariate rejection regions are determined by applying the optimization procedures of section 3 over the search space $V = h(W)$.

Analogously, for a test of $H_J = \cap_{i \in J} H_i$, $J \subset \{1, ..., k\}$, the joint permutation distribution of $(T_i : i \in J)$ and the respective search space and optimal rejection regions are determined using only the data on endpoints $i \in J$.

When testing the intersection hypotheses H_J by tests based on these permutation distributions, the closed testing procedure controls the FWER under the following additional exchangeability assumption, which is similar to the marginals-determine-the-joint condition given in [47].

Assumption 1. (Exchangeability) Let $K \subseteq \{1, ..., k\}$ be the index set of all true null hypotheses. The joint distribution of the observations on endpoints $i \in K$ is assumed to be identical in both treatment groups.

FWER control follows, because under Assumption 1 the conditional null distribution of $(T_i : i \in K)$ given the respective table margins for the data on the endpoints $i \in K$, say $\tilde{m}_K$, is its permutation distribution. So the type I error rate of the permutation test for H_K is controlled conditional on $\tilde{m}_K$, and since this holds for any realization of $\tilde{m}_K$ it also holds unconditionally. Type I error rate control of the test for H_K is sufficient for FWER control of the closed test.

See [48–50] for further discussion on testing multiple hypotheses using permutation tests and involved assumptions. In particular, Westfall and Troendle [48] argue that in absence of any marginal treatment effect it may be reasonable to also assume exchangeability. They further note that in absence of the exchangeability assumption, the rejection of an exchangeability null hypothesis would at least imply some treatment effect, though it could be on the correlation rather than on the actual parameter of interest and also it may not allow for a clear interpreation of the direction of the treatment effect. As illustration consider the following example with two binary endpoints.

Example 7. Assume two binary endpoints according to the above framework with true joint distributions given by $q_{s_1s_2}^{(g)}, g \in \{Trt, Ctr\}, s_1s_2 \in \mathcal{S} = \{11, 10, 01, 00\}$ and marginal null hypotheses $H_i : OR_i \leq 1, i = 1, 2$. Assume that that for each endpoint the marginal success probabilities are identical for both groups, i.e. $q_{11}^{(Trt)} + q_{10}^{(Trt)} = q_{11}^{(Ctr)} + q_{10}^{(Ctr)}$ and $q_{11}^{(Trt)} + q_{01}^{(Trt)} = q_{11}^{(Ctr)} + q_{01}^{(Ctr)}$. So the true distributions are under the global null hypothesis $H_1 \cap H_2$. Without any further assumption, though, the correlation structures could be different between the two groups such that, e.g., $q_{11}^{(Trt)} > q_{11}^{(Ctr)}$ and $q_{00}^{(Trt)} > q_{00}^{(Ctr)}$, even though $H_1 \cap H_2$ holds. In this setting, a level α permutation test could reject the exchangeability null hypothesis $H_{exch} : q_{s_1s_2}^{(Trt)} = q_{s_1s_2}^{(Ctr)}$ for all $s_1s_2 \in \mathcal{S}$ with a probability larger than α.

Note that Assumption 1 is not required when the Bonferroni-type tests are applied to the marginal distributions of multiple Fisher's exact test statistics. Further note that in any case only information on the observed table margins $\tilde{m}$ is required to perform the proposed tests, such that in a clinical trial the rejection region can be defined before unblinding the data.

4.2 Relation to the conditional inference framework

The above statistical model for multiple binary endpoints can be seen as a generalization of the model for a single binary endpoint discussed in section 1.5. As such the proposed tests fit in the general framework of conditional inference, as will be laid out in this section.

The density of $Y^{(j,g)}$ is $f(y) = P(Y^{(j,g)} = y) = \prod_{s=1}^{d} \left(q_s^{(g)}\right)^{y_s}$ for $y \in \{z \in \{0,1\}^d : \sum_{s=1}^{d} z_s = 1\}$ and 0 else. The following treatment of the density applies to, both, the treatment and the control group and so the superscript (g) is omitted for ease of notation. The density can be written in exponential family form as

$$f(y) = \prod_{s=1}^{d} (q_s)^{y_s} = \exp\left(\sum_{s=1}^{d} y_s \log q_s\right) =$$

$$= \exp\left(\sum_{s=1}^{d-1} (y_s \log q_s - y_s \log q_d + y_s \log q_s) + y_d \log q_d\right) =$$

$$= \exp\left(\sum_{s=1}^{d-1} y_s \log\left(\frac{q_s}{q_d}\right) + \log(q_d) \sum_{s=1}^{d} y_s\right) =$$

$$= \exp\left(\sum_{s=1}^{d} y_s \log\left(\frac{q_s}{q_d}\right) + \log(q_d)\right)$$

where we use $\log(q_d/q_d) = 0$ and $\sum_{s=1}^{d} y_s = 1$ in the last step.

Analogous to the log-odds for a single binary endpoint, the natural parameters for this density are $\theta_s = \log\left(\frac{q_s}{q_d}\right)$, or, as vector, $\theta = (\theta_1, ..., \theta_d)$. Further note that $q_d = 1 - \sum_{s=1}^{d-1} q_s = 1 - \sum_{s=1}^{d-1} \exp(\theta_s) q_d$ and, hence, $q_d = \left(1 + \sum_{s=1}^{d-1} \exp(\theta_s)\right)^{-1}$. Then

$$f(y) = \exp\left(y^T\theta - \log\left(1 + \sum_{s=1}^{d-1} \exp(\theta_s)\right)\right)$$

The parameter vector $\theta^{(g)} = \left(\theta_s^{(g)} = \log\left(\frac{q_s^{(g)}}{q_d^{(g)}}\right) : s = 1, ..., d\right)$ for a patient in group g is explained in terms of the regression model $\theta^{(g)} = \beta^{(0)} + x_{jg}\beta^{(1)}$ to jointly model the success probabilities in the two groups. Here, $\beta^{(0)}$ and $\beta^{(1)}$ are $d \times 1$ vectors and $x_{jg} = 1$ for $g = Trt$ and $x_{jg} = 0$ for $g = Ctr$. $\beta^{(0)}$ is a nuisance parameter, while $\beta^{(1)}$ captures the treatment effect. For an interpretation of the coefficient vector $\beta^{(1)}$, note that $\exp(\beta_s^{(1)}) = (q_s^{(Trt)}/q_d^{(Trt)})/(q_s^{(Ctr)}/q_d^{(Ctr)})$. This is akin to an odds ratio, however the ratios are calculated with respect to the common reference class d. Note that from the constraint $\sum_{s=1}^{d} q_s = 1$, the constraint for the reference class $\theta_d = 0$ is inherited, and consequently $\beta_d^{(0)} = \beta_d^{(1)} = 0$.

Similar to equation (1) in the one endpoint case, the joint density of the observations from both groups is

$$f\left(\left(y^{(j,g)} : j = 1, ..., n_g, g \in \{Trt, Ctr\}\right); x, \beta^{(0)}, \beta^{(1)}\right) =$$

$$= P_{x,\beta^{(0)},\beta^{(1)}}\left(Y^{(1,Trt)} = y^{(1,Trt)}, \ldots, Y^{(n_{Ctr},Ctr)} = y^{(n_{Ctr},Ctr)}\right) =$$

$$\prod_{g\in\{Trt,Ctr\}} \prod_{j=1}^{n_g} f\left(y^{(j,g)}; x_{jg}, \beta^{(0)}, \beta^{(1)}\right) =$$

$$= \exp\left(\sum_{g\in\{Trt,Ctr\}} \sum_{j=1}^{n_g} {y^{(j,g)}}^T \beta^{(0)}\right) \exp\left(\sum_{g\in\{Trt,Ctr\}} \sum_{j=1}^{n_g} {y^{(j,g)}}^T x_{jg}\beta^{(1)}\right) A =$$

$$= \exp\left(\left(\sum_{g\in\{Trt,Ctr\}} \sum_{j=1}^{n_g} {y^{(j,g)}}^T\right)\beta^{(0)}\right) \exp\left(\left(\sum_{g\in\{Trt,Ctr\}} \sum_{j=1}^{n_g} {y^{(j,g)}}^T x_{jg}\right)\beta^{(1)}\right) A$$

with $A = \prod_{g\in\{Trt,Ctr\}} \prod_{j=1}^{n_g} \left(1 + \sum_{s=1}^{d-1} \exp(\beta^{(0)} + x_{jg}\beta^{(1)})\right)^{-1}$.

By the factorization theorem, the vector of total frequencies across both groups $M = \sum_{g\in\{Trt,Ctr\}} \sum_{j=1}^{n_g} Y^{(j,g)}$ is a sufficient statistic for $\beta^{(0)}$. The vector of frequencies in the treatment group $Y^{(Trt)} = \sum_{j=1}^{n_g} Y^{(j,Trt)}$ is a sufficient statistic for $\beta^{(1)}$. (With other choices of encoding the two groups in x_{jg}, other equivalent sufficient statistics are found.) By conditioning on M, the unknown nuisance parameter $\beta^{(0)}$ can be removed and tests based on the conditional distribution of $Y^{(Trt)}$ can be defined to test hypotheses about $\beta^{(1)}$. Note that $Y^{(Trt)}$ itself is not a suitable test statistic for the directional tests that are of interest here, because large treatment effects may result in some entries of $Y^{(Trt)}$ to be large and others to be small. The aggregation in terms of marginal Fisher's exact test statistics $T = h(Y^{(Trt)})$ provides a multivariate test statistic with better relation to the marginal treatment effects. Also, it is of some convenience to use subsets of the same vector of test statistics throughout the full closed testing procedure.

Following the above derivation, it should be further noted that instead of using ratios $\frac{q_s^{(Trt)}}{q_s^{(Ctr)}}$ in (17) an equivalent representation that uses the odds-ratio analoga given by $\exp(\beta_s^{(1)})$ is

$$P(Y^{(Trt)} = y^{(Trt)}|M = \tilde{m}) = \frac{1}{N'} \prod_{s=1}^{d-1} \frac{1}{y_s^{(Trt)}!(\tilde{m}_s - y_s^{(Trt)})!} \left(\frac{q_s^{(Trt)}/q_d^{(Trt)}}{q_s^{(Ctr)}/q_d^{(Ctr)}} \right)^{y_s^{(Trt)}}$$

with N' the according normalization constant. This representation follows from noting $y_d^{(Trt)} = n_{Trt} - \sum_{s=1}^{d-1} y_s^{(Trt)}$ in (17) or from performing the derivation with the exponential family form of the multinomial distribution.

4.3 A medical example with two endpoints

This section is entirely dedicated to a numeric example to illustrate the proposed testing procedures and give some comparison to the methods described in the literature. An exemplary analysis similar to the one presented here can be performed using the R functions in the appendix section 8.1 and the commands in the appendix section 8.2.

Example 8. Consider a trial for a new treatment of patent ductus arteriosus in preterm infants similar to that described in [60]. In this study, the drug indomethacin was compared to the more recent competitor ibuprofen with respect to several endpoints. The primary endpoint was ductal closure and a major secondary endpoint was low urine output. For our example we will consider both as primary endpoints. Only marginal frequencies

were reported in the original study. For the example, the data is assumed to exhibit the same marginal observed success rates as in [60] and joint frequencies that could be expected under the assumption of independence between the two endpoints. Further, for illustration a true distribution of events close to the observed distribution is assumed.

The example data and the assumed alternative are shown in Table 4. The marginal odds ratios for the endpoints urine output and ductal closure under this alternative are $OR_{urine} = 6.6$ and $OR_{duct.} = 1.3$, respectively. For this example, the aim is to establish superiority of the new treatment over the control treatment in at least one of the two endpoints, at a global 2.5% level of significance. Thus, the elementary null hypotheses are $H_1 : OR_{urine} \leq 1$ and $H_2 : OR_{duct.} \leq 1$. The global intersection null hypothesis is $H_0 = H_{urine} \cap H_{duct.}$. Following the discussion above, exchangeability under H_0 is assumed.

Table 4: Exemplary data for observations on two binary endpoints with marginal distributions similar to the data in [60]. The example data is used to illustrate the analysis using optimal exact conditional tests.

	Observed frequencies		Assumed true distribution	
	Treatment	Control	Treatment	Control
Success both endpoints	80	57	0.83	0.68
Success urine output only	13	12	0.14	0.15
Success ductal closure only	1	10	0.02	0.13
No success	0	2	0.01	0.04

To test the global intersection null hypothesis H_0, the test statistics from the marginal Fisher's exact tests for the two endpoints are used. For the vector of these test statistics, the bivariate rejection regions for the unweighted Bonferroni test, Tarone's test [44], the Hommel and Krummenauer improvement thereof (HKT) [45], weighted consonant Bonferroni tests optimized according to (10)-(12), the Bonferroni test resulting from the according greedy algorithm (section 3.4), the minP test [52], the optimal tests using the joint distribution with or without the consonance constraint, solving (6)-(8), as well as the test resulting from the greedy algorithm for multivariate rejection regions were caclulated.

As a first step, Figure 4 illustrates the result of the pre-processing for the example data

set, both, with and without the consonance constraint. In this example the initial search space contains 386 points. Due to the pre-processing, the search space can be reduced to 195 points. If consonance is enforced, only 123 eligible points remain.

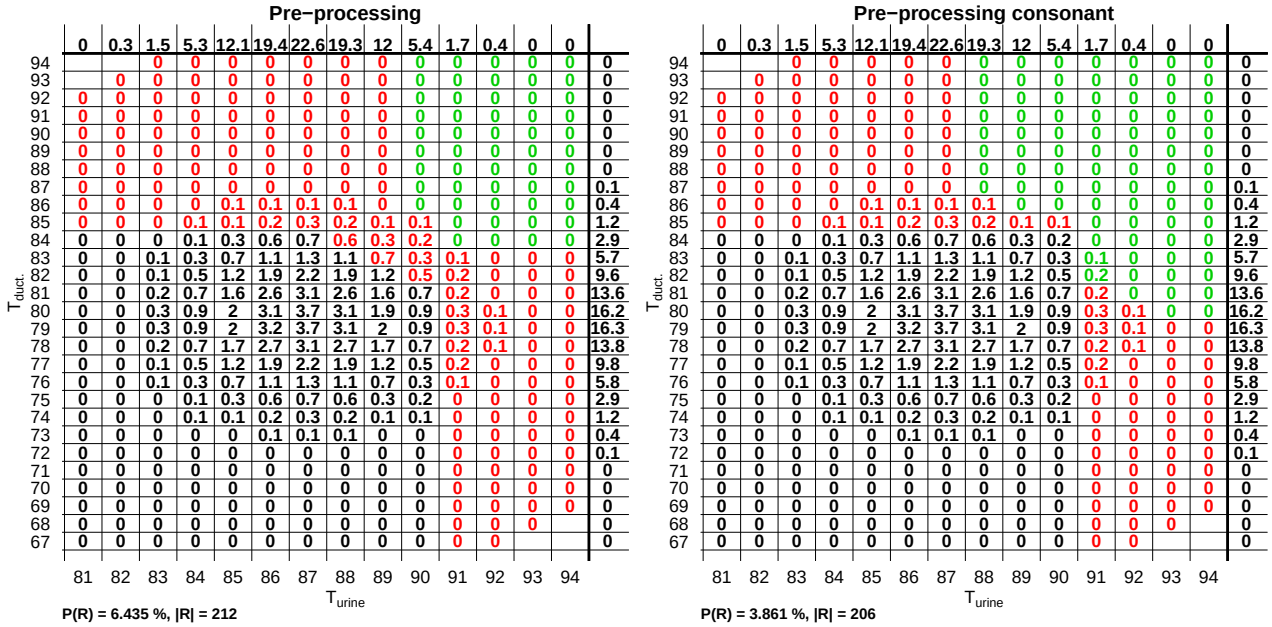

Figure 4: Effect of pre-processing on the effective search space in the example, without consonance constraint (left panel) and with the consonance constraint (right panel). Points removed in the first pre-processing step are indicated in black colour. Points removed in the second pre-processing step, because they are identified to be definitely contained in an optimal rejection region, are indicated in green colour. The remaining search space is visualized in red colour.

For the considered tests and conditional on the example data, Table 5 shows the actual type I error rate under H_0, the power under the alternative specified in Table 4 and the number of points a rejection region contains. For the Bonferroni-type tests and for the minP test, the critical boundaries are included in the table. Further, the effect of pre-processing is described in the table by showing the number of elements before and after the first and second pre-processing step. The rejection regions for the proposed optimal tests are visualized in Figures 5 and 6, together with the conditional joint distribution of the test statistics under the null hypothesis and under the alternative, respectively.

Table 5 shows that in this example conservatism is greatly reduced and the conditional power increased when using optimal tests as compared to a basic Bonferroni test. Using

the test with optimal alpha exhaustion even results in an actual type I error rate virtually identical to 2.5%, and an increase in power of more than 10 percentage points compared to the Bonferroni test. Using the rejection region which optimizes the power under the true alternative leads to an additional increase by several percentage points.

The marginal Fisher's exact tests reject at local level 2.5% if $T_{urine} \geq 91$ and $T_{duct.} \geq 85$. The observed values for the test statistics in the example are $t_{urine} = 93$ and $t_{duct.} = 81$. The point $(93, 81)$ is contained in all considered rejection regions (see Figures 5 and 6), and so all examined tests reject H_0. Following the application of the closed testing principle all tests also reject H_{urine}, concluding that the proportion of patients with low urine output is lower under the new medication.

In this example, the rejection region with maximal number of elements results in a consonant closed test. However, the tests maximizing alpha exhaustion or power are not consonant here, as both regions contain points with simultaneously $T_{urine} < 91$ and $T_{duct.} < 85$. The lower row in Figure 5 and 6 shows the respective rejection regions, when the additional constraint of consonance is imposed. The consonance constraint has little impact on the test with optimal power. The test optimized for alpha exhaustion even has better conditional power under the additional constraint of consonance. There is virtually no trade-off in terms of exhaustion of the nominal level, because in the example there is a large number of points in the joint distribution with very small probability mass. By adding or removing some of these points, the actual level can be set in fine enough steps to diminish the effect of discreteness. However, most of these points also have very small

---

Figure 5 *(following page)*: Optimal rejection regions for the example data set. The figure shows the conditional joint distribution of the Fisher's exact test statistics, T_{urine} and $T_{duct.}$, for the two endpoints under the null hypothesis. Probabilities are given in percent and rounded to one decimal place. Cells with entries 0 have a small positive probability, empty cells have probability zero. The upper and the right margin show the marginal distributions of T_{urine} and $T_{duct.}$. The rejection regions following optimization with different optimization goals are indicated in red colour. The probability mass of the rejection region ($P(R)$) and the number of elements in the region ($|R|$) are displayed below each graphic. The nominal significance level is 2.5% for all tests.

Greedy algorithm

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.41 %, |R| = 187

Optimal area

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.483 %, |R| = 191

Optimal alpha exhaustion

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.5 %, |R| = 120

Optimal power

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.488 %, |R| = 183

Optimal alpha exhaustion with consonance

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.5 %, |R| = 157

Optimal power with consonance

T_{duct} \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
87	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
86	0	0	0	0	0.1	0.1	0.1	0.1	0	0	0	0	0	0	0.4
85	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
84	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
83	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.7
82	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.6
81	0	0	0.2	0.7	1.6	2.6	3.1	2.6	1.6	0.7	0.2	0	0	0	13.6
80	0	0	0.3	0.9	2	3.1	3.7	3.1	1.9	0.9	0.3	0.1	0	0	16.2
79	0	0	0.3	0.9	2	3.2	3.7	3.1	2	0.9	0.3	0.1	0	0	16.3
78	0	0	0.2	0.7	1.7	2.7	3.1	2.7	1.7	0.7	0.2	0.1	0	0	13.8
77	0	0	0.1	0.5	1.2	1.9	2.2	1.9	1.2	0.5	0.2	0	0	0	9.8
76	0	0	0.1	0.3	0.7	1.1	1.3	1.1	0.7	0.3	0.1	0	0	0	5.8
75	0	0	0	0.1	0.3	0.6	0.7	0.6	0.3	0.2	0	0	0	0	2.9
74	0	0	0	0.1	0.1	0.2	0.3	0.2	0.1	0.1	0	0	0	0	1.2
73	0	0	0	0	0	0.1	0.1	0.1	0	0	0	0	0	0	0.4
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0
	0	0.3	1.5	5.3	12.1	19.4	22.6	19.3	12	5.4	1.7	0.4	0	0	

P(R) = 2.5 %, |R| = 170

probabilities under the alternative, so that maximal alpha exhaustion does not directly translate into larger power. The results on the example data serve for illustration, though, and do not allow for immediate generalization. A systematic study of the unconditional properties of the proposed tests is given in the next section.

Figure 6 *(following page)*: Optimal rejection regions for the example data set. The figure shows the conditional joint distribution of the Fisher's exact test statistics, T_{urine} and $T_{duct.}$, for the two endpoints under the alternative specified in Table 4. See the caption of Figure 5 for further details.

Greedy algorithm

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 79.014 %, |R| = 187

Optimal area

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 75.031 %, |R| = 191

Optimal alpha exhaustion

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2	
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 81.94 %, |R| = 120

Optimal power

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 88.279 %, |R| = 183

Optimal alpha exhaustion with consonance

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2	
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 87.772 %, |R| = 157

Optimal power with consonance

$T_{duct.}$ \ T_{urine}	81	82	83	84	85	86	87	88	89	90	91	92	93	94	
94			0	0	0	0	0	0	0	0	0	0	0	0	0
93		0	0	0	0	0	0	0	0	0	0	0	0	0	0
92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
91	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
89	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
88	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
87	0	0	0	0	0	0	0	0	0	0	0.1	0.2	0.1	0.1	0.5
86	0	0	0	0	0	0	0	0	0	0.1	0.3	0.4	0.4	0.2	1.5
85	0	0	0	0	0	0	0	0	0.1	0.3	0.7	1	0.9	0.4	3.5
84	0	0	0	0	0	0	0	0	0.2	0.6	1.4	2	1.8	0.7	6.8
83	0	0	0	0	0	0	0	0.1	0.3	1	2.2	3.3	2.9	1.2	10.9
82	0	0	0	0	0	0	0	0.1	0.4	1.3	3	4.4	3.9	1.6	14.8
81	0	0	0	0	0	0	0	0.1	0.5	1.5	3.4	5	4.4	1.8	16.7
80	0	0	0	0	0	0	0	0.1	0.5	1.5	3.2	4.8	4.2	1.7	15.9
79	0	0	0	0	0	0	0	0.1	0.4	1.2	2.6	3.8	3.3	1.3	12.7
78	0	0	0	0	0	0	0	0.1	0.3	0.8	1.8	2.6	2.2	0.8	8.5
77	0	0	0	0	0	0	0	0	0.1	0.5	1	1.4	1.2	0.4	4.7
76	0	0	0	0	0	0	0	0	0.1	0.2	0.5	0.7	0.5	0.2	2.2
75	0	0	0	0	0	0	0	0	0	0.1	0.2	0.3	0.2	0.1	0.8
74	0	0	0	0	0	0	0	0	0	0	0.1	0.1	0.1	0	0.3
73	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.1
72	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
68	0	0	0	0	0	0	0	0	0	0	0	0	0		0
67	0	0	0	0	0	0	0	0	0	0	0	0			0

P(R) = 87.772 %, |R| = 170

Table 5: Characteristics of different rejection regions for testing the global null hypothesis $H_0 = H_{urine} \cap H_{duct.}$ in the example data set. The table shows type I error rate (level) and power to reject H_0 in percent, both calculated conditionally on the observed total success numbers. Further columns show the number of points included in a rejection region ($|R|$) and the critical boundaries ($c_{urine}, c_{duct.}$) for the two test statistics for those tests with a rectangular cone-type rejection region such that H_0 is rejected if $T_{urine} \geq c_{urine}$ or $T_{duct.} \geq c_{duct.}$. $|V|$, $|V^{(1)}|$ and $|V^{(2)}|$ are the number of elements in the search space, initially and after the first and second pre-processing step, respectively. The number of iterations that were required in the branch and bound algorithm to identify an optimal solution is given in the last column. The nominal type I error rate for all tests is 2.5%. Tests that result in identical rejection regions for the example are summarized jointly in one row.

Test	Level	Power	$\|R\|$	c_{urine}	$c_{duct.}$	$\|V\|$	$\|V^{(1)}\|$	$\|V^{(2)}\|$	Iterations
Bonferroni/Tarone/HKT	0.98	67.4	177	92	86				
Bonferroni optimal alpha/optimal power	2.27	87.2	186	91	87				
Bonferroni optimal area/Greedy/minP	2.17	68.5	188	92	85				
Optimal alpha	2.50	81.9	120			386	212	159	357591
Optimal area	2.48	75.0	191			386	212	159	5084
Optimal power	2.49	88.3	183			386	212	159	6865
Consonant optimal alpha	2.50	87.8	157			386	206	123	45317
Consonant optimal area	2.48	75.0	191			386	206	123	1160
Consonant optimal power	2.50	87.8	170			386	206	123	2342
Greedy algorithm	2.41	79.0	187						

5 Power of the optimal procedures

The unconditional power of different closed testing procedures based on either of the discussed intersection hypothesis tests was studied for the setting of two and three binary endpoints. These intersection hypothesis tests were the Bonferroni test, Tarone's test, the HKT test, optimally weighted consonant Bonferroni tests with optimization aims (10)-(12), the greedy algorithm derived Bonferroni test, the minP test, optimal tests using the joint permutation distribution with the optimization aims (6)-(8) and the greedy algorithm test for joint distributions. For the case of two endpoints, also the optimal joint distribution-based tests with the additional constraint of consonance were studied. In all cases, one-sided Fisher's exact tests were applied for the local elementary hypothesis tests and the vector of the elementary test statistics was used as multivariate test statistic in the intersection hypothesis tests.

The between-groups differences in these scenarios were parametrized by the marginal success rates $p_i^{(Trt)}, i = 1, 2, (3)$ in the treatment group and $p_i^{(Ctr)}, i = 1, 2, (3)$ in the control group. Further, a common product-moment correlation ρ between the binary observations within each observational unit was assumed.

For immediate illustration of the general findings, results for a selected scenario with three endpoints are included in the main text in Table 6. Detailed results of the power study are tabulated in the appendix sections 8.3 and 8.4.

5.1 Scenarios with two endpoints

For two endpoints, scenarios with $n \in \{5, 10, 15, 20\}$ observational units per group and a nominal family wise significance level of $\alpha = 0.025$ were considered. For each sample size, scenarios with a treatment effect in one endpoint only, and scenarios with an effect in both endpoints were considered. For $n = 15$, additional scenarios with $\alpha = 0.05$ and with no effect in either endpoint were included. $p_i^{(Trt)}$ and $p_i^{(Ctr)}$ were chosen such that for efficacious endpoints $(p_i^{(Trt)} + p_i^{(Ctr)})/2 = 0.5$ and such that the power to reject the hypothesis at an unadjusted level α using a single Fisher's exact test is approximately 0.6. E.g. for the setting with $n = 15$, $\alpha = 0.025$ and an effect in both endpoints this results in $p_1^{(Trt)} = p_2^{(Trt)} = 0.735$ and $p_1^{(Ctr)} = p_2^{(Ctr)} = 0.265$, which correspond to a local power of 0.605 for each marginal Fisher's exact test. This way of setting the marginal success

rates was chosen in order to obtain scenarios that would be reasonable in a real trial and to allow for comparison of the results across different settings for n and α.

The product-moment correlation between endpoints within each treatment was assumed to be identical and either $\rho = 0$ or $\rho = 0.5$ when studying the case of an effect in both endpoints or no effect in any endpoint. For the scenarios with an effect in exactly one endpoint the correlation was always set to 0. This was due to the fact that for two binary random variables with unequal marginal success rates the maximal possible correlation is bounded below 1, and for the included settings it is typically even bounded below 0.5. E.g. with $p_1^{(Trt)} = 0.735$ and $p_2^{(Trt)} = 0.265$, the maximal possible correlation would be 0.36.

For those tests that aim to directly maximize the power, assumptions on the alternative need to be specified. For each scenario the optimal power tests were calculated under three different assumptions matching the three overall scenarios of an effect in one endpoint, an effect in both endpoints with $\rho = 0$ and an effect in both endpoints with $\rho = 0.5$. In this way, the true alternative is always included, and in addition the characteristics of the tests under assumptions that deviate from the truth can be assessed.

The power was calculated numerically. The rejection region of each test was calculated for each possible combination of total numbers of observational units with success in no endpoint, endpoint 1 only, endpoint 2 only, or both endpoints, i.e. for each possible row margin $\tilde{m}$ in the notation of section 4.1. The rejection probability conditional on the specific marginal success numbers was calculated from the conditional joint distribution of the test statistics under the specified alternative as given by (16). The unconditional power was then calculated as the weighted sum of the conditional power values, with weights equal to the probability to observe the respective marginal success numbers. With two endpoints, the number of possible row margin vectors is $(2n+1)^3/6 + (2n+1)^2/2 + (2n+1)/3$. For $n = 15$, this results in 5456. The maximal number of iterations in the branch and bound algorithm was set to 20000. The algorithm requires approximately 1 second of computation time for 10000 iterations at a 2.4 GHZ CPU. The optimal solution was found within this number of iterations, except for a few instances in scenarios with $n = 15$ or $n = 20$. For these rare cases the best feasible solution available after the maximal number of iterations was used to define the rejection region.

All results are tabulated in detail in the appendix tables A1 to A19 in section 8.3.

General findings from the numeric calculations are summarized below together with those on three endpoints.

5.2 Scenarios with three endpoints

For the case of three endpoints, six different scenarios with $n = 10$ and $\alpha = 0.025$ were studied. In the first two scenarios, no effect in any endpoint was assumed with all marginal success rates equal to 0.254, the correlation was either $\rho = 0$ or $\rho = 0.5$. In the next scenario, the marginal success rate for the first endpoint in the treatment group was $p_1^{(Trt)} = 0.746$, all other marginal success rates were 0.254 and the correlation was 0. In the further two scenarios, ρ was again either 0 or 0.5 and a common effect in all endpoints was assumed such that $p_i^{(Trt)} = 0.746$ and $p_i^{(Ctr)} = 0.254$ for all $i = 1, 2, 3$. These values for the marginal success rates result in a local power for one marginal Fisher's exact test of 0.41, and the power to reject the global intersection null hypothesis is between 0.6 and 0.9 for most tests. Therefore the selected scenarios should be a useful representation of reasonably powered small sample trials. In the final scenario, $\rho = 0.5$ and different effects in all three endpoints of $p_1^{(Trt)} = 0.8$, $p_2^{(Trt)} = 0.7$, $p_3^{(Trt)} = 0.6$ and $p_i^{(Ctr)} = 0.2$ for all $i = 1, 2, 3$ were assumed. The corresponding local power values of Fisher's exact tests are 0.64, 0.43 and 0.25.

The power was calculated by simulation. For each scenario, 2000 random data samples were drawn from the assumed multivariate Bernoulli distributions using the function rmvbin in the library bindata in R [61].

For tests optimizing the power, the assumed alternative was identical to the true alternative in all scenarios with three endpoints. The branch and bound algorithm was used with a maximal number of $5 \cdot 10^5$ iterations. This was sufficient in the case of correlated endpoints. However, in scenarios with zero correlation and especially when aiming to maximize the exhaustion of the nominal significance level, in many instances no confirmed optimal solution was found for the rejection region of the global intersection hypothesis test. As before, the best available feasible, though potentially suboptimal, solution was used in these cases.

The results for the first five scenarios are given in the appendix tables A20 to A24 in section 8.4. The results for the final scenario are shown in Table 6.

Table 6: Power with three binary endpoints with different effect sizes. The scenario for this table is $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.8$, $p_2^{(Trt)} = 0.7$, $p_3^{(Trt)} = 0.6$, $p_1^{(Ctr)} = p_2^{(Ctr)} = p_3^{(Ctr)} = 0.2$, $\rho = 0.5$. The calculations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$. The table shows the probabilities in percent that the closed testing procedure rejects the global intersection hypothesis $\cap_{i=1}^3 H_i$, at least one elementary null hypothesis (Any H_i), all elementary hypotheses simultaneously (All H_i), or particularly H_1, H_2 or H_3. The proportion of simulated instances in which a confirmed optimal solution was found by the branch and bound algorithm for all four intersection hypothesis tests in the closed testing procedure is given in the column labelled C(%). To further quantify the number of required iterations, the median (q50), the 90 % quantile (q90) and the maximum (Max) of the number of iterations are shown in the last three columns. A value of $5 \cdot 10^5$ indicates that the respective quantity to achieve an optimal solution would be above $5 \cdot 10^5$, but is here bounded by the maximal number of iterations.

Test	$\cap_{i=1}^3 H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	52.8	52.8	16.1	48.7	33.5	23.0				
Tarone	55.9	55.9	16.1	51.3	34.6	23.2				
HKT	55.9	55.9	16.1	51.3	34.6	23.2				
Bonferroni optimal alpha	72.5	72.5	16.5	63	40.8	27.4				
Bonferroni optimal area	71.8	71.8	16.5	62.5	40.6	27.4				
Bonferroni optimal power	73.5	73.5	16.5	64.2	41.1	27.1				
Bonferroni greedy algorithm	72.3	72.3	16.5	63.4	40.7	27.1				
minP	72.5	72.5	16.4	63.2	40.5	27.2				
Optimal alpha	78.6	66.0	16.5	57.5	38.6	27.0	99.5	44	2717.6	5e+05
Optimal area	80.6	73.5	16.5	64	41.2	27.3	100	36	595.2	133909
Optimal power	84.5	72.2	16.5	63.4	40.1	27.0	100	12	144	24591
Greedy algorithm	79.2	73.8	16.5	64.2	41.2	27.5				

5.3 Resulting power

Overall, the unweighted Bonferroni procedure had the lowest power in the numeric study. Tarone's test mostly provided only small improvements over the unweighted Bonferroni test in the studied scenarios. Especially for two endpoints this is not too surprising, because in this case these tests can have different results only when exactly one of the two marginal Fisher's exact tests cannot become significant at level α. This event has low probability, unless the sample size is very small. Indeed, with $n = 5$ Tarone's test was clearly more powerful than the unweighted Bonferroni test. The Hommel and Krummenauer improvement of Tarone's test resulted in slightly more power for some scenarios, but mostly these tests behaved identically.

Conservatism was reduced and the power was notably increased by an order of 10 percentage points when the local boundaries in Bonferroni-type tests were chosen more flexibly, meeting one of the proposed optimization criteria or using the greedy algorithm. When the effects in both endpoints were equal, there was almost no difference in the performance of these approaches. In the case with an effect in one endpoint only, however, optimizing power according to the power criterion (12) under the assumption of the true effects provided some additional advantage.

The minP test had highly similar power values as the weighted Bonferroni tests. This is of particular interest, as the intersection null hypotheses tested with the Bonferroni tests are of the more general type about marginal effects, while the minP test first of all tests hypotheses of exchangeability (compare section 4.1) and thus might have reduced interpretability without providing an advantage in terms of power in the studied settings.

The optimal tests based on the joint distribution provided a substantial improvement over the Bonferroni-type tests and over the minP test in the order of another 10 percentage points. Optimizing the power under the true alternative resulted in the largest power and the power of this, by definition, truly optimal test can be seen as a benchmark for the other tests. When there was an effect in only one endpoint, optimizing exhaustion of the nominal level or optimizing power under the wrong assumptions in some scenarios resulted in power similar to that of the unweighted Bonferroni test, though. In contrast, maximizing the number of points in the rejection region gave more robust results, with power values above that of the weighted Bonferroni tests.

For the optimal tests using the joint distribution, enforcing consonance in the scenarios with two endpoints did not lead to a notable improvement of the power to reject at least one elementary hypothesis, but at the same time decreased the power to reject the global intersection null hypothesis. Thus, the consonance property may be argued not be of special concern when optimizing the rejection region following the proposed framework.

The tests obtained through the greedy algorithms, both for the Bonferroni approach and the joint distribution rejection region, performed surprisingly well. In most scenarios these tests had power similar to or above that of the other tests in their class.

Differences in the power characteristics between the tests were mostly observed for the decision on the global intersection hypothesis. When deciding on the elementary hypotheses, all tests other than Bonferroni, Tarone's test and HKT showed similar power values, with some exceptions in the case of an effect in just one endpoint. The probability to reject all elementary hypotheses simultaneously was almost identical for most of the studied tests. This is another consequence of the discreteness of the elementary tests, by which the set of values for the multivariate test statistic, that lead to local rejection of all two or three elementary null hypotheses simultaneously, is often entirely contained in most optimal rejection regions. For illustration recall the numeric example of section 4.3. There, the set $\{(t_{urine}, t_{duct.}) : t_{urine} \geq 91, t_{duct.} \geq 85\}$ was contained in all optimal rejection regions. Even the simple Bonferroni rejection region almost completely contained this set, missing only the single point $(T_{urine} = 91, T_{duct.} = 85)$.

For all studied tests other than Bonferroni, Tarone's and HKT, the power to reject a specific elementary hypothesis was typically very close to the local power that a single Fisher's exact test would have under the chosen marginal success rates. The power to reject at least one elementary hypothesis was even larger, unless only one endpoint showed a true effect. This observation implies that carefully accounting for the discreteness of the tests allows one to greatly reduce the cost of multiple testing. In the studied scenarios, the FWER was controlled, while testing two or three instead of one hypothesis was possible almost without reduction of power due to the multiplicity correction.

6 Discussion

The analysis of small clincial trials is often challenging [6]. Asymptotic methods may lack type I error rate control when the sample size is small, or when there are few events in the case of binary endpoints. Exact tests guarantee type I error rate control, but they are typically overly conservative due to discreteness. At the same time it is required to make best use of the available information, if the overall number of observations is limited, which favors the analysis of multiple endpoints. This was a motivation to propose the methods to find optimal rejection regions for multivariate exact tests described in this work.

Most multiple testing adjustments lead to multivariate rejection regions of a certain restricted shape, e.g. a rectangular cone in the case of Bonferroni tests or the minP test. Within a class of shapes, optimization can be performed, however any such shape-restriction leads to increased conservatism when applied to discrete distributions, and as a consequence it has the potential to reduce the power of the test. This limitation can be avoided by allowing for arbitrarily shaped rejection regions. Still, some constraints are required to allow for unambiguous interpretation of the results. In contrast to earlier suggestions for optimal tests with discrete statistics [53, 54], in this work it is considered important that the test decision is monotonic in the value of the test statistic, leading to an additional constraint in the optimization framework. This condition avoids disputable decisions that would arise, if a relatively small observed effect would result in rejecting a null hypothesis while a stronger effect would not. It further ensures that for a given rejection region, larger treatment effects result in an increase in power. This may be of particular importance when planning a trial for a minimally relevant treatment effect while indeed expecting a much stronger effect. Finally, the simulation results of this work compared to those in [53] clearly show that the monotonicity constraint (or some similar constraint) is required to provide a powerful multiple testing procedure.

Owing to the increased freedom of choosing an (almost) arbitrarily shaped rejection region, it is important to pre-specify how the rejection region will be selected. In any case, the rejection region must be defined before information on the treatment effect is revealed. In a blinded experiment this means to define the rejection region before unblinding the treatment allocation. It is further recommend that all steps of the procedure leading to

an optimized rejection region are specified in the study protocol in a reproducible way, to ensure integrity of the analysis.

The actual choice of the optimization objective may depend on the availability of reasonable assumptions on the alternative. The rejection region optimized for the true alternative can result in a far more powerful test than those resulting from other optimization criteria. As the true alternative is unknown, though, optimizing power under one assumed alternative may not always be the best choice. However, the optimal power test can serve as a useful benchmark to judge the performance of the other optimal tests. Also, different alternatives could be considered simultaneously and weighed according to some prior belief to provide a Baysian-type objective function. The implementation of this approach is straightforward with the described methods, as on the technical level it simply uses yet another objective function. In contrast, if the study is a follow-up confirmation of a previous study, it may be possible to get a good estimate of the true alternative and optimizing power for this alternative will be the first choice.

A reduced power of discrete tests is often attributed to conservativeness of the test decision under the null hypothesis. Therefore, an obvious choice would be to maximize exhaustion of the type I error rate. This approach indeed results in actual levels close to the nominal one, but it does not necessarily result in high power. This may be due to the fact that in optimizing type I error rate exhaustion, there is no need to choose elements with high probability under the alternative, and further there is no need to choose a large number of elements at all, as long as their probability mass under the null has the right sum. Thus, directly optimizing the number of elements may often be preferable.

A useful alternative is the alpha-consistent test using the greedy algorithm proposed in section 3.4. In the scenarios included in the numeric power calculations, this test is often close to optimal. Even though alpha-consistency in itself is not required when deciding on a hypothesis at a pre-specified level of significance, the possibility to calculate p-values for this test may be seen as an advantage in some applications.

In an actual application the assumption of exchangeability under the null hypothesis, which was discussed in section 4.1, should be kept in mind. If this assumption is not made, the optimal Bonferroni-type tests should be preferred over tests using a permutation distribution. Also among these tests, the one using the greedy algorithm showed robust performance and good power. Further it is alpha-consistent and provides conso-

nant procedures by construction and can therefore be recommended as a good general choice.

Taken together, optimizing the rejection region for multivariate exact tests offers a notable advantage over more simple methods in terms of the power to reject a global intersection null hypothesis and, to a lesser extent, the power to reject some elementary null hypothesis. In the small sample setting, where this approach can have the greatest impact, numeric solutions of the discrete optimization problems are found within appreciably short computation times. Application of the optimal exact tests may require an additional effort at the planning stage, which is worthwhile if the aim is to make best use of multiple exact hypothesis tests from a small data sample.

7 References

[1] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. Evidence based medicine: what it is and what it isn't. *British Medical Journal*, 312(7023):71–72, 1996.

[2] ICH E9 Expert Working Group et al. Statistical principles for clinical trials: ICH harmonized tripartite guideline. *Statistics in Medicine*, 18(15):1903–1942, 1999.

[3] Arrigo Schieppati, Jan-Inge Henter, Erica Daina, and Anita Aperia. Why rare diseases are an important medical and social issue. *The Lancet*, 371(9629):2039–2041, 2008.

[4] European Parliament. Regulation (ec) no 141/2000 of the european parliament and of the council of 16. 1999 on orphan medicinal products. *Official Journal of the European Communities*, L18:1–5, 2000.

[5] 21 U.S.C. 360aa, 360bb, 360cc, 360dd, 371. Code of federal regulations, Title 21: Food and drugs, part 316 - orphan drugs. 57 FR 62085, Dec. 29, 1992.

[6] European Medicines Agency, Committe for Medicinal Products for Human use. Guideline on clinical trials in small populations. 2006.

[7] David R Cox and David V Hinkley. *Theoretical statistics*. Chapman & Hall/CRC Press, 1974.

[8] Erich L Lehmann. *Testing statistical hypotheses*. Springer, 2nd edition, 1986.

[9] Alan Agresti. *Categorical data analysis*. John Wiley & Sons, 2nd edition, 2002.

[10] Graham JG Upton. A comparison of alternative tests for the 2× 2 comparative trial. *Journal of the Royal Statistical Society. Series A (General)*, pages 86–105, 1982.

[11] Alan Agresti. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in medicine*, 20(17-18):2709–2722, 2001.

[12] Stian Lydersen, Morten W Fagerland, and Petter Laake. Recommended tests for association in 2× 2 tables. *Statistics in medicine*, 28(7):1159–1175, 2009.

[13] Ronald A Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, pages 39–82, 1935.

[14] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[15] Frank Yates. Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235, 1934.

[16] Alan Agresti. A survey of exact inference for contingency tables. *Statistical Science*, pages 131–153, 1992.

[17] Jun Shao. *Mathematical statistics*. Springer, 1999.

[18] KD Tocher. Extension of the neyman-pearson theory of tests to discontinuous variates. *Biometrika*, 37(1/2):130–144, 1950.

[19] HO Lancaster. Significance tests in discrete distributions. *Journal of the American Statistical Association*, 56(294):223–234, 1961.

[20] Karim F Hirji, Shu-Jane Tan, and Robert M Elashoff. A quasi-exact test for comparing two binomial proportions. *Statistics in Medicine*, 10(7):1137–1153, 1991.

[21] Graham JG Upton. Fisher's exact test. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 395–402, 1992.

[22] Karl Kieburtz, Andrew Feigin, Michael McDermott, Peter Como, David Abwender, Carol Zimmerman, Charlyne Hickey, Constance Orme, Kathy Claude, Jennie Sotack, et al. A controlled trial of remacemide hydrochloride in huntington's disease. *Movement disorders*, 11(3):273–277, 1996.

[23] G Schifitto, N Sacktor, K Marder, MP McDermott, JC McArthur, K Kieburtz, S Small, LG Epstein, Neurological AIDS Research Consortium, et al. Randomized trial of the platelet-activating factor antagonist lexipafant in hiv–associated cognitive impairment. *Neurology*, 53(2):391–391, 1999.

[24] Hannah M Linden, Brenda F Kurland, Lanell M Peterson, Erin K Schubert, Julie R Gralow, Jennifer M Specht, Georgiana K Ellis, Thomas J Lawton, Robert B Livingston, Philip H Petra, et al. Fluoroestradiol positron emission tomography reveals

differences in pharmacodynamics of aromatase inhibitors, tamoxifen, and fulvestrant in patients with metastatic breast cancer. *Clinical Cancer Research*, 17(14):4799–4805, 2011.

[25] Luc Duchateau and Paul Janssen. Small vaccination experiments with binary outcome: the paradox of increasing power with decreasing sample size and/or increasing imbalance. *Biometrical journal*, 41(5):583–600, 1999.

[26] George A Barnard. A new test for 2× 2 tables. *Nature*, 156:177, 1945.

[27] George A Barnard. Significance tests for 2× 2 tables. *Biometrika*, 34(1/2):123–138, 1947.

[28] George A Barnard. Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):115–149, 1949.

[29] RD Boschloo. Raised conditional level of significance for the 2× 2-table when testing the equality of two probabilities. *Statistica Neerlandica*, 24(1):1–9, 1970.

[30] Devan V Mehrotra, Ivan SF Chan, and Roger L Berger. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics*, 59(2):441–450, 2003.

[31] Roger L Berger and Dennis D Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427): 1012–1016, 1994.

[32] Michael A Proschan and Martha Nason. Conditioning in 2× 2 tables. *Biometrics*, 65(1):316–322, 2009.

[33] European Medicines Agency, Committe for Medicinal Products for Human use. Points to consider on multiplicity issues in clinical trials. 2002.

[34] Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[35] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[36] Eckart Sonnemann. *Allgemeine Lösungen multipler Testprobleme*. Universität Bern. Institut für Mathematische Statistik und Versicherungslehre, 1982.

[37] K Ruben Gabriel. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250, 1969.

[38] Eckart Sonnemann and Helmut Finner. Vollständigkeitssätze für multiple Testprobleme. In *Multiple Hypothesenprüfung/Multiple Hypotheses Testing*, pages 121–135. Springer, 1988.

[39] Richard M Bittman, Joseph P Romano, Carlos Vallarino, and Michael Wolf. Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, 96(2):399–410, 2009. doi: 10.1093/biomet/asp006.

[40] Helmut Finner and Klaus Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, pages 1194–1213, 2002.

[41] Alex Dmitrienko, Ajit C Tamhane, and Frank Bretz. *Multiple testing problems in pharmaceutical statistics*. CRC Press, 2009.

[42] Frank Bretz, Torsten Hothorn, and Peter Westfall. *Multiple comparisons using R*. CRC Press, 2010.

[43] John J Gart, Kenneth C Chu, and Robert E Tarone. Statistical issues in interpretation of chronic bioassay tests for carcinogenicity. *Journal of the National Cancer Institute*, 62(4):957–974, 1979.

[44] Robert E Tarone. A modified Bonferroni method for discrete data. *Biometrics*, pages 515–522, 1990.

[45] Gerhard Hommel and Frank Krummenauer. Improvements and modifications of tarone's multiple test procedure for discrete data. *Biometrics*, 54(2):673–681, 1998. doi: 10.2307/3109773.

[46] Arthur J Roth. Multiple comparison procedures for discrete test statistics. *Journal of statistical planning and inference*, 82(1):101–117, 1999.

[47] Haiyan Xu and Jason C Hsu. Applying the generalized partitioning principle to control the generalized familywise error rate. *Biometrical Journal*, 49(1):52–67, 2007.

[48] Peter H Westfall and James F Troendle. Multiple testing with minimal assumptions. *Biometrical Journal*, 50(5):745–755, 2008.

[49] Violeta Calian, Dongmei Li, and Jason C Hsu. Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical Journal*, 50 (5):756–766, 2008.

[50] Bernhard Klingenberg, Aldo Solari, Luigi Salmaso, and Fortunato Pesarin. Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, 65(2):452–462, 2009.

[51] Peter H Westfall and S Stanley Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley & Sons, 1993.

[52] Peter H Westfall and S Stanley Young. P value adjustments for multiple tests in multivariate binomial models. *Journal of the American Statistical Association*, 84 (407):780–786, 1989.

[53] Roee Gutman and Yosef Hochberg. Improved multiple test procedures for discrete distributions: New ideas and analytical review. *Journal of Statistical Planning and Inference*, 137(7):2380–2393, 2007. doi: 10.1016/j.jspi.2006.08.006.

[54] Jacob Paroush. Integer programming technique to construct statistical tests. *The American Statistician*, 23(5):43–44, 1969.

[55] Reiner Horst and Hoang Tuy. *Global optimization: Deterministic approaches*. Springer, 3rd edition, 1996.

[56] lp_solve reference guide. `http://lpsolve.sourceforge.net/5.5/`, accessed: 18 July 2016.

[57] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. `http://www.R-project.org/`, accessed: 18 July 2016.

[58] Michel Berkelaar and others. *lpSolve: Interface to Lp_solve v. 5.5 to solve linear/integer programs*, 2014. R package version 5.6.10, `http://CRAN.R-project.org/package=lpSolve`, accessed: 18 July 2016.

[59] Jelle J Goeman and Aldo Solari. The sequential rejection principle of familywise error control. *The Annals of Statistics*, pages 3782–3810, 2010.

[60] Paola Lago, Tiziana Bettiol, Sabrina Salvadori, Isabella Pitassi, Andrea Vianello, Lino Chiandetti, and Onofrio S Saia. Safety and efficacy of ibuprofen versus indomethacin in preterm infants treated for patent ductus arteriosus: a randomised controlled trial. *European journal of pediatrics*, 161(4):202–207, 2002.

[61] Friedrich Leisch, Andreas Weingessel, and Kurt Hornik. *bindata: Generation of Artificial Binary Data*, 2012. R package version 0.9-19, `http://CRAN.R-project.org/package=bindata`, accessed: 18 July 2016.

8 Appendix

8.1 Selected R functions to calculate optimal exact tests for binary endpoints

```
#Auxiliary functions
D.fn<-function(ko,sort.it=FALSE) {
        #set keep.empty=TRUE to keep these constraints. This is necessary, if my branch
            and bound algorithm is used, because
        #it uses the n x n constraint matrix as lookup table for cond(ii).
        #constraint matrix for criterion (ii) (no holes)
        if(!is.matrix(ko) & !is.data.frame(ko)) ko<-matrix(ko,ncol=1)
        if(sort.it==TRUE) for(i in 1:dim(ko)[2]) ko<-ko[order(ko[,i]),] #so it also works
             for one-dimensional problems
        n<-dim(ko)[1]
        D.ii<-matrix(FALSE,n,n)
        #if(n > 0) {
        not.empty<-rep(TRUE,n)
        for(i in 1:n) {
                for(j in 1:n) if(all(ko[j,]>=ko[i,])) D.ii[i,j]<- TRUE
        }
        D.ii
}

namen.fn<-function(anz.dim) {
        #create binary numbers from 0 to 2^anz.dim - 1
        x<-0:(2^anz.dim - 1)
        x
        M<-t(sapply(x,dec2bin,digits=anz.dim))
        #sort, so that we get 0,0,0 ; 1,0,0 ; 0,1,0 ; ...
        if(anz.dim>1) out<-M[order(rowSums(M)),] else out<-t(M)  #because in this case
            the t() in sapply is not required
        out
}

dec2bin<-function(x,digits=8) {
        b<-rep(0,digits)
        for(k in 1:digits) {
                b[k]<-x%%2
                x<-(x - b[k])/2
        }
        b
}

#Get array from a vectorized form
vek2array.fn<-function(vek,d,name.shift=0) { #previously shift=-1, but we don't need this
     option anymore
        anz.dim<-length(d)
        dim.namen<-vector(length=anz.dim,mode="list")
        for(i in 1:anz.dim) dim.namen[[i]]<-(1:d[i])+name.shift
        array(vek,dim=d,dimnames=dim.namen)
}

#Find the joint permutation distribution of Fisher exact test statistics
#The main output is in vectorized form, so we can handle arbitrary number of endpoints

#Beliebig viele Dimensionen
mult.fisher.2J.fn<-function(tab=NULL,n=NULL,m=NULL,M=NULL,p1=NULL,p2=NULL,margin.prob=
    FALSE,calc.marg.p=FALSE) {
        #The input is either a Jx2 table with J=2^(number of endopints) via 'tab'
        #or the margins of this table are provided as 'n' (sample sizes in the two groups
            ) and m (marginal number of successes,
        #e.g. for 2 endpoints: no succ., succ. A only, succ. B only, succ. A and B)
        if(!is.null(tab)) {
                m<-rowSums(tab)
                n<-colSums(tab)
        }
        #m...row margins
        #n...column margins (number of observations in the two groups)
```

```
#The matrix M is such that M%*%tab[,1] gives the marginal test statistics and M%*
    %m gives the vector of total number of successes for the endpoints
#By default, it is assumed that m is such that it starts with the number of
    patients without any success,
#followed by the entries for success in exactly one endpoint, then simultaneous
    success in exactly two endpoints and so on.
#so, e.g. for two dimensions M=matrix(c(0,1,0,1,0,0,1,1),byrow=TRUE,nrow=2)
if(is.null(M)) {
        anz.dim<-round(log2(length(m))) #round, just in case that there are
            numerical issues
        M<-t(namen.fn(anz.dim))
} else {
        anz.dim<-dim(M)[1]
}

k<-length(m)
x<-rep(NA,k)
N<-sum(m)
cur.lim<-matrix(NA,nrow=k,ncol=2)
stop<-FALSE
j<-1:k
u<-0
ws<-NULL

#Bound for the test statistics so that we look only at the possible range of
    values for the test statistics
m.marg<-as.numeric(M%*%m) #marginal number of successess for each endpoint
#now use the usual formula for 2x2 tables:
low.marg<-up.marg<-rep(NA,anz.dim)
for(i in 1:anz.dim) {
        #lowest value the i-th test stat can take:
        low.marg[i]<-max(0,m.marg[i]-n[2])
        #largest value the i-th test stat can take:
        up.marg[i]<-min(m.marg[i],n[1]) #maximum
}

d<- up.marg-low.marg + 1 #this is the number of the nonzero elements of the
    marginal distribution in each dimension

#prepare vectorized output
WA.vek<-Z.vek<-F.vek<-rep(0,prod(d))   # array(0,dim=d) #+1 because 0 is in
    general a possible value for the test statistics
cp<-cumprod(d[-length(d)]) #requried to set indices
WS<-NA
WS.vek<-NA
if(margin.prob==TRUE) ws.obs.margins<-0 else ws.obs.margins<-NA #probability, to
    observe a table with the provided margins, this may be used when calculating
    unconditional power

#Going through all permutations
while(!stop) {
        #for i>u: calculate cur.lims anew and set x[i] to the lower limit. x are
            the observed frequencies for group 1 in the current permutation
        if(u+1<=k-1) {
                for(i in (u+1):(k-1)) {
                        x.sum<-ifelse(i==1,0,sum(x[1:(i-1)])) #x sum set so far #
                            sum(c(0,x[0:(i-1)]))
                        m.sum<-sum(m[(i+1):k]) #remaining margin sum
                        low<-max(n[1]-x.sum-m.sum,0) #c(0,x[0])=0
                        up<-min(n[1]-x.sum,m[i])
                        cur.lim[i,]<-c(low,up)
                        x[i]<-low
                }
        }
        x[k]<-n[1]-sum(x[1:(k-1)])
        cur.lim[k,]<-c(x[k],x[k]) #here, only one possibility remains
        #Calculate probabiliy for current permutation
        mgl<-prod(choose(m,x))

        #Current Vector of test statistics
        T.cur<-M%*%x
```

```
        #Current vector of coordinates. The connection between coordinates and
            test statistics is X_i = T_i + 1 - low.marg_i
        #The coordinates start with 1 for all dimensions and are used by
            subsequent functions, e.g. to check condition (ii)
        X.cur<-T.cur + 1 - low.marg

        #Coordinates for vectorized output
        ko<-sum(cp*(X.cur[-1]-1))+X.cur[1]

        Z.vek[ko]<-Z.vek[ko]+mgl

        #WS, Probability under the alternative
        if(!is.null(p1) & !is.null(p2) ) {
                x2<-m-x
                p.u.log<-sum(x*(log(p1)-log(p2)))-sum(lfactorial(x))-sum(
                    lfactorial(x2))
                WA.vek[ko]<-WA.vek[ko]+exp(p.u.log)
                #unconditional probability of these table margins under given
                    alternative p1 and p2:
                if(margin.prob==TRUE) ws.obs.margins<-ws.obs.margins+exp(
                    dmultinom(x=x,prob=p1,log=TRUE)+dmultinom(x=x2,prob=p2,log=
                    TRUE))
        }

        #Stop or next iteration
        if(all(x==cur.lim[,2])) {
                stop<-TRUE
        } else {
                #increase x[i] by 1, where i=max{j=1,...,k : x[j]<cur.lim[j,2]}
                u<-max(j[x<cur.lim[,2]])
                x[u]<-x[u]+1
        }
}

#Distribution under H0
F.vek<-Z.vek/sum(Z.vek)

#Make dimnames, and in the same loop, make the coordinate vectors X1, X2, ... for
     the vectorized output
dim.namen<-vector(length=anz.dim,mode="list")
koord<-matrix(NA,ncol=anz.dim,nrow=prod(d))
colnames(koord)<-paste("X",1:anz.dim,sep="")

#caclulate marginal p-values (one sided), these can be used to find p-value based
     rejection regions
marg.pval<-matrix(NA,ncol=anz.dim,nrow=prod(d))
colnames(marg.pval)<-paste("pval",1:anz.dim,sep="")

for(i in 1:anz.dim) {
        dim.namen[[i]]<-low.marg[i]:up.marg[i]     #0:(d[i]-1)

        #Coordinates, use recycling of vectors by R
        koord[,i]<-rep(1:d[i],each=prod( c(1,d)[1:i] ) )   #c(1,d), so in the
            first dimension each entry is repeated once, in the other dimensions
            the number of repeates is the number of values in the previous
            dimension.

        #Marginal p-values:
        if(calc.marg.p==TRUE) {
                t.temp<-low.marg[i]:up.marg[i]
                p.temp<-phyper(q=t.temp,m=n[1],n=n[2],k=m.marg[i],lower.tail=
                    FALSE)+dhyper(x=t.temp,m=n[1],n=n[2],k=m.marg[i])
                marg.pval[,i]<-rep(p.temp,each=prod( c(1,d)[1:i] ) )
        }
}

#Distributions under H0 and alternative in array form:
#Null distribution
F<-array(F.vek,dim=d,dimnames=dim.namen)
#Joint distribution of T under the alternative p1,p2
if(!is.null(p1) & !is.null(p2) ) {
        WS.vek<-WA.vek/sum(WA.vek)
```

```
                WS<-array(WS.vek,dim=d,dimnames=dim.namen)
        }

        #Vektorized output, the LP requires this type of input anyways and it works for
            arbitrary dimensions

        T<-matrix(NA,nrow=dim(koord)[1],ncol=dim(koord)[2])
        for(i in 1:dim(koord)[2]) T[,i]<-koord[,i] + low.marg[i] -1

        colnames(T)<-paste("T",1:anz.dim,sep="")
        vek.out<-data.frame(koord,T,index=1:prod(d),alpha=F.vek,power=WS.vek,solution=0,E
            =0,marg.pval)

        #if(!is.null(tab)) obs.prob<-prod(choose(m,tab[,1])) else obs.prob<-NA #number of
             Permutations that give the observed table (if provided)
        #list(vek.out=vek.out,d=d,anz.dim=anz.dim,M=M,n=n,m=m,marginal.succ=m.marg,F=F,WS
            =WS,p1=p1,p2=p2,ws.obs.margins=ws.obs.margins,total.perm=total.perm,sumZ=sum(
            Z.vek),sum(Z.vek)==total.perm,dim.out=dim(vek.out),dim.koord=dim(koord),l.F.
            vek=length(F.vek))
        list(vek.out=vek.out,d=d,anz.dim=anz.dim,M=M,n=n,m=m,marginal.succ=m.marg,F=F,WS=
            WS,p1=p1,p2=p2,ws.obs.margins=ws.obs.margins)

}

#Branch and bound function for the paper
bb.fn<-function(alpha.contributions,obj,D,alpha=0.025,maxiter=10000) {
        #alpha.contributions is the vector of contributions to the type I error rate (
            alpha.contributions%*%solution is the actual type I error rate)
        #obj is the vector of contributions to the objective function (obj%*%solution is
            the value of the objective function)
        #D is the look up matrix (see text)

        #set length of the solution vector
        d.x<-length(alpha.contributions)

        presolution<-rep(0,d.x) #if some feasible solution is known it may be used here
            instead of the trivial solution
        max.low<-0

        #the function to add a node:
        index<-1:d.x
        add.fn<-function(n,z) {
                x<-n[-c(1:3)]
                #find first not yet defined entry and if x meets cond. (i) return low, up
                    , branched out, x
                i<-min(index[x == -1])
                #set new entry and all others according to cond(ii)
                if(z==0) {
                        x[D[,i]]<-0
                        return(c(sum(obj[x==1]),sum(obj[x!=0]),all(x>=0),x))
                }
                if(z==1) {
                        x[D[i,]]<-1
                        #check cond(i), this is only required for z=1.
                        if(sum(alpha.contributions[x==1])>alpha) return(NULL) else return
                            (c(sum(obj[x==1]),sum(obj[x!=0]),all(x>=0),x))
                }
        }

        #prepare matrix of nodes
        #the columns are: lower bound, upper bound, fully branched (0/1), and then the
            solution vector
        #for a branched out solution, low=up.
        #for a solution 1 means included, 0 means not included and -1 means not yet
            determined

        N<-matrix(
                c(max.low,max.low,1,presolution,
                0,0,0,rep(-1,d.x)),byrow=TRUE,nrow=2)

        nc<-dim(N)[2]
        active<-N[,3]==0
```

```
        N.act<-matrix(N[active,],ncol=nc)
        N.finished<-matrix(N[!active,],ncol=nc)

        iter<-0L
        while(dim(N.act)[1]>0 & iter<maxiter) {
                iter<-iter+1L
                #select of all not fully branched nodes that with largest lower bound
                #as alternative we might alternatingly also select that with largest
                    upper bound.
                #so they get removed before everything has branched out.
                #In the example, this takes the same number of iterations, but is slower,
                     probably because N can gets larger.
                #index 1 or 2:
                #indi<-1+iter%%2
                indi<-1
                j<-which(N.act[,indi]==max(N.act[,indi]))[1]
                n<-N.act[j,]
                #add new nodes and remove the one just branched (j), rbind(NULL,n) gives
                    a matrix, so it N always stays a matrix
                N<-rbind(N.finished,N.act[-j,],add.fn(n,0),add.fn(n,1))
                #update current best lower bound
                max.low<-max(N[,1])
                #remove all nodes with up<max.low, N can become 0 x (3+d.x) Matrix, but
                    it will not disappear
                N<-matrix(N[N[,2]>=max.low,],ncol=nc)
                #select active nodes
                active<-N[,3]==0
                N.act<-matrix(N[active,],ncol=nc)
                N.finished<-matrix(N[!active,],ncol=nc)
        }
        optimization.finished<-dim(N.act)[1]==0
        if(optimization.finished) {
                solution<-N[1,-c(1:3)]
        } else {
                #else we still provide a feasible solution, namely that with the largest
                    lower bound
                j<-which(N[,1]==max(N[,1]))[1]
                solution<-N[j,-c(1:3)]
                solution[solution==-1]<-0 #this is always feasible because of the filling
                     up in add.fn
        }
        list(solution=solution,obj.value=sum(obj[solution==1]),nodes=N,n.solutions=dim(N)
            [1],iterations=iter,maxiter=maxiter,optimization.finished=optimization.
            finished) #,status=as.numeric(!optimization.finished))
}

##
#just the pre-processing
preprocess.2J.fn<-function(V,anz.dim=sum(grepl("X",names(V))),alpha=0.025,d) {

        #1) remove lines with prob 0
        keep<-V$alpha>0
        X<-V[keep,]

    #check, if X is empty. If so, skip remaining parts and report V.
    if(dim(X)[1]>0) {

        #2) actual pre-processing, cond (i) und (ii)
        K<-dim(X)[1] #number of points (rows)
        #condition (i)
        for(i in 1:K) {
                set<-rep(TRUE,K)
                for(j in 1:anz.dim) set<-set&(X[,j]>=X[i,j]) #the coordinates have to be
                    in the first columns of V
                alpha.min<-sum(X$alpha[set])
                if(alpha.min<=alpha) X$E[i]<-1 #eligible points, according to cond(i)
        }
        #only keep eligible
        X1<-X[X$E==1,]
        K1<-dim(X1)[1]
        max.alpha<-sum(X1$alpha)
        #condition (ii)
```

```
        if(K1>0) {
                for(i in 1:K1) {
                        set<-rep(TRUE,K1)
                        for(j in 1:anz.dim) set<-set&(X1[,j]<=X1[i,j])
                        max.level.ohne.i<-max.alpha-sum(X1$alpha[set])
                        max.level.mit.i<-max.level.ohne.i + X1$alpha[i]
                        if(max.level.mit.i<=alpha) X1$E[i]<-2 #points, which are
                            contained in eversy optimal rej. region
                }
                X2<-X1[X1$E==1,]
        } else {
                X2<-X1
        }
        alpha.frei<-alpha-sum(X1$alpha[X1$E==2])
        X1$E
        V$E[keep]<-X$E
        V$E[keep][X$E==1]<-X1$E
    }
        R<-vek2array.fn(V$E,d=d)
        #plot.muf.fn(F=A$F,region=r.test,gr=1.25,r=1,main="Optimal area",print.area="prob
            .area",gr.label=1,label.line=4,gr.ax=1.25)

        list(V=V,V2=X2,R=R,alpha.frei=alpha.frei)
}

#plot function

#Plot Function, works for two dimensions only
plot.muf.fn<-function(F,r=2,gr=0.75,gr.label=gr,region=NULL,color=1:10,alpha=0.025,farbe=
    c(2,1,1,2),rm.zero.lines=TRUE,print.level=TRUE,label.line=3,font.label=2,font.text=2,
    no.numbers=FALSE,gr.ax=1,transpose=FALSE,subscripts=c(1,2),lwdborder=2,...) {
        if(transpose==TRUE) {
                F<-t(F)
                if(!is.null(region)) region<-t(region)
        }

        if(rm.zero.lines==TRUE) {
                r.zero<-rowSums(F)>0
                c.zero<-colSums(F)>0
                F<-F[r.zero,c.zero]
                if(!is.null(region)) region<-region[r.zero,c.zero]
        }
        M<-addmargins(F)
        #remove the sum "100" in the uppermost corner
        M[dim(M)[1],dim(M)[2]]<-0
        x<-1:(dim(M)[1])
        y<-1:(dim(M)[2])

        if(!is.null(region)) region[F==0]<-0

        plot.new()
        plot.window(xlim=c(min(x),max(x)),ylim=c(min(y),max(y)))
        title(xlab=bquote(T[.(subscripts[1])]), ylab=bquote(T[.(subscripts[2])]),...)

        abline(v=x-0.5,lwd=c(rep(1,length(x)-1),lwdborder))
        abline(h=y-0.5,lwd=c(rep(1,length(y)-1),lwdborder))

        rownames(M)[rownames(M)=="Sum"]<-""
        colnames(M)[colnames(M)=="Sum"]<-""

        axis(1,at=x,labels=rownames(M),tcl=0,lwd=0,cex.axis=gr.ax)
        axis(2,at=y,labels=colnames(M),tcl=0,lwd=0,las=2,cex.axis=gr.ax)

        for(i in 1:length(x)) {
                for(j in 1:length(y)) {
                        if(!is.null(region) & i!=length(x) & j!=length(y)) text.farbe<-
                            color[region[i,j]+1] else text.farbe<-farbe[1]
                        if(i==length(x) | j==length(y)) text.farbe<-farbe[2] #margin
                            Farbe
                        if(M[i,j]!=0) text(x[i],y[j],round(M[i,j]*100,r),cex=gr,col=text.
                            farbe,font=font.text)
                }
```

```
        }

        if(!is.null(region) & print.level==TRUE) {
                level<-sum(F[region>0])
                mtext(side=1,text=paste("P(R) = ",round(level*100,3),"%,","|R| = ",sum(
                    region>0)),adj=0,line=label.line,cex=gr.label,col=farbe[3],font=font.
                    label)
        }


}

##
dhyper.nc<-function(n1.,n.1,n..,OR=1,dhyp=TRUE) {
        n2.<-n..-n1.
        n.2<-n..-n.1
        p.fn<-function(x) choose(n.1,x)*choose(n.2,n1.-x)*OR^x
        start<-max(0,n.1-n2.)
        ende<-min(n1.,n.1)
        if(OR==1 & dhyp==TRUE) {
                prob<-dhyper(x=start:ende,m=n1.,n=n..-n1.,k=n.1) #more accurate, because
                    for OR=1 the sum is known and the calculation can be done without
                    multiplying and dividing huge numbers
        } else {
                s<-sapply(start:ende,p.fn)
                S<-sum(s)
                prob=s/S
        }

        out<-data.frame(t=start:ende,prob=prob,F=cumsum(prob),F.rev=rev(cumsum(rev(prob))
            ))
        #names(out)<-start:ende
        out
}

#This function finds Bonferroni adjusted critical boundaries for one-sided Fisher's exact
     test using a greedy algorithm.
#The critical boundaries are such that T>=c means rejection
#In the output distribution tables, alpha=P_H0(T=t), P=P_H0(T>=t), Q=P_HA(T>=t)

marg.fisher.greedy.fn<-function(m.list,n.list,OR.0.vek=NULL,OR.1.vek=NULL,alpha=0.025) {
        #m.list ... list of row margins of the k 2x2 tables, the first entry is the
            number of successes. Large values are in contradiction to H0.
        #n.list ... list of column margins, i.e. the sample sizes in the two groups.
        #OR.0.vek ... a vector of the odds ratios lambda_i to specify H_i: OR <= lambda_i
            .
        #OR.1.vek ... a vector of the odds ratios lambda'_i under the alternative. Used
            to calculate the local marginal power.
        #alpha ... The global significance level
        k<-length(m.list)

        if(is.null(OR.0.vek)) OR.0.vek<-rep(1,k)
        if(is.null(OR.1.vek)) OR.1.vek<-rep(1,k)

        prob<-vector(length=k,mode="list")
        names(prob)<-paste("Test",1:k,sep=".")
        #d<-d.total<-rep(NA,k)
        d<-rep(NA,k)
        #d.total is the number of values the test stat can take, while d is the number of
             values that are actually considered after pre-processing
        #prepare information for one Fisher test for a 2x2 table:

        for(i in 1:k) {
                m<-m.list[[i]]
                n<-n.list[[i]]
                N<-sum(n)
                if(N!=sum(m)) stop("Row and colum margins do not have the same sum.")
                prob.0<-dhyper.nc(n1. = m[1],n.1 = n[1],n.. = N,OR=OR.0.vek[i])
                prob.1<-dhyper.nc(n1. = m[1],n.1 = n[1],n.. = N,OR=OR.1.vek[i])

                #prepare the distribution tables
                prob.i<-data.frame(t=prob.0$t,alpha=prob.0$prob,P=prob.0$F.rev,Q=prob.1$F
```

```
                    .rev,EP=i) #cum prob under H0: P(T>=t), under H1: Q(T>=t)
                prob.i<-rbind(prob.i,c(Inf,0,0,0,i))
                prob[[i]]<-prob.i
                #number of points to consider for each test stat:
                d[i]<-dim(prob.i)[1]

        }

        pr<-rep(1,k)
        ind.c<-d #index of critical boundaries c, c is initiated at Inf
        loc.level<-rep(0,k)
        stop<-FALSE
        #e<-rep(NA,k)
        while(!stop) {
                for(i in 1:k) pr[i]<-prob[[i]]$alpha[ind.c[i]-1]
                j<-which(pr==min(pr))[1] #selected test
                ind.c[j]<-ind.c[j]-1
                for(i in 1:k) loc.level[i]<-prob[[i]]$P[ind.c[i]]

                if(sum(loc.level)>alpha) {
                        ind.c[j]<-ind.c[j]+1 #reverse the last step
                        stop<-TRUE
                }
        }

        crit.b<-rep(Inf,k)
        #marginal alpha and power
        alpha.i<-power.i<-rep(0,k)
        names(crit.b)<-names(alpha.i)<-names(power.i)<-names(prob)

        for(i in 1:k) {
                crit.b[i]<-prob[[i]]$t[ind.c[i]]
                alpha.i[i]<-prob[[i]]$P[ind.c[i]]
                power.i[i]<-prob[[i]]$Q[ind.c[i]]
                prob[[i]]$R<-as.numeric(prob[[i]]$t>=crit.b[i])
        }

        list(distributions=prob,crit.boundaries=crit.b,
                alpha.local=alpha.i,alpha.sum=sum(alpha.i),power.local=power.i)
                #disj.power.ind=1-prod(1-power.i),conj.power.ind=prod(power.i))
}
```

8.2 Exemplary analysis commands using R

```
tab1<-matrix(c(0,13,1,80,2,12,10,57),ncol=2) #Example ductus arteriosa
rownames(tab1)<-c("no␣succ.","succ.␣urine␣only","succ.␣duct.␣only","succ.␣both")
colnames(tab1)<-c("trt","ctrl")

P<-data.frame(p1=c(0.01,0.14,0.02,0.83),p2=c(0.04,0.15,0.13,0.68))
rownames(P)<-rownames(tab1)

alpha<-0.025

#Optimal test using the joint distribution

#Find permutation distribution
A<-mult.fisher.2J.fn(tab1,p1=P$p1,p2=P$p2,margin.prob=TRUE,calc.marg.p=TRUE)

#Pre-processing
pp<-preprocess.2J.fn(A$vek.out,d=A$d)

#Add solution points found in pre-processing
pp$V$solution[pp$V$E==2]<-1

#Make lookup matrix
D.mat<-D.fn(ko=pp$V2[,1:2])

#Find optimal solution with respect to power
OPT<-bb.fn(alpha.contributions=pp$V2$alpha,obj=pp$V2$power,D=D.mat,alpha=pp$alpha.frei,
    maxiter=10000)
```

```
#Add solution to the solution vector in V
pp$V$solution[pp$V2$index[OPT$solution==1]]<-1

#make array
R<-vek2array.fn(pp$V$solution,d=A$d)

#plot
subscr<-c("urine","duct.")
plot.muf.fn(F=A$F,region=R,subscripts=subscr,main="Example␣under␣H0,␣optimal␣power")
plot.muf.fn(F=A$WS,region=R,subscripts=subscr,main="Example␣under␣HA,␣optimal␣power")

#Bonferroni test optimized by greedy algorithm

#marginal tables
M<-t(namen.fn(2))
urine<-rbind(A$M%*%tab1,(1-A$M)%*%tab1)[c(1,3),]
duct<-rbind(A$M%*%tab1,(1-A$M)%*%tab1)[c(2,4),]
rownames(urine)<-rownames(duct)<-c("succ.","no␣succ.")
#Note: success must be the first row for the greedy optimization function,
#in contrast to the joint distr function. This can easily be reversed in the function.

Bonf.greedy<-marg.fisher.greedy.fn(m.list=list(rowSums(urine),rowSums(duct)),
        n.list=list(colSums(urine),colSums(duct)),
        OR.0.vek=NULL,OR.1.vek=c(6.62,1.33),alpha=0.025)

Bonf.greedy
```

8.3 Simulation results for two endpoints

The appendix tables A1 to A19 show the power for the different closed testing procedures when comparing two treatment groups with respect to two binary endpoints. The tables show the probabilities in percent that the closed testing procedure rejects the global intersection hypothesis $H_1 \cap H_2$, at least one elementary null hypothesis (H_1 or H_2), both elementary null hypothesis (H_1 and H_2), or particularly H_1 or H_2.

The branch and bound algorithm was used with a maximal number of 20000 iterations. In the table headings, n is the sample size per group, α the nominal global significance level, $p_i^{(Trt)}$ and $p_i^{(Ctr)}$ the success probability in endpoint i in the treatment and control group, respectively, and ρ the common product moment correlation between the endpoints.

The number of instances in which a confirmed optimal solution was found or not found by the branch and bound algorithm, respectively, is given in the two columns labelled C (optimization complete) and NC (not complete). To further quantify the number of required iterations, the median (q50), the 90 % quantile (q90) and the maximum (Max) of the number of iterations are shown in the last three columns. A value of 20000 indicates that the respective quantity to achieve an optimal solution would be above 20000, but is here bounded by the maximal number of iterations. A value of 0 means that the optimization was finished by the pre-processing.

For tests optimizing the power, the assumed alternative is indicated in brackets next to the test label. There, “all” and “EP 1” refer to an assumed alternative identical to the true alterative in a scenario with an effect in both endpoints or one endpoint, respectively. ρ there indicates the assumed correlation between the endpoints.

Table A1: Power with two binary endpoints, $n = 5$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.865$, $p_2^{(Trt)} = 0.135$, $p_1^{(Ctr)} = 0.135$, $p_2^{(Ctr)} = 0.135$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	23.5	23.5	0	23.5	0					
Tarone	58.8	58.8	0	58.8	0					
HKT	58.8	58.8	0	58.8	0					
Bonferroni optimal alpha	59.4	59.4	0	59.4	0.1					
Bonferroni optimal power (all)	59.4	59.4	0	59.4	0.1					
Bonferroni optimal power (1)	60.1	60.1	0	60.1	0.1					
Bonferroni greedy algorithm	59.3	59.3	0	59.3	0.1					
minP	58.8	58.8	0	58.8	0					
Optimal alpha	60.2	58.4	0	58.3	0.1	286	0	0	3	8
Optimal area	61.8	60.1	0	60.1	0.1	286	0	0	3	8
Optimal power (all, $\rho = 0$)	57	55.2	0	55.1	0.1	286	0	0	3	8
Optimal power (all, $\rho = 0.5$)	57	55.2	0	55.1	0.1	286	0	0	2	4
Optimal power (EP 1, $\rho = 0$)	61.9	60.1	0	60.1	0.1	286	0	0	3	8
Consonant optimal alpha	59.4	59.4	0	59.4	0.1	286	0	0	3	7
Consonnat optimal area	60.1	60.1	0	60.1	0.1	286	0	0	2	7
Consonant optimal power (all, $\rho = 0$)	59.6	59.6	0	59.6	0.1	286	0	0	2	7
Consonant optimal power (all, $\rho = 0.5$)	59.6	59.6	0	59.6	0.1	286	0	0	2	3
Consonant optimal power (EP 1, $\rho = 0$)	60.1	60.1	0	60.1	0.1	286	0	0	2	7
Greedy algorithm	60.9	59.2	0	59.1	0.1					

Table A2: Power with two binary endpoints, $n = 5$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.865$, $p_2^{(Trt)} = 0.865$, $p_1^{(Ctr)} = 0.135$, $p_2^{(Ctr)} = 0.135$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	41.4	41.4	22.7	32	32					
Tarone	52.3	52.3	22.7	37.5	37.5					
HKT	52.3	52.3	22.7	37.5	37.5					
Bonferroni optimal alpha	76.2	76.2	36.1	57.9	54.4					
Bonferroni optimal power (all)	76.2	76.2	36.1	57.9	54.4					
Bonferroni optimal power (1)	72.7	72.7	36.1	60.1	48.7					
Bonferroni greedy algorithm	69.1	69.1	36.1	53.4	51.7					
minP	55.7	55.7	24	39.9	39.9					
Optimal alpha	81.5	75.9	36.1	57.8	54.1	286	0	0	3	8
Optimal area	83.3	78.2	36.1	60.1	54.2	286	0	0	3	8
Optimal power (all, $\rho = 0$)	88.3	81.2	36.1	58.9	58.3	286	0	0	3	8
Optimal power (all, $\rho = 0.5$)	88.3	81.2	36.1	58.9	58.3	286	0	0	2	4
Optimal power (EP 1, $\rho = 0$)	83.4	78	36.1	60.1	54	286	0	0	3	8
Consonant optimal alpha	76.2	76.2	36.1	57.9	54.4	286	0	0	3	7
Consonnat optimal area	78.2	78.2	36.1	60.1	54.2	286	0	0	2	7
Consonant optimal power (all, $\rho = 0$)	82.4	82.4	36.1	59.6	58.9	286	0	0	2	7
Consonant optimal power (all, $\rho = 0.5$)	82.4	82.4	36.1	59.6	58.9	286	0	0	2	3
Consonant optimal power (EP 1, $\rho = 0$)	78.2	78.2	36.1	60.1	54.2	286	0	0	2	7
Greedy algorithm	86.3	81.1	36.1	59.8	57.4					

Table A3: Power with two binary endpoints, $n = 5$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.865$, $p_2^{(Trt)} = 0.865$, $p_1^{(Ctr)} = 0.135$, $p_2^{(Ctr)} = 0.135$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	35.2	35.2	28.5	31.9	31.9					
Tarone	44.3	44.3	28.5	36.4	36.4					
HKT	44.3	44.3	28.5	36.4	36.4					
Bonferroni optimal alpha	72.1	72.1	45.1	59.6	57.6					
Bonferroni optimal power (all)	72.1	72.1	45.1	59.6	57.6					
Bonferroni optimal power (1)	67.5	67.5	45.1	60.1	52.5					
Bonferroni greedy algorithm	62.8	62.8	45.1	54.6	53.3					
minP	63.4	63.4	39.6	51.5	51.5					
Optimal alpha	73.3	72.1	45.1	59.6	57.6	286	0	0	3	8
Optimal area	73.6	72.4	45.1	60.1	57.4	286	0	0	3	8
Optimal power (all, $\rho = 0$)	74.1	72.8	45.1	59.5	58.4	286	0	0	3	8
Optimal power (all, $\rho = 0.5$)	74.1	72.8	45.1	59.5	58.4	286	0	0	2	4
Optimal power (EP 1, $\rho = 0$)	73.6	72.4	45.1	60.1	57.4	286	0	0	3	8
Consonant optimal alpha	72.1	72.1	45.1	59.6	57.6	286	0	0	3	7
Consonnat optimal area	72.4	72.4	45.1	60.1	57.4	286	0	0	2	7
Consonant optimal power (all, $\rho = 0$)	72.9	72.9	45.1	59.6	58.4	286	0	0	2	7
Consonant optimal power (all, $\rho = 0.5$)	72.9	72.9	45.1	59.6	58.4	286	0	0	2	3
Consonant optimal power (EP 1, $\rho = 0$)	72.4	72.4	45.1	60.1	57.4	286	0	0	2	7
Greedy algorithm	74	72.7	45.1	59.9	58					

Table A4: Power with two binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.792$, $p_2^{(Trt)} = 0.208$, $p_1^{(Ctr)} = 0.208$, $p_2^{(Ctr)} = 0.208$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	59.5	59.5	0.2	59.5	0.3					
Tarone	59.9	59.9	0.2	59.8	0.3					
HKT	59.9	59.9	0.2	59.8	0.3					
Bonferroni optimal alpha	56.6	56.6	0.2	56.5	0.4					
Bonferroni optimal power (all)	60.2	60.2	0.2	60	0.4					
Bonferroni optimal power (1)	60.2	60.2	0.2	60.1	0.4					
Bonferroni greedy algorithm	60.2	60.2	0.2	60	0.4					
minP	60.1	60.1	0.2	59.9	0.4					
Optimal alpha	56.9	49.4	0.2	49.3	0.4	1771	0	4	25	81
Optimal area	67.1	60	0.2	59.8	0.4	1771	0	4	24	78
Optimal power (all, $\rho = 0$)	57.7	50.7	0.2	50.5	0.4	1771	0	3	42	186
Optimal power (all, $\rho = 0.5$)	57.6	50.5	0.2	50.4	0.4	1771	0	2	8	22
Optimal power (EP 1, $\rho = 0$)	70	60.2	0.2	60	0.4	1771	0	3	58	275
Consonant optimal alpha	60	60	0.2	59.8	0.4	1771	0	0	0	18
Consonnat optimal area	60	60	0.2	59.8	0.4	1771	0	0	0	18
Consonant optimal power (all, $\rho = 0$)	60	60	0.2	59.9	0.4	1771	0	0	0	21
Consonant optimal power (all, $\rho = 0.5$)	60	60	0.2	59.9	0.4	1771	0	0	0	10
Consonant optimal power (EP 1, $\rho = 0$)	60.2	60.2	0.2	60.1	0.4	1771	0	0	0	21
Greedy algorithm	65.7	60.2	0.2	60	0.4					

Table A5: Power with two binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.792$, $p_2^{(Trt)} = 0.792$, $p_1^{(Ctr)} = 0.208$, $p_2^{(Ctr)} = 0.208$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	83.6	83.6	36.1	59.8	59.8					
Tarone	83.6	83.6	36.1	59.8	59.8					
HKT	83.6	83.6	36.1	59.8	59.8					
Bonferroni optimal alpha	83.9	83.9	36.1	60	60					
Bonferroni optimal power (all)	83.9	83.9	36.1	60	60					
Bonferroni optimal power (1)	83.9	83.9	36.1	60.1	59.9					
Bonferroni greedy algorithm	83.9	83.9	36.1	60	60					
minP	83.9	83.9	36.1	60	60					
Optimal alpha	94.3	82.6	36.1	59.5	59.2	1771	0	4	25	81
Optimal area	94.3	84	36.1	60.1	60	1771	0	4	24	78
Optimal power (all, $\rho = 0$)	96.2	83.7	36.1	59.9	59.9	1771	0	3	42	186
Optimal power (all, $\rho = 0.5$)	96.2	83.7	36.1	59.9	59.9	1771	0	2	8	22
Optimal power (EP 1, $\rho = 0$)	94.2	83	36.1	60.1	59	1771	0	3	58	275
Consonant optimal alpha	84	84	36.1	60.1	60.1	1771	0	0	0	18
Consonnat optimal area	84	84	36.1	60.1	60.1	1771	0	0	0	18
Consonant optimal power (all, $\rho = 0$)	84	84	36.1	60.1	60.1	1771	0	0	0	21
Consonant optimal power (all, $\rho = 0.5$)	84	84	36.1	60.1	60.1	1771	0	0	0	10
Consonant optimal power (EP 1, $\rho = 0$)	84	84	36.1	60.1	60.1	1771	0	0	0	21
Greedy algorithm	94.4	84	36.1	60.1	60.1					

Table A6: Power with two binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.792$, $p_2^{(Trt)} = 0.792$, $p_1^{(Ctr)} = 0.208$, $p_2^{(Ctr)} = 0.208$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	75.2	75.2	44.3	59.8	59.8					
Tarone	75.2	75.2	44.3	59.8	59.8					
HKT	75.2	75.2	44.3	59.8	59.8					
Bonferroni optimal alpha	75.7	75.7	44.3	60	60					
Bonferroni optimal power (all)	75.7	75.7	44.3	60	60					
Bonferroni optimal power (1)	75.7	75.7	44.3	60.1	59.9					
Bonferroni greedy algorithm	75.7	75.7	44.3	60	60					
minP	75.7	75.7	44.3	60	60					
Optimal alpha	84.4	74.5	44.3	59.4	59.4	1771	0	4	25	81
Optimal area	83.7	75.4	44.3	60.1	59.7	1771	0	4	24	78
Optimal power (all, $\rho = 0$)	84.6	74.9	44.3	59.6	59.6	1771	0	3	42	186
Optimal power (all, $\rho = 0.5$)	84.6	74.9	44.3	59.6	59.6	1771	0	2	8	22
Optimal power (EP 1, $\rho = 0$)	83.6	75.2	44.3	60.1	59.5	1771	0	3	58	275
Consonant optimal alpha	75.8	75.8	44.3	60.1	60	1771	0	0	0	18
Consonnat optimal area	75.8	75.8	44.3	60.1	60	1771	0	0	0	18
Consonant optimal power (all, $\rho = 0$)	75.8	75.8	44.3	60.1	60	1771	0	0	0	21
Consonant optimal power (all, $\rho = 0.5$)	75.8	75.8	44.3	60.1	60	1771	0	0	0	10
Consonant optimal power (EP 1, $\rho = 0$)	75.8	75.8	44.3	60.1	60	1771	0	0	0	21
Greedy algorithm	83	75.8	44.3	60.1	60					

Table A7: Power with two binary endpoints, $n = 15$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.265$, $p_2^{(Trt)} = 0.265$, $p_1^{(Ctr)} = 0.265$, $p_2^{(Ctr)} = 0.265$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	0.8	0.8	0.01	0.4	0.4					
Tarone	0.87	0.87	0.01	0.44	0.44					
HKT	0.92	0.92	0.01	0.46	0.46					
Bonferroni optimal alpha	1.53	1.53	0.01	0.8	0.74					
Bonferroni optimal power (all)	1.49	1.49	0.01	0.74	0.76					
Bonferroni optimal power (1)	1.45	1.45	0.01	0.88	0.58					
Bonferroni greedy algorithm	1.38	1.38	0.01	0.72	0.66					
minP	1.32	1.32	0.01	0.66	0.66					
Optimal alpha	2.44	1.03	0.01	0.54	0.49	5456	0	17	137	889
Optimal area	2.21	1.55	0.01	0.86	0.71	5456	0	17	136	874
Optimal power (all, $\rho = 0$)	2.36	0.88	0.01	0.45	0.43	5456	0	37	649	4438
Optimal power (all, $\rho = 0.5$)	2.37	0.86	0.01	0.45	0.42	5456	0	8	43	143
Optimal power (EP 1, $\rho = 0$)	2.33	1.18	0.01	0.87	0.32	5456	0	43	1765	14350
Consonant optimal alpha	1.61	1.61	0.01	0.82	0.8	5456	0	0	7	68
Consonnat optimal area	1.59	1.59	0.01	0.86	0.73	5456	0	0	7	68
Consonant optimal power (all, $\rho = 0$)	1.6	1.6	0.01	0.81	0.8	5456	0	0	7	226
Consonant optimal power (all, $\rho = 0.5$)	1.6	1.6	0.01	0.81	0.8	5456	0	0	3	45
Consonant optimal power (EP 1, $\rho = 0$)	1.59	1.59	0.01	0.88	0.72	5456	0	0	7	242
Greedy algorithm	2.12	1.47	0.01	0.76	0.72					

Table A8: Power with two binary endpoints, $n = 15$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.265$, $p_2^{(Trt)} = 0.265$, $p_1^{(Ctr)} = 0.265$, $p_2^{(Ctr)} = 0.265$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	0.77	0.77	0.08	0.43	0.43					
Tarone	0.85	0.85	0.08	0.47	0.47					
HKT	0.89	0.89	0.09	0.49	0.49					
Bonferroni optimal alpha	1.47	1.47	0.1	0.82	0.75					
Bonferroni optimal power (all)	1.45	1.45	0.1	0.77	0.79					
Bonferroni optimal power (1)	1.41	1.41	0.1	0.88	0.63					
Bonferroni greedy algorithm	1.34	1.34	0.1	0.76	0.69					
minP	1.33	1.33	0.09	0.71	0.71					
Optimal alpha	2.21	1.29	0.1	0.72	0.68	5456	0	17	137	889
Optimal area	2.04	1.52	0.1	0.86	0.76	5456	0	17	136	874
Optimal power (all, $\rho = 0$)	2.14	1.23	0.1	0.68	0.65	5456	0	37	649	4438
Optimal power (all, $\rho = 0.5$)	2.15	1.24	0.1	0.69	0.66	5456	0	8	43	143
Optimal power (EP 1, $\rho = 0$)	2.1	1.33	0.1	0.88	0.56	5456	0	43	1765	14350
Consonant optimal alpha	1.54	1.54	0.1	0.83	0.81	5456	0	0	7	68
Consonnat optimal area	1.53	1.53	0.1	0.86	0.77	5456	0	0	7	68
Consonant optimal power (all, $\rho = 0$)	1.53	1.53	0.1	0.83	0.81	5456	0	0	7	226
Consonant optimal power (all, $\rho = 0.5$)	1.53	1.53	0.1	0.83	0.81	5456	0	0	3	45
Consonant optimal power (EP 1, $\rho = 0$)	1.52	1.52	0.1	0.88	0.75	5456	0	0	7	242
Greedy algorithm	1.96	1.51	0.1	0.83	0.78					

Table A9: Power with two binary endpoints, $n = 15$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.735$, $p_2^{(Trt)} = 0.265$, $p_1^{(Ctr)} = 0.265$, $p_2^{(Ctr)} = 0.265$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	47.6	47.6	0.5	47.5	0.6					
Tarone	48.5	48.5	0.5	48.4	0.6					
HKT	49.6	49.6	0.5	49.5	0.7					
Bonferroni optimal alpha	56	56	0.5	55.7	0.9					
Bonferroni optimal power (all)	59.3	59.3	0.5	59	0.8					
Bonferroni optimal power (1)	60.6	60.6	0.5	60.4	0.8					
Bonferroni greedy algorithm	60.5	60.5	0.5	60.2	0.8					
minP	60.3	60.3	0.5	60.1	0.8					
Optimal alpha	47	40.6	0.5	40.3	0.9	5456	0	17	137	889
Optimal area	63.8	58.9	0.5	58.6	0.9	5456	0	17	136	874
Optimal power (all, $\rho = 0$)	53.6	46.9	0.5	46.6	0.9	5456	0	37	649	4438
Optimal power (all, $\rho = 0.5$)	52.8	46	0.5	45.7	0.9	5456	0	8	43	143
Optimal power (EP 1, $\rho = 0$)	70.9	60.4	0.5	60.1	0.8	5456	0	43	1765	14350
Consonant optimal alpha	58.9	58.9	0.5	58.5	0.9	5456	0	0	7	68
Consonnat optimal area	58.8	58.8	0.5	58.4	0.9	5456	0	0	7	68
Consonant optimal power (all, $\rho = 0$)	60.2	60.2	0.5	59.9	0.9	5456	0	0	7	226
Consonant optimal power (all, $\rho = 0.5$)	60.2	60.2	0.5	59.9	0.9	5456	0	0	3	45
Consonant optimal power (EP 1, $\rho = 0$)	60.7	60.7	0.5	60.4	0.9	5456	0	0	7	242
Greedy algorithm	64.2	60.4	0.5	60	0.9					

Table A10: Power with two binary endpoints, $n = 15$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.735$, $p_2^{(Trt)} = 0.735$, $p_1^{(Ctr)} = 0.265$, $p_2^{(Ctr)} = 0.265$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	72.3	72.3	34.8	53.6	53.6					
Tarone	72.3	72.3	34.8	53.6	53.6					
HKT	72.3	72.3	34.8	53.6	53.6					
Bonferroni optimal alpha	82.7	82.7	36.5	60	59.2					
Bonferroni optimal power (all)	82.7	82.7	36.5	60	59.2					
Bonferroni optimal power (1)	82.7	82.7	36.5	60.4	58.8					
Bonferroni greedy algorithm	82.7	82.7	36.5	59.8	59.3					
minP	81.5	81.5	35.9	58.8	58.6					
Optimal alpha	92.6	82.3	36.5	59.5	59.3	5456	0	17	137	889
Optimal area	93	84.3	36.5	60.4	60.4	5456	0	17	136	874
Optimal power (all, $\rho = 0$)	95.7	83.9	36.5	60.2	60.1	5456	0	37	649	4438
Optimal power (all, $\rho = 0.5$)	95.7	83.8	36.5	60.2	60.1	5456	0	8	43	143
Optimal power (EP 1, $\rho = 0$)	92.2	82.4	36.5	60.4	58.5	5456	0	43	1765	14350
Consonant optimal alpha	84.3	84.3	36.5	60.4	60.4	5456	0	0	7	68
Consonnat optimal area	84.3	84.3	36.5	60.4	60.4	5456	0	0	7	68
Consonant optimal power (all, $\rho = 0$)	84.3	84.3	36.5	60.4	60.4	5456	0	0	7	226
Consonant optimal power (all, $\rho = 0.5$)	84.3	84.3	36.5	60.4	60.4	5456	0	0	3	45
Consonant optimal power (EP 1, $\rho = 0$)	84.3	84.3	36.5	60.4	60.4	5456	0	0	7	242
Greedy algorithm	93.2	84.3	36.5	60.4	60.4					

Table A11: Power with two binary endpoints, $n = 15$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.735$, $p_2^{(Trt)} = 0.735$, $p_1^{(Ctr)} = 0.265$, $p_2^{(Ctr)} = 0.265$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	64	64	42.5	53.3	53.3					
Tarone	64	64	42.5	53.3	53.3					
HKT	64	64	42.5	53.3	53.3					
Bonferroni optimal alpha	74.7	74.7	44.6	60	59.2					
Bonferroni optimal power (all)	74.7	74.7	44.6	60	59.2					
Bonferroni optimal power (1)	74.6	74.6	44.6	60.4	58.8					
Bonferroni greedy algorithm	74.6	74.6	44.6	59.9	59.3					
minP	75.7	75.7	44.5	60.1	60					
Optimal alpha	83.6	74.3	44.6	59.5	59.4	5456	0	17	137	889
Optimal area	83.4	76.1	44.6	60.4	60.2	5456	0	17	136	874
Optimal power (all, $\rho = 0$)	84.6	75.4	44.6	60	59.9	5456	0	37	649	4438
Optimal power (all, $\rho = 0.5$)	84.6	75.4	44.6	60	59.9	5456	0	8	43	143
Optimal power (EP 1, $\rho = 0$)	83.2	75	44.6	60.4	59.2	5456	0	43	1765	14350
Consonant optimal alpha	76.2	76.2	44.6	60.4	60.4	5456	0	0	7	68
Consonnat optimal area	76.2	76.2	44.6	60.4	60.4	5456	0	0	7	68
Consonant optimal power (all, $\rho = 0$)	76.2	76.2	44.6	60.4	60.4	5456	0	0	7	226
Consonant optimal power (all, $\rho = 0.5$)	76.2	76.2	44.6	60.4	60.4	5456	0	0	3	45
Consonant optimal power (EP 1, $\rho = 0$)	76.2	76.2	44.6	60.4	60.4	5456	0	0	7	242
Greedy algorithm	82.9	76.2	44.6	60.4	60.4					

Table A12: Power with two binary endpoints, $n = 15$, $\alpha = 0.05$, $p_1^{(Trt)} = 0.297$, $p_2^{(Trt)} = 0.297$, $p_1^{(Ctr)} = 0.297$, $p_2^{(Ctr)} = 0.297$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	1.7	1.7	0.03	0.86	0.86					
Tarone	1.72	1.72	0.03	0.87	0.87					
HKT	1.77	1.77	0.03	0.9	0.9					
Bonferroni optimal alpha	3.4	3.4	0.04	1.78	1.66					
Bonferroni optimal power (all)	3.3	3.3	0.04	1.67	1.67					
Bonferroni optimal power (1)	3.18	3.18	0.04	2	1.22					
Bonferroni greedy algorithm	2.96	2.96	0.04	1.56	1.44					
minP	2.94	2.94	0.04	1.49	1.49					
Optimal alpha	4.96	2.39	0.04	1.25	1.18	5456	0	22	160.5	1002
Optimal area	4.53	3.5	0.04	1.9	1.64	5456	0	22	156.5	998
Optimal power (all, $\rho = 0$)	4.85	2.15	0.04	1.11	1.08	5456	0	55	927	8955
Optimal power (all, $\rho = 0.5$)	4.86	2.17	0.04	1.12	1.09	5456	0	13	66	241
Optimal power (EP 1, $\rho = 0$)	4.78	2.68	0.04	1.99	0.73	5454	2	61	2446	20000
Consonant optimal alpha	3.65	3.65	0.04	1.87	1.83	5456	0	0	33	206
Consonnat optimal area	3.57	3.57	0.04	1.91	1.7	5456	0	0	33	206
Consonant optimal power (all, $\rho = 0$)	3.61	3.61	0.04	1.83	1.82	5456	0	0	45	1008
Consonant optimal power (all, $\rho = 0.5$)	3.61	3.61	0.04	1.83	1.82	5456	0	0	8	71
Consonant optimal power (EP 1, $\rho = 0$)	3.59	3.59	0.04	2	1.63	5456	0	0	52.5	1122
Greedy algorithm	4.36	3.37	0.04	1.73	1.68					

Table A13: Power with two binary endpoints, $n = 15$, $\alpha = 0.05$, $p_1^{(Trt)} = 0.297$, $p_2^{(Trt)} = 0.297$, $p_1^{(Ctr)} = 0.297$, $p_2^{(Ctr)} = 0.297$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	1.6	1.6	0.25	0.93	0.93					
Tarone	1.63	1.63	0.25	0.94	0.94					
HKT	1.67	1.67	0.26	0.96	0.96					
Bonferroni optimal alpha	3.23	3.23	0.33	1.85	1.71					
Bonferroni optimal power (all)	3.18	3.18	0.33	1.76	1.75					
Bonferroni optimal power (1)	3.05	3.05	0.33	2	1.38					
Bonferroni greedy algorithm	2.86	2.86	0.33	1.67	1.53					
minP	2.94	2.94	0.31	1.62	1.62					
Optimal alpha	4.67	2.88	0.33	1.64	1.58	5456	0	22	160.5	1002
Optimal area	4.25	3.37	0.33	1.98	1.72	5456	0	22	156.5	998
Optimal power (all, $\rho = 0$)	4.54	2.82	0.33	1.59	1.56	5456	0	55	927	8955
Optimal power (all, $\rho = 0.5$)	4.57	2.82	0.33	1.59	1.56	5456	0	13	66	241
Optimal power (EP 1, $\rho = 0$)	4.43	2.98	0.33	2	1.32	5454	2	61	2446	20000
Consonant optimal alpha	3.45	3.45	0.33	1.92	1.87	5456	0	0	33	206
Consonnat optimal area	3.39	3.39	0.33	1.98	1.75	5456	0	0	33	206
Consonant optimal power (all, $\rho = 0$)	3.42	3.42	0.33	1.88	1.86	5456	0	0	45	1008
Consonant optimal power (all, $\rho = 0.5$)	3.42	3.42	0.33	1.89	1.87	5456	0	0	8	71
Consonant optimal power (EP 1, $\rho = 0$)	3.39	3.39	0.33	2	1.72	5456	0	0	52.5	1122
Greedy algorithm	4.13	3.35	0.33	1.88	1.81					

Table A14: Power with two binary endpoints, $n = 15$, $\alpha = 0.05$, $p_1^{(Trt)} = 0.703$, $p_2^{(Trt)} = 0.297$, $p_1^{(Ctr)} = 0.297$, $p_2^{(Ctr)} = 0.297$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	45.3	45.3	1	44.9	1.4					
Tarone	45.4	45.4	1	45.1	1.4					
HKT	45.8	45.8	1	45.4	1.4					
Bonferroni optimal alpha	55.9	55.9	1.2	55.2	1.9					
Bonferroni optimal power (all)	58.7	58.7	1.2	58.1	1.8					
Bonferroni optimal power (1)	60.8	60.8	1.2	60.4	1.7					
Bonferroni greedy algorithm	56.3	56.3	1.2	55.8	1.7					
minP	56.3	56.3	1.2	55.8	1.7					
Optimal alpha	48.3	42.4	1.2	41.7	1.9	5456	0	22	160.5	1002
Optimal area	62.6	59.1	1.2	58.4	2	5456	0	22	156.5	998
Optimal power (all, $\rho = 0$)	53.2	46.6	1.2	45.8	2	5456	0	55	927	8955
Optimal power (all, $\rho = 0.5$)	52.7	46.2	1.2	45.4	2	5456	0	13	66	241
Optimal power (EP 1, $\rho = 0$)	69	60.9	1.2	60.2	1.9	5454	2	61	2446	20000
Consonant optimal alpha	57.8	57.8	1.2	57	2	5456	0	0	33	206
Consonnat optimal area	59.2	59.2	1.2	58.4	2	5456	0	0	33	206
Consonant optimal power (all, $\rho = 0$)	59.9	59.9	1.2	59.1	2	5456	0	0	45	1008
Consonant optimal power (all, $\rho = 0.5$)	59.8	59.8	1.2	59	2	5456	0	0	8	71
Consonant optimal power (EP 1, $\rho = 0$)	61.1	61.1	1.2	60.4	2	5456	0	0	52.5	1122
Greedy algorithm	61.8	59	1.2	58.2	2					

Table A15: Power with two binary endpoints, $n = 15$, $\alpha = 0.05$, $p_1^{(Trt)} = 0.703$, $p_2^{(Trt)} = 0.703$, $p_1^{(Ctr)} = 0.297$, $p_2^{(Ctr)} = 0.297$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	69.5	69.5	34	51.8	51.8					
Tarone	69.5	69.5	34	51.8	51.8					
HKT	69.5	69.5	34	51.8	51.8					
Bonferroni optimal alpha	81.8	81.8	36.4	59.7	58.6					
Bonferroni optimal power (all)	81.8	81.8	36.4	59.7	58.6					
Bonferroni optimal power (1)	81.8	81.8	36.4	60.4	57.8					
Bonferroni greedy algorithm	81.7	81.7	36.4	59.6	58.5					
minP	79.3	79.3	35.3	57.5	57.2					
Optimal alpha	91.2	82.3	36.4	59.6	59.2	5456	0	22	160.5	1002
Optimal area	89.3	83.9	36.4	60.4	60	5456	0	22	156.5	998
Optimal power (all, $\rho = 0$)	93.5	83.3	36.4	59.9	59.8	5456	0	55	927	8955
Optimal power (all, $\rho = 0.5$)	93.5	83.3	36.4	59.9	59.8	5456	0	13	66	241
Optimal power (EP 1, $\rho = 0$)	89.7	81.5	36.4	60.4	57.6	5454	2	61	2446	20000
Consonant optimal alpha	84.2	84.2	36.4	60.4	60.3	5456	0	0	33	206
Consonnat optimal area	84.1	84.1	36.4	60.4	60.2	5456	0	0	33	206
Consonant optimal power (all, $\rho = 0$)	84.3	84.3	36.4	60.4	60.3	5456	0	0	45	1008
Consonant optimal power (all, $\rho = 0.5$)	84.3	84.3	36.4	60.4	60.3	5456	0	0	8	71
Consonant optimal power (EP 1, $\rho = 0$)	84.2	84.2	36.4	60.4	60.2	5456	0	0	52.5	1122
Greedy algorithm	90.2	84.1	36.4	60.3	60.2					

Table A16: Power with two binary endpoints, $n = 15$, $\alpha = 0.05$, $p_1^{(Trt)} = 0.703$, $p_2^{(Trt)} = 0.703$, $p_1^{(Ctr)} = 0.297$, $p_2^{(Ctr)} = 0.297$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	61.2	61.2	41.7	51.4	51.4					
Tarone	61.2	61.2	41.7	51.4	51.4					
HKT	61.2	61.2	41.7	51.4	51.4					
Bonferroni optimal alpha	73.9	73.9	44.5	59.8	58.7					
Bonferroni optimal power (all)	73.9	73.9	44.5	59.8	58.7					
Bonferroni optimal power (1)	73.8	73.8	44.5	60.4	58					
Bonferroni greedy algorithm	73.8	73.8	44.5	59.7	58.6					
minP	71.9	71.9	43.4	57.7	57.6					
Optimal alpha	81.4	74	44.5	59.4	59.1	5456	0	22	160.5	1002
Optimal area	79	75.8	44.5	60.4	59.9	5456	0	22	156.5	998
Optimal power (all, $\rho = 0$)	81.8	74.8	44.5	59.7	59.6	5456	0	55	927	8955
Optimal power (all, $\rho = 0.5$)	81.9	74.5	44.5	59.6	59.4	5456	0	13	66	241
Optimal power (EP 1, $\rho = 0$)	80.3	74.4	44.5	60.4	58.6	5454	2	61	2446	20000
Consonant optimal alpha	76	76	44.5	60.3	60.2	5456	0	0	33	206
Consonnat optimal area	75.8	75.8	44.5	60.4	60	5456	0	0	33	206
Consonant optimal power (all, $\rho = 0$)	76	76	44.5	60.3	60.2	5456	0	0	45	1008
Consonant optimal power (all, $\rho = 0.5$)	76	76	44.5	60.3	60.2	5456	0	0	8	71
Consonant optimal power (EP 1, $\rho = 0$)	75.8	75.8	44.5	60.4	60	5456	0	0	52.5	1122
Greedy algorithm	78.9	75.8	44.5	60.3	60					

Table A17: Power with two binary endpoints, $n = 20$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.701$, $p_2^{(Trt)} = 0.299$, $p_1^{(Ctr)} = 0.299$, $p_2^{(Ctr)} = 0.299$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	54.8	54.8	0.7	54.6	0.9					
Tarone	54.9	54.9	0.7	54.6	0.9					
HKT	54.9	54.9	0.7	54.7	0.9					
Bonferroni optimal alpha	51.6	51.6	0.7	51.1	1.2					
Bonferroni optimal power (all)	55.4	55.4	0.7	55	1.1					
Bonferroni optimal power (1)	60.7	60.7	0.7	60.5	1					
Bonferroni greedy algorithm	58.7	58.7	0.7	58.5	1					
minP	58.7	58.7	0.7	58.4	1					
Optimal alpha	47.2	41.4	0.7	40.9	1.2	12341	0	58	740	10549
Optimal area	62.5	58.5	0.7	58	1.2	12341	0	57	732	10522
Optimal power (all, $\rho = 0$)	49.9	44	0.7	43.5	1.2	11662	679	208	9305	20000
Optimal power (all, $\rho = 0.5$)	49.4	43.4	0.7	43	1.2	12341	0	32	257	1287
Optimal power (EP 1, $\rho = 0$)	71.1	60.6	0.7	60.2	1.1	10521	1820	341	20000	20000
Consonant optimal alpha	56.1	56.1	0.7	55.6	1.2	12341	0	0	79	1893
Consonnat optimal area	58.4	58.4	0.7	58	1.2	12341	0	0	78	1889
Consonant optimal power (all, $\rho = 0$)	59.2	59.2	0.7	58.7	1.2	12341	0	0	278	7871
Consonant optimal power (all, $\rho = 0.5$)	59.1	59.1	0.7	58.6	1.2	12341	0	0	26	385
Consonant optimal power (EP 1, $\rho = 0$)	60.9	60.9	0.7	60.5	1.2	12337	4	0	461	20000
Greedy algorithm	62.4	59.3	0.7	58.9	1.2					

Table A18: Power with two binary endpoints, $n = 20$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.701$, $p_2^{(Trt)} = 0.701$, $p_1^{(Ctr)} = 0.299$, $p_2^{(Ctr)} = 0.299$, $\rho = 0$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	79.4	79.4	36.2	57.8	57.8					
Tarone	79.4	79.4	36.2	57.8	57.8					
HKT	79.4	79.4	36.2	57.8	57.8					
Bonferroni optimal alpha	82.1	82.1	36.6	59.4	59.3					
Bonferroni optimal power (all)	82.4	82.4	36.6	59.6	59.4					
Bonferroni optimal power (1)	82	82	36.6	60.5	58.1					
Bonferroni greedy algorithm	82	82	36.5	59.3	59.2					
minP	81.6	81.6	36.4	59.2	58.9					
Optimal alpha	92	82.1	36.6	59.5	59.2	12341	0	58	740	10549
Optimal area	91.7	84.2	36.6	60.4	60.3	12341	0	57	732	10522
Optimal power (all, $\rho = 0$)	95.2	83.7	36.6	60.2	60.1	11662	679	208	9305	20000
Optimal power (all, $\rho = 0.5$)	95.2	83.7	36.6	60.2	60.1	12341	0	32	257	1287
Optimal power (EP 1, $\rho = 0$)	89.4	81	36.6	60.5	57.1	10521	1820	341	20000	20000
Consonant optimal alpha	84	84	36.6	60.3	60.3	12341	0	0	79	1893
Consonnat optimal area	84.3	84.3	36.6	60.4	60.4	12341	0	0	78	1889
Consonant optimal power (all, $\rho = 0$)	84.4	84.4	36.6	60.5	60.5	12341	0	0	278	7871
Consonant optimal power (all, $\rho = 0.5$)	84.4	84.4	36.6	60.5	60.5	12341	0	0	26	385
Consonant optimal power (EP 1, $\rho = 0$)	84	84	36.6	60.5	60.1	12337	4	0	461	20000
Greedy algorithm	92.5	84.3	36.6	60.4	60.4					

Table A19: Power with two binary endpoints, $n = 20$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.701$, $p_2^{(Trt)} = 0.701$, $p_1^{(Ctr)} = 0.299$, $p_2^{(Ctr)} = 0.299$, $\rho = 0.5$.

Test	$H_1 \cap H_2$	H_1 or H_2	H_1 and H_2	H_1	H_2	C	NC	q50	q90	Max
Bonferroni	71.2	71.2	44.1	57.6	57.6					
Tarone	71.2	71.2	44.1	57.6	57.6					
HKT	71.2	71.2	44.1	57.6	57.6					
Bonferroni optimal alpha	74.6	74.6	44.5	59.6	59.5					
Bonferroni optimal power (all)	74.5	74.5	44.5	59.6	59.4					
Bonferroni optimal power (1)	74.2	74.2	44.5	60.5	58.3					
Bonferroni greedy algorithm	73.8	73.8	44.5	59.2	59					
minP	74.4	74.4	44.5	59.5	59.3					
Optimal alpha	82.1	73.4	44.5	59.1	58.9	12341	0	58	740	10549
Optimal area	81.4	76	44.5	60.4	60.1	12341	0	57	732	10522
Optimal power (all, $\rho = 0$)	83.8	75.2	44.5	59.9	59.8	11662	679	208	9305	20000
Optimal power (all, $\rho = 0.5$)	83.8	75.1	44.5	59.9	59.8	12341	0	32	257	1287
Optimal power (EP 1, $\rho = 0$)	81.2	73.7	44.5	60.5	57.8	10521	1820	341	20000	20000
Consonant optimal alpha	75.8	75.8	44.5	60.2	60.2	12341	0	0	79	1893
Consonnat optimal area	76.1	76.1	44.5	60.4	60.2	12341	0	0	78	1889
Consonant optimal power (all, $\rho = 0$)	76.3	76.3	44.5	60.4	60.4	12341	0	0	278	7871
Consonant optimal power (all, $\rho = 0.5$)	76.3	76.3	44.5	60.4	60.4	12341	0	0	26	385
Consonant optimal power (EP 1, $\rho = 0$)	75.9	75.9	44.5	60.5	59.9	12337	4	0	461	20000
Greedy algorithm	81.6	76.1	44.5	60.3	60.3					

8.4 Simulation results for three endpoints

The appendix tables A20 to A24 show the power for the different closed testing procedures when comparing two treatment groups with respect to three binary endpoints. The tables show the probabilities in percent that the closed testing procedure rejects the global intersection hypothesis $\cap_{i=1}^{3} H_i$, at least one elementary null hypothesis (Any H_i), all elementary hypotheses simultaneously (All H_i), or particularly H_1, H_2 or H_3.

The branch and bound algorithm was used with a maximal number of $5 \cdot 10^5$ iterations. In the table headings, n is the sample size per group, α the nominal global significance level, $p_i^{(Trt)}$ and $p_i^{(Ctr)}$ the success probability in endpoint i in the treatment and control group, respectively, and ρ the common product moment correlation between the endpoints.

The proportion of simulated instances in which a confirmed optimal solution was found by the branch and bound algorithm for all four intersection hypothesis tests in the closed testing procedure is given in the column labelled C(%). To further quantify the number of required iterations for finding the rejection region of the global intersection hypothesis test, the median (q50), the 90 % quantile (q90) and the maximum (Max) of the number of iterations are shown in the last three columns. A value of $5 \cdot 10^5$ indicates that the respective quantity to achieve an optimal solution would be above $5 \cdot 10^5$, but is here bounded by the maximal number of iterations.

For tests optimizing the power, the assumed alternative was identical to the true alternative.

Table A20: Power with three binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.254$, $p_2^{(Trt)} = 0.254$, $p_3^{(Trt)} = 0.254$, $p_1^{(Ctr)} = 0.254$, $p_2^{(Ctr)} = 0.254$, $p_3^{(Ctr)} = 0.254$, $\rho = 0$. The caclulations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$.

Test	$\cap_{i=1}^3 H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	0.4	0.4	0	0.2	0	0.1				
Tarone	0.7	0.7	0	0.4	0.1	0.2				
HKT	0.7	0.7	0	0.4	0.1	0.2				
Bonferroni optimal alpha	1.6	1.6	0	0.6	0.6	0.4				
Bonferroni optimal area	1.6	1.6	0	0.6	0.6	0.4				
Bonferroni optimal power	1.6	1.6	0	0.6	0.6	0.4				
Bonferroni greedy algorithm	1.4	1.4	0	0.7	0.4	0.3				
minP	1.2	1.2	0	0.6	0.4	0.3				
Optimal alpha	2.5	0.4	0	0.1	0.1	0.2	73.3	14624.5	5e+05	5e+05
Optimal area	2.2	1.4	0	0.7	0.4	0.3	90.3	2970	479137.3	5e+05
Optimal power	2.7	0.4	0	0.2	0.1	0.2	73.3	14619.5	5e+05	5e+05
Greedy algorithm	2.1	1.3	0	0.6	0.4	0.3				

Table A21: Power with three binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.254$, $p_2^{(Trt)} = 0.254$, $p_3^{(Trt)} = 0.254$, $p_1^{(Ctr)} = 0.254$, $p_2^{(Ctr)} = 0.254$, $p_3^{(Ctr)} = 0.254$, $\rho = 0.5$. The caclulations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$.

Test	$\cap_{i=1}^{3} H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	0.4	0.4	0	0.3	0	0.1				
Tarone	0.6	0.6	0	0.4	0.1	0.2				
HKT	0.6	0.6	0	0.4	0.1	0.2				
Bonferroni optimal alpha	1.7	1.7	0	0.8	0.4	0.6				
Bonferroni optimal area	1.7	1.7	0	0.8	0.4	0.6				
Bonferroni optimal power	1.7	1.7	0	0.8	0.4	0.6				
Bonferroni greedy algorithm	1.5	1.5	0	0.8	0.3	0.6				
minP	1.2	1.2	0	0.6	0.2	0.6				
Optimal alpha	2.6	1.1	0	0.5	0.2	0.6	99.8	10	683.2	5e+05
Optimal area	2.6	1.5	0	0.8	0.4	0.6	100	10	341	425584
Optimal power	2.5	1	0	0.4	0.2	0.5	99.8	9.5	683.4	5e+05
Greedy algorithm	2.5	1.6	0	0.8	0.3	0.6				

Table A22: Power with three binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.746$, $p_2^{(Trt)} = 0.254$, $p_3^{(Trt)} = 0.254$, $p_1^{(Ctr)} = 0.254$, $p_2^{(Ctr)} = 0.254$, $p_3^{(Ctr)} = 0.254$, $\rho = 0$. The caclulations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$.

Test	$\cap_{i=1}^3 H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	21.4	21.4	0	21.4	0	0				
Tarone	37	37	0	37	0.2	0.2				
HKT	37	37	0	37	0.2	0.2				
Bonferroni optimal alpha	34.9	34.9	0	34.2	0.4	0.5				
Bonferroni optimal area	34.6	34.6	0	34	0.4	0.5				
Bonferroni optimal power	40.6	40.6	0	40.1	0.3	0.4				
Bonferroni greedy algorithm	40.1	40.1	0	39.8	0.2	0.4				
minP	39.7	39.7	0	39.4	0.2	0.4				
Optimal alpha	42.4	22.2	0	21.8	0.4	0.2	42.8	5e+05	5e+05	5e+05
Optimal area	46	39.2	0	38.7	0.5	0.4	71.8	53071	5e+05	5e+05
Optimal power	57.5	40.3	0	40	0.3	0.2	98.8	1025.5	36389.9	5e+05
Greedy algorithm	44.1	38.4	0	37.8	0.4	0.5				

Table A23: Power with three binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.746$, $p_2^{(Trt)} = 0.746$, $p_3^{(Trt)} = 0.746$, $p_1^{(Ctr)} = 0.254$, $p_2^{(Ctr)} = 0.254$, $p_3^{(Ctr)} = 0.254$, $\rho = 0$. The caclulations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$.

Test	$\cap_{i=1}^3 H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	52.4	52.4	5.7	27.7	29.8	28.5				
Tarone	53.4	53.4	5.8	28.1	30.2	28.9				
HKT	53.4	53.4	5.8	28.1	30.2	28.9				
Bonferroni optimal alpha	76	76	6.6	37.5	40	39.3				
Bonferroni optimal area	75.5	75.5	6.6	37.2	40.2	39				
Bonferroni optimal power	76.3	76.3	6.6	37.5	40.1	39.6				
Bonferroni greedy algorithm	75.2	75.2	6.6	37.4	40.3	38.5				
minP	73.1	73.1	6.3	35.9	39	37.6				
Optimal alpha	85.8	68.5	6.6	35.5	37.5	33.8	44	5e+05	5e+05	5e+05
Optimal area	91.7	77.3	6.6	37.8	40.9	40	81.3	31456	5e+05	5e+05
Optimal power	95.8	76	6.6	37	40.3	39.1	64.3	132740	5e+05	5e+05
Greedy algorithm	93.3	77.4	6.6	37.8	40.9	40				

Table A24: Power with three binary endpoints, $n = 10$, $\alpha = 0.025$, $p_1^{(Trt)} = 0.746$, $p_2^{(Trt)} = 0.746$, $p_3^{(Trt)} = 0.746$, $p_1^{(Ctr)} = 0.254$, $p_2^{(Ctr)} = 0.254$, $p_3^{(Ctr)} = 0.254$, $\rho = 0.5$. The caclulations are based on 2000 random samples, the maximal number of iterations in the branch and bound algorithm was $5 \cdot 10^5$.

Test	$\cap_{i=1}^3 H_i$	Any H_i	All H_i	H_1	H_2	H_3	C(%)	q50	q90	Max
Bonferroni	43.5	43.5	17.1	30.9	30.9	31.4				
Tarone	44.1	44.1	17.1	31.1	31.2	31.6				
HKT	44.1	44.1	17.1	31.1	31.2	31.6				
Bonferroni optimal alpha	64.6	64.6	18.1	41.2	41	40.2				
Bonferroni optimal area	64.3	64.3	18.1	41.2	40.8	40.1				
Bonferroni optimal power	64.7	64.7	18.1	41.1	40.8	40.5				
Bonferroni greedy algorithm	63.9	63.9	18.1	41.1	40.9	39.6				
minP	63.6	63.6	18.1	40.4	40.5	40				
Optimal alpha	74.3	59.5	18.1	38.5	38.8	38.6	99.6	43.5	3166.7	5e+05
Optimal area	74.5	64	18.1	41.1	40.8	40	100	32	608.1	180941
Optimal power	76.9	61.2	18.1	39.3	39.5	39.5	99.8	34	1646.1	5e+05
Greedy algorithm	74.3	65.3	18.1	41.2	41.1	41				