Aus dem Institut für Biometrie und Medizinische Informatik der Medizinischen Fakultät
der Otto-von-Guericke-Universität Magdeburg

Multiple investigations in clinical trials

Habilitationsschrift

zur Erlangung des akademischen Grades

Dr. rer. nat. habil.
(doctor rerum naturalium habilitatus)

an der Medizinischen Fakultät

der Otto-von-Guericke-Universtität Magdeburg

vorgelegt von Ekkehard Glimm
aus Herford
Magdeburg 2013

Contents

List of Abbreviations

ANOVA	analysis of variance
COPD	chronic obstructive pulmonary disease
CTP	closed test principle
DLBCL	diffuse large B-cell lymphomas
DNA	desoxyribonucleic acid
EEG	electroencephalogram
FEV_1	forced expiratory volume from 1 second of exhaling
FDR	false discovery rate
FWER	familywise error rate
HbA1C	glycated haemoglobin A1C
LSD	least significant difference
MANOVA	multivariate analysis of variance
MLE	maximum-likelihood estimate
mRS	modified Rankin scale
MSE	mean-squared error
NIHSS	National Institute of Health Stroke Scale
OS	overall survival
PC	principal component
PFS	progression-free survival
SGRQ	St. Georges Respiratory Questionnaire
TQT	thorough QT (refers to studies analyzing the QT interval from an electrocardiogram)
UMVCUE	uniformly minimum variance conditionally unbiased estimate

Foreword

This text summarizes my research activities in 14 years following my PhD thesis. During this time, I have worked on the application of mathematical statistics in the life sciences, and to a very large part in the design and analysis of clinical trials. The work is inspired by the experience that multiplicity is ubiquitous in clinical trials and that medical statisticians benefit greatly from a well-stacked toolbox that allows them to get the right instrument for converting a research question into a statistical decision problem.

Within this common theme, the papers included in this cumulative habilitation cover a range of topics. Thus, they require a framework introducing the wider medical context as well as an outline of some technical foundations that they share. To meet these requirements, I start by discussing definitions and origin of multiplicity in clinical trials in chapter 1. Chapter 2 introduces the statistical models that are needed to deal with the challenges posed by multiplicity. The subsequent chapters on multiple testing, multivariate inference and estimation are split into a *foundation* section which discusses state of the art methodology and a *contributions* section summarizing the papers of this cumulative habilitation.

Chapter 1

Introduction

This work is about the design of clinical trials and the analysis of data arising from them. Most clinical trials are performed to assess the effect of a treatment on a disease. The *endpoints* of a clinical trial are measures quantifying this treatment effect. Such endpoints are derived from the responses to treatment observed on patients participating in the trial. For example, in a cancer trial, the response maybe the survival time, and the corresponding endpoint the observed hazard ratio of survival under a new treatment versus survival under the standard-of-care. In a trial for a new diabetes drug, the endpoint will often be the average percent change from baseline of glycated haemoglobin A1C (HbA1C) after 12 weeks of treatment[1]. For clinical trials aiming at admission of a drug to the market, health authorities typically require a pre-specified confirmatory analysis for the primary endpoint. The design and analysis of clinical trials is therefore often based on the *primary endpoint paradigm*, which states that only a single, univariate measure of interest, called primary endpoint, should be subject to the formal confirmatory statistical assessment performed at the end of the trial to decide about the treatment benefit. Any additional investigations of other endpoints are then labeled as *secondary* and are subject to a less formalized exploratory analysis.

In practice, the primary endpoint paradigm is often difficult to implement. Frequently there is some disagreement about the most appropriate primary endpoint. The reasons for this can be manifold. Here are some examples:

[1] The word *endpoint* is usually used without any more formal definition. To give an example, suppose the percent change from baseline of HbA1C after 12 weeks of treatment is measured in patients on a new treatment and patients in a control group. Suppose further that the *endpoint* is the ratio of these changes between treatment and placebo. Finally, assume that this assessment is done in two populations of patients: 1. the intention-to-treat-population and 2. the per-protocol population. Depending on the context, the two ratios will sometimes be considered as two endpoints or as two investigations of the same endpoint. In the context of this work, we will deal in general with situations where more than one endpoint is of interest, irrespective of ”type” of endpoint.

- In an ophthalmological trial, one health authority required the visual acuity as the primary endpoint, whereas the health authority of another country required the number of recurrences of eye inflammations.

- In many time-to-event trials, the primary endpoint is a composite of several events, e.g. death due to any cause, or non-fatal stroke, or hospitalization due to major myocardial event. Such composites are sometimes questioned and additional investigations of their individual components will often be of interest.

- The benefit-risk-profile of a drug can often not be assessed adequately with a single endpoint. For example, platelet-reducing drugs for the prevention of heart failure reduce the probability of thrombosis, but increase the risk of stroke.

- In earlier phases of clinical development, when a candidate compound is investigated for its potential benefit, trials are usually relatively small and short. Such trials do not yield a sufficient amount of data to reliably assess a clinical endpoint like survival time. In this case, *biomarkers* (e.g. measurements of laboratory parameters) are often investigated as surrogates. Usually, there are many biomarkers that need to be investigated simultaneously.

- The analysis of safety data usually comprises many types of adverse events. It is rarely possible to condense these into a single primary endpoint.

In addition to these cases, multiple endpoints also arise from repeated measures of the same quantity under different circumstances:

- The primary population of a clinical trial is not always obvious. Efficacy of a drug is typically investigated with the intention-to-treat population (all patients enrolled in the study, including those for whom some kind of protocol violation, i.e. a deviation from the treatment pre-specified in the study protocol, has occurred), whereas the per-protocol-population (only those patient whose treatment followed the study protocol) is often considered more appropriate for safety. As another example, it is sometimes suspected that a subpopulation of patients may benefit more from a new treatment than the rest of the study population and it is unclear whether the investigation of the full population or a subpopulation should be considered the primary comparison.

- Measurements of response to treatment are often taken repeatedly in time. For example, in glucose-lowering treatments for diabetic patients, change from baseline may be considered after 4, 6 or 12 weeks of treatment. The response to treatment at one of these timepoints or a summary measure of the time effect (e.g. the area under the curve or the slope of a linear time trend) are then candidates for the primary endpoint definition.

- In a similar fashion, data from clinical trials is sometimes investigated repeatedly in time when responses from only a fraction of all patients who have or will be recruited are available. Such trials are called *group-sequential trials* if the interim analyses are preplanned before the start of patient recruitment with the intention to stop the trial in case of conclusive evidence for the presence or absence of a treatment benefit. Group-sequential trials are a special case of *adaptive trials*. The latter allow additional changes of the study conduct, like selection of treatment arms or changes of overall sample size. In both cases, it is necessary to adjust statistical inference for the fact that the treatment benefit is estimated repeatedly in time.

- Several antihypertensive drugs are combination therapies which combine two monotherapies (e.g. two different blood-pressure lowering compounds) in a single pill. In order to gain approval, such treatments need to demonstrate an advantage over both monotherapies, and in some cases over placebo treatment as well. No single of these comparisons can be the sole primary comparison in this situation.

- In many trials, different doses or treatment regimens of the same drug are investigated with none of them being clearly of primary interest.

In all of these examples, the limitation of the confirmatory statistical analysis to a univariate primary endpoint seems undesirable. One may be tempted to simply declare several of these endpoints as equally important ("multiple primary endpoints"), and analyze them separately with methods from univariate statistical analysis. This is indeed a possible strategy, but care is needed with the correct interpretation of its results. Firstly, the statistical properties of selected extreme results (like e.g. the smallest p-value from several statistical tests) are not the same as those of the corresponding, seemingly identical non-selected results (e.g. the p-value from the same test, but this time not selected after the application of several statistical tests to the data). If such selected extreme results are treated as if they were from a single analysis, the consequences are:

- Inflation of nominal type I errors in statistical tests of hypotheses,
- selection biases and random extremes in point estimates of treatment effects,
- deviations from nominal coverage probabilities in confidence intervals of treatment effects.

These complications are referred to as the problem of *multiplicity* of endpoints.

Secondly, such simultaneous univariate analyses also fail to exploit relations between variables. They only consider the marginal variation in variables and do not make use of any patterns observed in their joint behavior.

To illustrate the effect of multiplicity, consider the following example:

Example 1: Type I error inflation from multiple tests in the investigation of the EEG

Läuter et al. (1996) investigate data from 19 depressive patients acquired at the beginning and at the end of a six week therapy with an antidepressant. The analysis considers the changes of absolute theta power of electroencephalograms (EEG) during the therapy in nine selected channels. Table 1.1 shows the data together with the observed mean, standard deviation and the p-value of a univariate two-sided t-test for the null hypothesis of no change in response separately for every channel.

Table 1.1: change of absolute theta-power in 19 depressive patients

Patient	ch3	ch4	ch5	ch6	ch7	ch8	ch17	ch18	ch19
1	-3.54	-3.11	-0.24	0.42	-0.49	2.13	-4.15	2.87	1.34
2	5.72	5.07	6.87	5.96	8.2	4.87	5.48	5.57	6.33
3	0.52	-0.18	0.9	0.6	1.27	1.28	-0.95	1.74	0.79
4	0	0.74	1.1	0.13	0.19	0.07	0.8	0.25	-0.66
5	2.07	0.76	3.51	0.6	3.71	1.86	1.49	3.11	1.8
6	1.67	4.18	2.77	4.55	1.8	4.79	4.51	3.24	3.99
7	9.13	12.92	3.44	4.8	0.48	1.63	9.94	1.34	1.53
8	-0.43	-1.59	-0.31	-0.61	-1.04	-0.13	-0.61	-0.61	-0.43
9	-0.56	0.92	-1.22	0.67	-0.97	-0.98	0	-1.22	-0.91
10	1.28	0.92	1.89	1.77	1.83	0.91	1.4	1.1	-0.12
11	3.21	3.41	3.92	0.85	2.77	0.79	3.31	1.2	2.15
12	1.47	2.38	1.22	1.71	-0.25	-1.52	1.46	2.26	2.01
13	-2.14	-2.44	-2.01	-1.95	0.31	-3.3	-1.04	-1.28	-1.59
14	-1.4	0.37	-1.1	0.18	-1.71	0.37	-0.12	-0.98	0.06
15	-0.29	0.31	-0.13	-0.89	-0.65	-1.06	0.71	0.35	-1.29
16	2.58	3.16	2.67	3.77	1.26	3.61	6.12	2.83	4.39
17	0.85	3.02	-0.59	2.17	0.65	1.25	1.2	-0.24	0.36
18	-1.71	-1.4	-1.83	-2.01	0.24	-0.61	-1.77	-0.12	-3.66
19	-1.89	0.76	-1.15	-0.95	-1.43	0.24	-0.77	-7.14	2.81
mean	0.871	1.589	1.037	1.146	0.851	0.853	1.422	0.751	0.995
stdev	2.950	3.514	2.358	2.252	2.276	2.069	3.261	2.635	2.359
p	0.215	0.064	0.071	0.04	0.121	0.089	0.074	0.23	0.083

The smallest of the p-values from the 9 tests occurs in channel 6 ($p = 0.04$). Considered in isolation, the change in channel 6 is statistically significant at the usual 5% level. Since, however, this is the minimum p-value from 9 simultaneous investigations, it overstates the evidence against the null hypothesis of no treatment effect on the EEG. A resampling test (Westfall and Young, 1993) can be used to quantify the true evidence against the null hypothesis H_0 that the treatment has no effect in any of the channels: If this hypothesis were true, then each permutation of the signs would yield an equally likely outcome. Hence, one can go through all $k = 1, \ldots, 2^{19} = 524288$ sign permutations $\mathbf{x}_i^{(k)} = \pm\mathbf{x}_i, i = 1, \ldots, 19$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{i9})'$ is the vector of the responses in the 9 channels from patient $i = 1, \ldots, 19$, obtain the minimum p-value from the 9 two-sided t-tests of every channel performed on each of the permutations, and record the fraction of cases in which the minimum p-value is smaller than the observed p-value from the original data. This results in an *adjusted p-value*, i.e. the probability of observing a minimum p-value of 0.04 or even smaller under H_0 (Westfall and Young's *min-p-test*, 1993). Here we get $p_{adj} = 0.17$, far larger than the unadjusted p-value of 0.04.

By using the distribution of a quantity derived from permutations as the empirical null distribution, Westfall and Young's min-p-test automatically takes into account the correlations between the 9 channels. Thus, it correctly adjusts for the selection bias inherent in the multiplicity problem. However, the minimum p-value is not an efficient quantity to detect differences between the pre- and post-treatment responses in this case. The stable multivariate tests by Läuter et al. (1996) combine the information from all channels in a more efficient way while also avoiding the bias arising from selecting the most significant of many univariate tests without proper adjustment. □

As the situations discussed so far indicate, multiplicity of endpoints is both ubiquitous in clinical trials and associated with several different research questions. This work will focus on statistical hypothesis testing. Many clinical trials are performed in order to detect differences between different treatment options. For example, most clinical trials performed in the pharmaceutical industry are designed to investigate if a new treatment under development performs better in terms of efficacy than previously available treatment options. For this purpose, the *null hypothesis* of no treatment effect is assessed by an appropriately chosen quantity (the so-called *test statistic*). If the test statistic takes a value which is larger than a critical value, the null hypothesis can be rejected and it can be concluded that there are statistically significant differences between the treatments (see e.g. Lehmann and Romano, 2005, for the theory

of statistical hypothesis testing). Multiplicity implies either that there are many such null hypotheses $H_1, \ldots, H_p$ or that the single null hypothesis H_0 consists of a statement about many treatment effects (e.g. the effects of the antidepressant on the different theta channels in the EEG in example 1) simultaneously.

The various problems posed by multiplicity have been the subject of a lot of research from various schools of thought. This work aims at clarifying relations between some of these, but an exhaustive, complete overview of all aspects of multiplicity is beyond the scope of this text. To give an example of a topic which is not covered, many Bayesian statisticians have worked on dealing with random highs via *shrinkage estimation* (e.g. Gelman et al., 2004).

To discuss the concepts of multiplicity adjustment and multivariate analysis in the context of statistical testing, we will use the following terminology:

- We assume that there are p *elementary (null) hypotheses* $H_i, i = 1, \ldots, p$. These constitute the atoms of the research questions at hand. For example, in a clinical trial investigating two doses of a new drug versus the standard-of-care in subpopulations of patients characterized by a genetic marker (M+ or M-), the hypothesis of no difference between new drug and standard-of-care in dose 1 and subpopulation M+ would be one of the H_i's.

- $H_0 = \cap_{i=1}^{p} H_i$ is called the *global* hypothesis. In the example, it would be the statement that there is no difference between new drug and standard-of-care in any dose or subpopulation.

- Other intersections $H_I = \cap_{I \subset \{1,\ldots,p\}} H_i$ will be called *intersection* hypotheses.

In the following, we will distinguish between two situations:

1. Elementary hypotheses are of specific interest:

   In most clinical trials, each of the corresponding elementary hypotheses H_i is of individual interest. In confirmatory phase III trials aimed at approval of a drug to the market, health authorities will often require so-called *strong control* of the type I error of erroneously declaring any of the tested endpoints as efficacious. That is, they require that the probability $P(H_i \text{ is rejected } | H_i \text{ is true }) \leq \alpha$ for all H_i, where H_i denotes the hypothesis that endpoint i is not efficacious. As a typical example, consider the case where two doses of a new drug are compared with the standard-of-care. In this case, it is not sufficient to conclude that the new drug is better than the standard-of-care (without further specifying

which dose it is that is actually better). A statistically qualified statement must be made about *which* of the doses is better.

Multiple comparison procedures (Hochberg and Tamhane, 1987; Hsu, 1996) have been developed to deal with this situation. In recent years, research in this area has focussed on stepwise multiple comparison procedures (Dmitrienko et al., 2009). These will be discussed in more detail in chapter 3.

2. Elementary hypotheses are not of specific interest:

   When there are many endpoints (dozens to millions), the individual endpoints are usually not of individual interest. Instead, it is explicitly or implicitly assumed that groups of endpoints (or more generally *variables* as they are called in multivariate analysis) are manifestations of some underlying, unobservable latent quantity (Läuter, 1992). Only the combined evidence from these variables is relevant for the decision about efficacy of a treatment. Consequently, regarding statistical testing, only H_0 is relevant. Gene expression data, readings from many channels in the EEG etc, are examples.

   Multivariate analysis methods are often suitable for such data. Substantial methodological progress in multivariate analysis was achieved in the fifties to seventies (Anderson, 1958, 1981, 2004; Ahrens and Läuter, 1974, 1981; Mardia, Kent and Bibby, 1979; Srivastava and Khatri, 1979). Interest in these methods was then revived around the year 2000 when advances in genomics created a need for high dimensional data analysis (Speed et al., 2003). Chapter 4 deals with multivariate analysis.

Typically, in the first situation there are few hypotheses, and the analysis is regarded as a multiplicity problem calling for a multiple comparison procedure. In contrast, the second situation is usually a consequence of the large number of hypotheses and is seen as a multivariate problem. Hence, one might be tempted to postulate a natural association:

number of hypotheses	elementary hypotheses of interest?	analysis method
few	yes	multiple comparison procedure
many	no	multivariate analysis

In many applications, these associations hold true, but they are not strict.

As an example of a case where few individual endpoints are not of individual interest, con-

sider the following case: A respiratory drug was investigated with the forced expiratory volume from 1 second of exhaling (FEV_1) as the primary endpoint in two independent clinical trials. Two trials were performed because health authorities often request confirmation of the treatment benefit in two independent trials. The two trials also investigated a quality-of-life related score from the St. George's Respiratory Questionnaire (SGRQ) as a secondary endpoint. Since the sample size of each individual trial would not be sufficient to reliably evaluate this more variable endpoint, it was agreed to pool the two trials for the sake of its analysis. In this situation, although formally there are two endpoints that in theory could be considered in two separate hypotheses (namely treatment effect on SGRQ in study 1 and 2), it was agreed that there is really only one hypothesis and that the possibility of a difference in treatment effect between the two studies can be ruled out a priori. Hence, this can be viewed as a situation where there are two endpoints (treatment effect in study 1 and 2), but we are not interested in an isolated statement about each of them.

As another example, in so-called *thorough QT (TQT) trials*, the effect of a new drug on the QT-interval is assessed. The QT-interval is a quantity derived from the electrocardiogram. A prolongation of this interval indicates an adverse effect of the drug on the heart function. TQT trials are performed in healthy volunteers who are randomly assigned to three groups receiving placebo, the new drug, and an active control, respectively. The QT-interval is measured pre-treatment and at $k > 1$ times post-treatment. The number of time points relevant for the assessment of the QT-interval is usually small, typically only 3 to 6 time points. The primary endpoint of a QT study is the change in the QT-interval length before and after treatment (ΔQT). The difference between the ΔQT of the new treatment and the ΔQT of placebo is called $\Delta\Delta QT$. The primary aim of the study is to establish that there is no substantial prolongation of the QT-interval due to the new drug, i.e. that $\Delta\Delta QT$ is below a pre-specified threshold, often 10 ms. The active control group is included to establish *assay sensitivity*: The $\Delta\Delta QT$ of the new drug with placebo is considered valid only if a corresponding, expected prolongation is demonstrated for the active control. Hence, the active control is a substance known to cause a QT prolongation, and assay sensitivity is considered established if at any point in time post treatment, the $\Delta\Delta QT$ of active control and placebo is *above* a threshold, e.g. 5 ms. The corresponding statistical decision problem addresses the elementary null hypotheses $H_t : \Delta\Delta QT \leq 5$ ms for time points $t = 1, \ldots, k$. Rejection of at least one of these hypotheses establishes assay sensitivity. Hence, we are only testing the global null hypothesis $H_0 = \cap_{t=1}^{k} H_t$. There is no need for

additional statements about the H_t's. For merely declaring assay sensitivity is not necessary to know at which specific point in time an increase of ΔQT was caused by the active control.

Regarding the analysis method used, efficiency (e.g. the power of statistical tests) is also a concern when deciding about the use of multiple comparison procedures or multivariate methods. If, for example, a treatment only affects one of many endpoints, then a multiple comparison procedure investigating the minimum p-value from several univariate tests of treatment versus control may yield a more efficient analysis than many multivariate methods, even if we are not interested in the individual variables. In practice however, these aspects are usually related to the number of variables: With just two or three variables, even if we claim not to be interested in the individual variable, a multiple comparison procedure like the simple Bonferroni adjustment is surprisingly ”hard to beat” in terms of efficiency (Srivastava, 2002, chapter 4). On the other hand, with thousands of gene expression measurements from a microarray, it is hardly conceivable that in reality a treatment affects only a single of these or that there is a genuine individual interest in more than a few genes that are of special interest due to prior information.

Finally, there are situations where questions about the individual importance of the elementary hypotheses cannot simply be answered by ”yes” or ”no”. For example, we may be interested in individual statements about some of the endpoints, but only a joint statement about an entire group of others, or we may be interested in individual statements about several intersection hypotheses H_I, but not down to the level where these H_I's are the elementary hypotheses. Such situations will occur repeatedly in the following chapters.

Chapter 2

Mathematical foundations

2.1 Multivariate normal distribution

In many of the situations outlined in chapter 1 it can be assumed that the multiple endpoints have a joint normal distribution: Let $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_p)'$ be the vector of treatment effect estimates in p endpoints. Then $\hat{\mathbf{y}}$ has a p-dimensional normal distribution

$$\hat{\mathbf{y}} \sim N_p(\boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ is the vector of true treatment effects and $\boldsymbol{\Sigma}$ a covariance matrix.

In the simplest case, formula (2.1) may arise from the difference $\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C$ between the average responses $\bar{\mathbf{y}}_i = \sum_{k=1}^{n_i} \mathbf{y}_{ik}, i = C, T$ of individual patients k who are randomly assigned to investigated treatment T and control treatment C, where the individual responses per treatment are assumed to follow a normal distribution. However, it may also arise as the asymptotic distribution of treatment effect estimates from a wide variety of other models, e.g. the treatment effect estimate from a multivariate analysis of variance (MANOVA) model adjusting for several other covariates, the joint distribution of log-rank test statistics in a group-sequential time-to-event trial, the joint distribution of log-odds ratio estimates of adverse event counts from different types of side effects of a drug, or the joint distribution of the estimated effect of a treatment in a full and in a subpopulation.

In general the true treatment effects $\boldsymbol{\beta}$ are unknown and the object of inference. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{i,j}$ is usually not of interest in itself, but it has an impact on point estimates, type I errors and confidence intervals of the treatment effect. It will almost never be

known completely, but it is not always entirely unknown either. Parts of it are usually known whenever endpoints are measures of the same underlying quantity and the correlation between them is caused by an overlap in patients. For example, if $\hat{\mathbf{y}}$ are the estimated average responses to treatment at different analysis time points in a group-sequential trial on a single normally distributed response which analyzes interim data at time j when n_j patients have completed treatment ($n_1 \leq n_2 \leq \ldots \leq n_k$), then $\sigma_{ij} = \sigma^2\sqrt{n_i/n_j}$. Thus $\boldsymbol{\Sigma}$ is known, apart from σ^2, the unknown variance of an individual response. Likewise, if several doses of a new treatment are compared with a common control and response to treatment is normally distributed with a common variance σ^2, the correlation between endpoints is $\sqrt{\frac{n_1}{n_0+n_1}}\sqrt{\frac{n_2}{n_0+n_2}}$ with sample sizes n_0, n_1, n_2, respectively, in the control group and dose groups 1 and 2.

Of course, if the multiple endpoints are genuinely different quantities like visual acuity and eye inflammations, the entire matrix $\boldsymbol{\Sigma}$ will be unknown.

The multivariate normal distribution assumption (2.1) is central to all of the publications summarized in this work.

2.2 Resampling-based Methods

The multivariate normal distribution provides an appropriate model for many types of multivariate data. Due to the central limit theorem, it often holds asymptotically, even if the original data is far from normally distributed. For example, the score statistics as well as likelihood ratio statistics are asymptotically normally distributed under very general assumptions (Pawitan, 2001). Sometimes, however, the multivariate normal distribution does not provide an adequate model. For example, asymptotic arguments are not relevant if sample sizes are very small. Furthermore, with multivariate data, it can be very difficult to estimate the covariance matrix $\boldsymbol{\Sigma}$ if there are many variables p. Resampling-based methods (Westfall and Young, 1993) are an alternative in these cases.

To illustrate the basic idea of resampling-based methods, assume that we have multivariate observations from n patients on p variables in two groups (for example, treated and untreated patients) with sample sizes n_1 and n_2, respectively, $n_1 + n_2 = n$. Let these be arranged in the two $n_i \times p$-matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and let $\boldsymbol{X} = \binom{\boldsymbol{X}_1}{\boldsymbol{X}_2}$ be the $n \times p$ matrix of all observations stacked into one matrix. It is furthermore assumed that there is a test statistic $t(\mathbf{X}) = t(\boldsymbol{X}_1, \boldsymbol{X}_2)$ for the global null hypothesis H_0 of no difference between the two groups. Under H_0, the distribution of

$t(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is not affected by whether a patient is assigned to group 1 or 2. Hence, any exchange of rows of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ does not change the distribution of the test statistic. To test H_0, we can thus proceed as follows:

1. Produce all $\frac{n!}{n_1!n_2!}$ permutations $\mathbf{X}^{(k)}$ of the rows of $\mathbf{X}$.
2. Calculate $t(\mathbf{X}^{(k)})$ from each permutation.
3. Reject H_0 if $\mathbf{X}$ is larger than the $\left((1-\alpha)\cdot\frac{n!}{n_1!n_2!}\right)$-largest value of the $t(\mathbf{X}^{(k)})$'s.

The set $\left\{t(\mathbf{X}^{(k)}), k = 1, \ldots, \frac{n!}{n_1!n_2!}\right\}$ is the empirical distribution of the random variable $t(\mathbf{X})$ under H_0 given the sample $\mathbf{X}$.

This method of statistical testing is very generally applicable. As long as the values of $t(\mathbf{X})$ can be ordered with respect to some notion of "closeness" to H_0, it can be used. In the description just given, it is assumed that large values of $t(\mathbf{X})$ provide evidence again H_0. Hence, the power of the procedure is high if deviations from H_0 lead to large values of $t(\mathbf{X})$. There is no need to derive the null distribution of $t(\mathbf{X})$, as the resampling process mimics this distribution empirically (under H_0). In this sense, correlations between $\mathbf{X}_1, \ldots, \mathbf{X}_p$ are automatically taken care of.

The method is very well suited for computer implementation. The permutation of rows of $\mathbf{X}$ can be done by attaching an index vector of group membership to $\mathbf{X}$ and just permuting this. This way, the method can be implemented very efficiently. If the number of permutations is too large, a random sample of all permutations can be performed, or the bootstrap (Efron and Tibshirani, 1993) can be used.

The method also has some disadvantages: First of all, it is designed for testing the global H_0 of no treatment effect. It is difficult to generalize to a shifted H_0 (like, for example, a non-inferiority hypotheses stating that differences between treatment and control are larger than some fixed value which is not 0). It also does not generalize naturally to point or interval estimation of treatment effects.

Furthermore, resampling can be combined with the closed test principle (see section 3.1) to decide about the statistical significance of single variables, but this requires an additional assumption: Assume that $\{1, \ldots, p\}$ is split into two mutually exclusively subsets S and non-S. If S consists of null variables only, then the conditional distribution of $t(\boldsymbol{X}_S)$ given $t(\boldsymbol{X}_{\text{non-}S})$ must be stochastically not larger when the non-subset variables do not fufill H_0 than when they do. This is a generalization of the *subset pivotality* condition by Westfall and Young (1993).

The latter requires that the conditional distribution of $t(\boldsymbol{X}_S)$ is the same for all values of the non-subset variables, i.e. $t(\boldsymbol{X}_S)$ and $t(\boldsymbol{X}_{\text{non-}S})$ are stochastically independent. In this case, the conditional null distribution of $t(\boldsymbol{X}_S)$ is the same irrespective of whether the non-subset variables do fufill H_0 or not.

In any case, this condition can be difficult to verify in practice. It is fulfilled for the multivariate Normal distribution when testing hypotheses about the mean, and the covariance matrix is the same for all rows of $\mathbf{X}$.

Chapter 3

Multiple comparison procedures

3.1 Foundations

Many of the problems arising in clinical trials are associated with few endpoints. As a typical case, assume that a trial's success depends on statistical evidence of a treatment benefit on the primary endpoint (for example overall survival time in an oncology trial or triglyceride level in a diabetes trial), but additional claims of treatment benefit may be attained from a secondary endpoint (e.g. weight loss in the diabetes trial case). Different doses of the treatment or the repeated testing of the endpoints may generate further multiplicity, but the situation is characterized by the fact that each of the elementary hypotheses $H_i, i = 1, \ldots, k$, is sufficiently interesting to be tested individually.

Familywise error rate (FWER) control is a key concept in this situation. A multiple testing method is said to control the FWER at level α if

$$P(H_I \text{ is rejected} \,|H_I \text{ is true}\,) \leq \alpha \tag{3.1}$$

for all sets of indices $I \subset \{1, \ldots, k\}$, where $H_I = \bigcap_{i \in I} H_i$. Hence, it requires that no true hypothesis or subset of true hypotheses must be rejected with a probability larger than α. This is also called *strong control* of the type I error. In contrast, *weak control* would only require to establish $P(H_K \text{ is rejected} \,|H_K \text{ is true}\,) \leq \alpha, \quad K = \{1, \ldots, k\}$.

Strong control places emphasis on each individual hypothesis H_i.

One of the most important techniques to achieve FWER control is the so-called closed test principle (CTP, Marcus, Peritz and Gabriel, 1976). Assume that a level-α-test is available for

each hypothesis $H_I, I \in P(K)$ where $P(K)$ is the power set of $K = \{1, \ldots, k\}$. Then the CTP states that H_I can be rejected if and only if all $H_J, I \subseteq J$ are rejected. In particular H_i is rejected if and only if all $H_I, i \in I$ are rejected. Marcus et al. (1976) show that this procedure provides strong control of the type I error at level α.

With the CTP, it is possible that H_I is rejected, but that no $H_i, i \in I$ can be rejected. A CTP-based multiple test procedure which avoids this is called *consonant*, i.e. it has the property that:

Rejection of H_I implies that there exists an $i \in I$ such that H_i is rejected.

Non-consonant multiple test procedures can get "stuck". The most famous example for a non-consonant procedure is Fisher's least significant difference method. Assume that we want to compare the means $\mu_i, i = 1, 2, 3$ of three populations from which we have samples with observations that are stochastically independent and normally distributed with equal but unknown variance. Then Fisher's LSD method consists of the following steps:

1. Test the global hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ with an F-test at level α.

2. If and only if H_0 is rejected, test the three hypothesis $H_{ij} : \mu_i = \mu_j, i \neq j$, each at level α.

This procedure controls the familywise type I error at level α. However, there is a positive probability that H_0, but none of H_{12}, H_{13}, H_{23} is rejected. Hence, there is sufficient evidence that the means μ_1, μ_2, μ_3 are not all equal, but not sufficient evidence to conclude that any pair of them is different.

In most practical clinical applications, consonance is a natural requirement. For example, a non-consonant procedure might lead to the conclusion that there is a treatment benefit on either triglyceride level or weight loss, but it is not possible to pin it down on one of the endpoints. Such an outcome would of course be perceived as useless.

Of course, a non-consonant procedure can always be converted into a consonant one by simply "throwing away" all rejections which are stuck at a non-elementary hypothesis. For example, assume that t_i is a test statistic and c_i a critical value chosen such that $P\left(t_i \geq c_i \,|H_i \text{ is true}\right) \leq \alpha$. Assume further that $H_i, i = 1, \ldots, k$ are the elementary hypotheses and $H_i, i \geq k + 1$ are intersections of these. By the CTP, rejection of H_i occurs if $t_{l(i)} \geq c_{l(i)}$ for all $l(i)$ where $l(i)$ are the indices of all intersection hypotheses that contain i, i.e. if H_i pertains to index set $I \in P(K)$, then the set of $l(i)$'s pertains to all $J \supseteq I$. Non-consonance means that there is a set of non-elementary hypotheses $\left\{H_{l_1}, \ldots, H_{l_q}\right\}$ which are nested according to the CTP such

that they are all rejected if $t_{l_1} \geq c_{l_1}, \ldots, t_{l_q} \geq c_{l_q}$, but for which $P(t_1 < c_1, \ldots, t_k < c_k, t_{l_1} \geq c_{l_1}, \ldots, t_{l_q} \geq c_{l_q}) > 0$. If all of the test statistics $t_1, \ldots, t_K$ have a continuous distribution with $P(t_i \in [b_l, b_u]) > 0$ for all $-\infty < b_l < b_u < \infty$, then of course it is always possible to find an infinite number of modifications of the critical values $c^*_{l_1} \geq c_{l_1}, \ldots, c^*_{l_q} \geq c_{l_q}$ with at least one strict inequality and $c^*_1 \leq c_1, \ldots, c^*_k \leq c_k$ with at least one strict inequality, such that the FWER is kept at α and

$$P(\text{one of } t_i \geq c^*_i, i = 1, \ldots, k, t_{l_1} \geq c^*_{l_1}, \ldots, t_{l_q} \geq c^*_{l_q}) >$$
$$P(\text{one of } t_i \geq c_i, i = 1, \ldots, k, t_{l_1} \geq c_{l_1}, \ldots, t_{l_q} \geq c_{l_q})$$

Hence under a few mild regularity conditions on the involved tests in a closed test procedure, for every non-consonant procedure there is a consonant procedure that has more power regarding rejection of the elementary hypotheses. Romano and Wolf (2007) show how such a consonant procedure can be constructed in a concrete case.

In summary,

- FWER control is an important principle in many clinical trials,
- consonance is often desirable, and
- the CTP is a general method to obtain FWER-controlling statistical tests from test procedures for the individual hypotheses.

These general considerations, however, still leave much room for concrete implementation in clinical trials. The CTP itself is a ”meta”-technique in the sense that it can be used with any valid set of statistical tests for the hypotheses indexed by $P(K)$. In particular, the closed test principle guarantees FWER control, but the power of the resulting procedure depends crucially on the concrete tests that are used for the individual hypotheses. Furthermore, the CTP can be tedious to implement when there are many elementary hypotheses, since the total number of hypotheses to be tested roughly doubles[1] with every additional elementary hypothesis.

One way of lifting this burden on computation is to restrict to methods which use *max test statistics* $t_1, \ldots, t_k$ for the elementary hypotheses. If doing so, the $2^k - 1$ tests use only k test statistics. For any intersection hypothesis H_I, the corresponding test statistic is $\max_{i \in I} t_i$. In

[1] It increases from $2^k - 1$ to $2^{k+1} - 1$.

order to fulfill the requirements of the CTP, we need to find 2^k critical values $c(I)$ such that

$$P_{H_I}\left(\max_{i\in I} t_i \geq c(I)\right) \leq \alpha. \tag{3.2}$$

The resulting procedure is consonant if $c(I) \leq c(J)$ for all $I \subseteq J$ (Hommel et al., 2007). If all t_i's have positive probability mass on the entire range of $[-\infty, \infty]$, then this condition is not only sufficient, but also necessary for consonance of the procedure.

A further simplification arises if the p-values p_i of the statistical tests for $H_i, i = 1, \ldots, k$ are used as the test statistics, and the Bonferroni principle is applied to obtain the critical values $c(I)$. In this case, we simply have $c_i = \alpha, i = 1, \ldots, k$ for the elementary hypotheses. The simplest ("unweighted") Bonferroni adjustment would use $c(I) = \alpha/card(I)$ for hypothesis H_I, where $card(I)$ is the number of elements in I. This way, we obtain the widely used Bonferroni-Holm procedure (Holm, 1979), represented in terms of the CTP. This principle can be generalized to *weighted Bonferroni-based stepwise multiple testing procedures*. Bretz et al. (2009) give an elegant generalized graphical framework for such approaches. While this simplifies the construction of complex multiple testing procedures tremendously, one has to keep in mind that the Bonferroni principle leads to conservative statistical tests and thus a power loss. When there are just a few elementary hypotheses, this power loss is usually very small. With many elementary hypotheses, however, it may be substantial.

3.2 Contributions

3.2.1 Application of graphical procedures in complex clinical trials (publication 1 in Appendix)

Maurer, Glimm and Bretz (2010) discuss how the requirements of a confirmative clinical trial with several hypotheses of potentially different importance can be translated into a multiple testing strategy. They introduce families of hypotheses to be tested following the principles described in the previous chapter. The families determine a hierarchy of primary and secondary hypotheses. Five situations leading to such hierarchies are frequently encountered in clinical trials:

1. **One primary with one descendant secondary endpoint for two comparisons.** a) Comparison of two treatment arms (e.g. two different doses of a drug) versus a control

arm. b) Comparison of a new treatment versus control in two subgroups of patients.

2. **Two primary endpoints with a secondary endpoint each.**

3. **A primary and secondary comparison of two equally important endpoints with correlation τ.** For the comparisons the same assumptions a) and b) are made as for case 1. This case arises when, for example, two doses of a new drug are investigated with respect to two equally important endpoints (e.g. visual acuity and number of inflammations in an ophthalmologic drug), and the higher dose is the primary target.

4. **Non-inferiority (primary) and superiority (secondary) testing** for a) two endpoints of equal importance with correlation τ, b) two treatment arms vs. control, and c) two subgroups. Here the correlation ρ between the non-inferiority and the superiority test is 1 if the same analysis population is used for the two tests (e.g. the "full analysis set") or $0 \leq \rho \leq 1$ if the set for non-inferiority is a subpopulation of the population used for the superiority test (e.g. the per protocol population).

5. **Group sequential testing of a primary and a secondary hypothesis with one interim analysis.** For the primary/secondary hierarchy we can consider here a) a primary and secondary endpoint, b) the total population and a subpopulation and c) non-inferiority and superiority testing in the same population.

In all these situations, the test statistics t_i follow the same asymptotic multivariate normal distributions. However, in some of the cases correlations between them can be calculated because they are functions of the sample sizes per group (generally speaking, this is the case when some observations are represented in several test statistics, e.g. if the results from the control group are used in the comparisons with both of two dose groups of a new treatment), whereas in others, they are unknown (usually when they represent different quantities measured on the same patient, e.g. HbA1c reduction and weight loss). The paper analyzes these correlations in detail. Regarding the statistical testing, most of the paper focuses on weighted Bonferroni-based stepwise procedures. These are generally applicable without knowledge of correlations. Some extensions are given for cases where correlations between test statistics can be derived from the trial design.

Furthermore, the paper defines the properties of *successiveness* and *consistency*. Similar to the way consonance prevents a closed test procedure from getting stuck before reaching a

useful conclusion about the elementary hypotheses, these properties are requirements which avoid test decisions that are not in line with the hierarchical structure of the primary and secondary families. A multiple testing procedure is called *successive* if a secondary hypothesis can only be rejected if at least one of its parent primary hypotheses is rejected. It is called *consistent* if it is successive and in addition the retention of a secondary hypothesis cannot preclude the rejection of a primary hypothesis. The paper discusses how these properties can be checked and how multiple testing procedures obeying them can be constructed.

3.2.2 Extension of graphical procedures to endpoints with known, or partly known correlations (publication 2 in Appendix)

Bretz, Posch, Glimm et al. (2011) extend the graphical procedures introduced by Bretz et al. (2009) to cases where the correlation between all or some of the test statistics are known. They show how this knowledge can be used to construct tests that are more powerful than Bonferroni-based procedures. More formally, if for each intersection hypothesis $H_J, J \subseteq I$ of a closed test procedure the joint distribution of the p-values $p_j, j \in J$ is known, a weighted min p-test can be defined in the following way: reject H_J if there exists a $j \in J$ such that $p_j \leq \alpha w_j(J) c_J$. The constant c_J satisfies

$$P_{H_J}\left(\bigcup_{j \in J} \{p_j \leq c_J \alpha w_j(J)\}\right) \leq \alpha. \tag{3.3}$$

Here, $w_j(J)$ are weights that determine the critical value for p_j in the step of the closed test procedure where H_J is tested (for example, $w_j(J) = 1/card(J)$ for the Bonferroni-Holm procedure). To exhaust the level as much as possible, c_J is chosen as large as possible subject to the inequality restriction (3.3). Obviously, if the p-values are continuously distributed, one can choose c_J such that the rejection probability is exactly α. For test statistics following the normal distributions described in section 2.1, the c_J's are functions of the correlations between the test statistics and increase with increasing correlations. For a Bonferroni-based closed test procedure, they are all 1. The distances from 1 can be interpreted as the "gains" from using knowledge about the correlations in comparison with a Bonferroni-based closed test procedure.

This seems a straightforward extension, but there are some complications beyond the mere technical difficulty of calculating the c_J's. In particular, while it is still simple to give a condition for consonance similar to the one following (3.2), the Bonferroni method's automatic guarantee

for achieving it is lost. Section 3.2 of the paper gives an example of this with two doses of a new drug tested against a control for non-inferiority in the per-protocol-population and for superiority in the intention-to-treat population. A modification is given to achieve consonance.

Furthermore, methods that exploit known correlations between test statistics are more powerful than Bonferroni-based methods with respect to the probability of rejecting at least one elementary hypothesis, but not necessarily with respect to rejecting any given hypothesis H_i, $i = 1, \ldots, k$ (see Maurer, Glimm and Bretz, 2010, p. 341 for an example).

3.2.3 Multiple testing in group sequential trials (publication 3 in Appendix)

Glimm, Maurer and Bretz (2010) discuss the important case of a primary and a secondary endpoint being investigated simultaneously in a group-sequential clinical trial. From a statistical perspective, this is one of the situations where part of the correlations between test statistics are known (namely, those between the tests of the same endpoint at several time points), whereas others (those between test statistics for primary and secondary endpoints) remain unknown. Hung and Wang (2007) had observed that the FWER is not kept by a hierarchical strategy that stops the trial when the primary endpoint is significant according to a type I error-controlling group-sequential approach, and then tests the secondary endpoint at full level α. Glimm et al. (2010) (and, independently of them, Tamhane, Liu and Mehta, 2010) show that an upper limit to the maximum type I error inflation in a two-stage clinical trial with multivariate normally distributed test statistics is given by $1 - \Phi_{\sqrt{t_{s,1}}}(u_{s,1}, u_{s,2})$, where $\Phi_\rho(\cdot)$ denotes the cumulative distribution function of the bivariate Normal distribution with mean $\mathbf{0}$, variances 1 and correlation ρ; $t_{s,1}$ is the information fraction available for the secondary endpoint at the interim analysis time point (typically, this will be the percentage of patients in whom this endpoint has already been observed at the interim), and $u_{s,1}$ and $u_{s,2}$ are the critical values for the interim and the final test, respectively, of the secondary endpoint. Knowledge of this upper bound allows the derivation of FWER-controlling group-sequential approaches. The approaches are based on the use of *alpha-spending* functions (see e.g. Jennison and Turnbull, 2000) which may be different for primary and for secondary endpoints. The operating characteristics of such group-sequential clinical trials (i.e. mathematically speaking: the probabilities with which primary and secondary hypotheses about the treatment effects can be rejected) then do not only depend on effect strength, variability and sample size, but also on the preplanned stopping strategies. For example, consider a group-sequential two-stage oncology trial with progression-

free survival (PFS) as the primary and overall survival (OS) as the secondary outcome. Two possible alternatives of conducting the trial are as follows:

1. If at the interim analysis, a statistically significant benefit is seen in PFS, stop the trial, irrespective of the result on OS.

2. Stop the trial at the interim analysis only if the treatment benefit is statistically significant for both PFS and OS. Otherwise, continue the trial to the final analysis.

The paper investigates these and various similar stopping strategies with respect to their operating characteristics. As is easily seen, there is no uniformly best approach for all possible configurations of primary and secondary effects sizes. The paper calculates rejection probabilities for the involved primary and secondary hypothesis under many scenarios. The main conclusions are:

1. If stopping of the trial for efficacy depends on the primary endpoint alone, it is advantageous to use an aggressive alpha-spending approach for the secondary endpoint (for example Pocock's approach (Pocock, 1977) for the secondary endpoint, when the primary endpoint is tested according to the O'Brien-Fleming approach (O'Brien and Fleming, 1979)).

2. If the trial continues in case of a significant benefit on the primary endpoint, but non-significance on the secondary endpoints, the use of similar alpha-spending methods (e.g. O'Brien-Fleming spending for both primary and secondary endpoints) will usually be preferable.

The paper illustrates planning and implementation of these methods in a trial comparing a new respiratory drug with an active control for the treatment of the chronic obstructive pulmonary disease (COPD).

3.2.4 Multiple testing in group-sequential time-to-event trials in Oncology (publication 4 in Appendix)

As already alluded to in the previous section, group-sequential time-to-event trials are particularly important in oncology, where the endpoints of interest are typically progression-free survival and overall survival. The log-rank test (e.g. Collett, 1994, section 2.5.2.) is usually used to test for a treatment effect. Group-sequential time-to-event trials are more complicated

to analyze than conventional group-sequential trials, because patients remain in the study until they have the event (e.g. they die). Hence, they are in the study for a time period which is unknown at the beginning of the clinical trial. In contrast to conventional, non-time-to-event trials, the variance of the effect size estimates (and thus the power of the trial) depends on the number of events rather than on the number of patients available. Because of this, group-sequential trials in oncology are usually *event-driven* which means that they continue until a certain number of events has been observed. In a group-sequential time-to-event trial, this means that, for example, the interim analysis is performed after 300 events and the final analysis after 600 events. Hence, if OS is the primary endpoint and interim and final analysis are planned for fixed numbers of deaths, then the number of PFS events at these points in time is a random variable.

Di Scala and Glimm (2010) discuss such trials. They consider the case of two doses of an experimental treatment compared with the standard-of-care in non-small cell lung cancer patients. The paper shows how the joint distribution of the log-rank test statistics for the two endpoints OS and PFS can be approximated by a normal distribution as introduced in chapter 2. Di Scala and Glimm then discuss four testing methods from adaptive design theory which guarantee FWER control in the group-sequential time-to-event case. These methods also allow the selection of treatment arms after the interim analysis, thus introducing an adaptive element into the group-sequential approach. The authors discuss criteria for such a selection. The criteria are based on weighted sums of predictive probabilities of success for the two endpoints PFS and OS. A treatment arm is dropped after the interim analysis if the predicted probability of success is low. Ultimately, this means that a treatment arm is deselected if its effect on disease progression and/or survival is disappointing. As discussed in chapter 1, such modifications introduce bias in the estimates of the retained treatment arm. The paper then describes how test statistics calculated from the two stages of the trial can be combined to yield valid inference on the selected treatment. Four methods originally developed for non-time-to-event data (König et al., 2008) are adapted to the survival analysis case. Di Scala and Glimm discuss the conditions under which these methods keep the FWER and compare the operating characteristics of the adjustment methods by simulation.

Chapter 4

Multivariate methods

4.1 Foundations

Chapter 1 already discussed some communalities and differences between multiple comparison procedures and multivariate statistical methods. This chapter focusses on the application of multivariate statistics in clinical trials.

From a mathematical point of view, multivariate analysis is a broader field than multiple comparison procedures. It encompasses a very wide range of methods like multivariate analysis of variance, principle components analysis, factor analysis, multiple-dimensional scaling and more. An important sub-division is between parametric (Anderson, 2003; Srivastava and Khatri, 1979; Srivastava, 2002) and non-parametric multivariate analysis (Puri and Sen, 1971).

Regarding statistical testing, parametric multivariate analysis and multiple endpoint analysis are both based on the same joint multivariate distributions of test statistics. Furthermore, multiplicity in clinical trials is often generated by multiple measurements per sample unit which is precisely the situation that led to the development of multivariate analysis.

In contrast to multiple testing procedures, multivariate analysis is less concerned with statements about elementary hypotheses. Often, these are not even considered at all. For example, Hotelling's classical T^2-test for two samples $\mathbf{y}_{ij}, j = 1, \ldots, n_i, i = 1, 2$ from p-dimensional normal distributions $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), i = 1, 2$, with unknown covariance matrix $\boldsymbol{\Sigma}$ (Hotelling, 1931) only considers the global null hypothesis $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \mathbf{0}$. This line of research does not attempt to make statements about the individual components $\mu_{ij}, j = 1, \ldots, p$, of $\boldsymbol{\mu}_i$. Furthermore, this test and its MANOVA generalizations are affine-invariant. For Hotelling's T^2-test this means that its power depends on the true mean and variances only via the Mahalanobis distance

$const(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ (where $const$ is a scalar function of the sample sizes in the treatment groups) and thus any linear transformation $\mathbf{Y}\boldsymbol{C}$ of the original observations $\mathbf{Y} = \left(\mathbf{y}_{ij}'\right)$ (arranged in an $n \times p$-matrix) with a positive definite $p \times p$ matrix $\boldsymbol{C}$ leads to the same test as the untransformed data $\mathbf{Y}$. Loosely speaking, these tests are looking for deviations from H_0 in any direction. From a mathematical point of view, this seems a natural requirement. Assume for example, that in a clinical trial patients are randomized to two different treatments, and that responses to treatment are the changes of various biomarkers due to treatment. Hotelling's T^2-test tests whether there are any differences between the two treatments at all, without any preferences for certain ways in which such differences may arise. However, this has two consequences which limits usefulness in practical applications:

1. The aim of a clinical trial is usually to demonstrate or refute a "treatment benefit". Therefore, it is usually desired to have more power for certain alternatives (e.g. for $\mu_{2j} - \mu_{1j} > 0$ for all $j = 1, \ldots, p$ if this implies a treatment benefit on all endpoints) than for others.

2. The methods tend to become unstable as the dimension p increases (Läuter, 1992). Again loosely speaking, there are too many directions in which deviations from H_0 can occur. In consequence, the power for detecting deviations from H_0 is spread across too big a space of alternative parameter values. If the sample size n is smaller than the dimension p, maintaining affine-invariance even becomes impossible and hence these tests exist only if $n > p$.

The second of these issues has been called the "curse of dimensionality" (Bellman, 1957). Several authors have addressed it by developing ANOVA-type test statistics for the mean vector $\boldsymbol{\mu}$ which avoid estimation of the entire covariance matrix $\boldsymbol{\Sigma}$, see e.g. Srivastava (2007, 2009), Srivastava and Du (2008), Box (1954) and Dempster (1958, 1960). In the subsequent sections of this chapter, we will however focus on a different approach that tackles both issues in parallel.

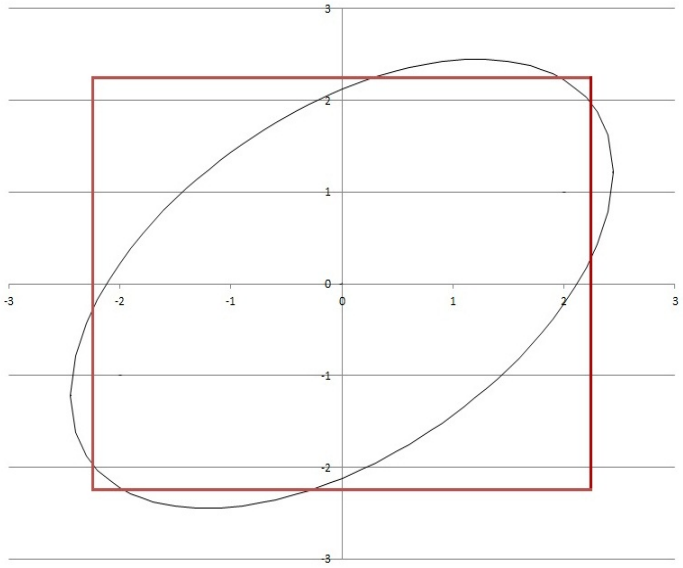
In case of few variables p, the advantages of multiple comparison procedures compared with multivariate tests usually outweigh their disadvantages. As discussed in chapter 3, multiple comparison procedures allow statements about individual endpoints. However, even if such statements are not relevant, for small dimensions power differences between the T^2-test and a Dunnett- or Bonferroni-test are small and the parameter constellations where the T^2-test has higher power than the Dunnett-tests are in parts of the alternative space that are not relevant.

Figure 4.1 gives an illustration for $p = 2$. It displays rejection regions for two test statistics

t_1 and t_2 that have distribution $\mathbf{t} = (t_1, t_2)' \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ is known. $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is to be tested. Then the obvious simplification of Hotelling's T^2-test is Scheffé's test which rejects H_0 if $t = \boldsymbol{t}'\boldsymbol{\Sigma}^{-1}\boldsymbol{t} \geq \chi^2_{1-\alpha}(2)$, where $\chi^2_{1-\alpha}(2)$ is the $(1-\alpha)$-quantile of the χ^2-distribution with 2 degrees of freedom. The graph displays the rejection regions of this test and the Bonferroni-adjusted two-sided test for H_0 (which rejects if $\max(t_1, t_2) > u_{1-\alpha/2}$ where $u_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution). The rejection region of the T^2-test consists of all points outside the ellipse. The Bonferroni-adjusted test's rejection region is given by all points outside of the box. We see that the area covered by the ellipse is generally smaller (hence, the T^2-test rejects "more" in general). However, if we restrict attention to the positive orthant $t_1 > 0, t_2 > 0$, then there is relative little difference between the rejection regions and there are even values of (t_1, t_2) where the Bonferroni test rejects, but the T^2-test does not.

Of course, the probability mass associated with regions in this plot - and thus the true power of those tests - depends on the true means $\boldsymbol{\mu}$. Still, the graph illustrates nicely that the simple Bonferroni adjustment may serve our purpose well, in particular when we suspect that treatment effects are positively correlated and roughly share the same direction of deviation from H_0, as is often reasonable to assume in practice.

Figure 4.1: Rejection regions of Bonferroni's and Scheffé's test in case of $p = 2$

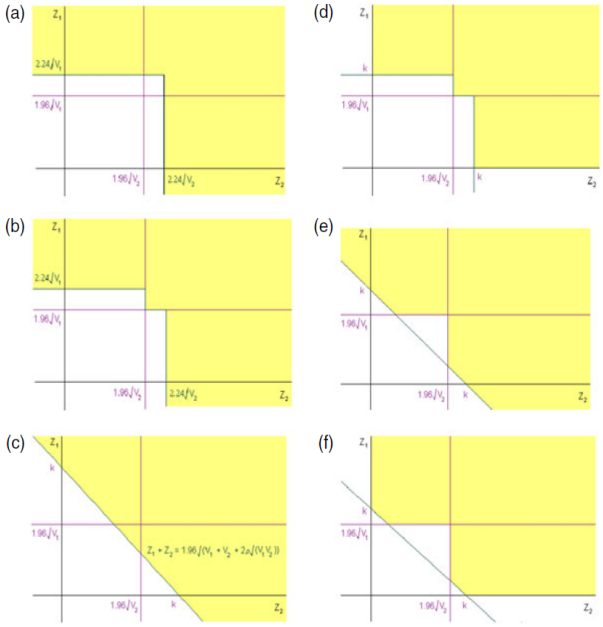


4.2 Contributions

4.2.1 Comparison of multiple and multivariate tests in case of two hypotheses (publication 5 in Appendix)

In the example just discussed, there is a test statistic t that rejects H_0 if and only if $t > c$, where c is such that $P_{H_0}(t > c) = \alpha$. Furthermore, this test statistic t has a representation $t = f(t_1, \ldots, t_p)$, where t_i is a test statistic for endpoint i and $f(.)$ is some function $\mathbf{R}^p \to \mathbf{R}$ defined on all possible values of t_i. The condition $t > c$ is equivalent to $\mathbf{t} = (t_1, \ldots, t_p)' \in S(c) \subset \mathbf{R}^p$, where $S(c)$ is a set of points in $\mathbf{R}^p$ leading to rejection of H_0. For low dimensions like $p = 2$ or 3, we can define a shape for the rejection region S (e.g. all values outside of a rectangle or of an ellipse), and then calculate its boundaries in such a way that the condition $P_{H_0}(\mathbf{t} = (t_1, \ldots, t_p)' \in S) \leq \alpha$ is kept. For $p = 2$, this is done in Su, Glimm, Whitehead and Branson (2011). Figure 4.2 is taken from the publication and shows the rejection regions considered. The paper deals with "one-sided" decisions: all rejection regions are unbounded in the upward direction on each axis.

Figure 4.2: Rejection regions of several directional multivariate tests in case of $p = 2$



The rejection regions (shaded) of six possible procedures for testing two elementary hypotheses (with $\alpha = 0.025$): (a) Bonferroni; (b) Simes; (c) combined; (d) restricted Simes; (e) consonant combined; (f) restricted consonant combined.

The paper investigates the properties of these multivariate tests. In the multivariate normal case, the probability that the observed vector of test statistics (t_1, t_2) is in the rejection region

depends on its bivariate normal distribution, i.e. on the true unknown means and the correlation between the test statistics. If we assume knowledge of these quantities, the power of the suggestions can be calculated. It is obvious that no choice of rejection region can be optimal simultaneously for all parameter constellations. Hence, the paper investigates which choice of rejection region is advantageous under what assumed constellation of true parameter values.

Furthermore, it investigates how well the methods perform if they are applied asymptotically to the log odds ratio estimated from binary data. As an example, it considers the success rate of a new cardiovascular treatment and a control treatment as judged by the modified Rankin scale (mRS) and the National Institute of Health Stroke Scale (NIHSS). The observational vector from every patient can take one of the four values $(0,0);(0,1);(1,0)$ or $(1,1)$ depending on whether the treatment was successful or not according to mRS and NIHSS, respectively. From this data, the correlation between the log odds ratios for comparing new treatment and control can be calculated. Regarding type I error control, simulations in the paper indicate that no type I error inflation arises from treating the estimated correlations as if they were the known ones in this context.□

The directional decisions mentioned in Su et al. (2011) are an important aspect of multivariate analysis. In the univariate situation of a single endpoint, the decision problem addressed by a statistical test is either $H_0 : \mu \leq 0$ versus $A : \mu > 0$ or $H_0 : \mu = 0$ versus $A : \mu \neq 0$, corresponding to a one- or a two-sided test, respectively. Sometimes, the one-sided case is written in a simplified way as $H_0 : \mu = 0$ versus $A : \mu > 0$. This somewhat imprecise practice arose, because most practically applied tests are either likelihood-ratio tests of H_0 versus A (e.g. the t-test) or are asymptotically equivalent to it. The likelihood ratio test of parameter θ is in general defined as $\frac{\max_{H_0} ll(\boldsymbol{\theta})}{\max_A ll(\boldsymbol{\theta})}$ where $ll(\boldsymbol{\theta})$ is the likelihood[1]. In the typical univariate setting, $\max_{H_0} ll(\boldsymbol{\theta})$ occurs at $\mu = 0$. Hence, the univariate tests of $H_0 : \mu = 0$ versus $A : \mu > 0$ and $H_0 : \mu \leq 0$ versus $A : \mu > 0$ are identical, rendering the distinction between them irrelevant.

The important aspect here, however, is that there are only two directions in which the parameter μ can deviate from 0. In multivariate statistics, there is an infinity of directions. Therefore, no uniformly most powerful test exists and various suggestions are superior to each other depending on where the true parameter values $\boldsymbol{\mu}$ are. As mentioned, Hotelling's T^2 test and its MANOVA generalizations like Wilks' Λ and the Hotelling-Lawley trace statistic are affine-

[1] $\boldsymbol{\theta}$ may contain nuisance parameters. For example, in case of the t-test, $\boldsymbol{\theta} = (\mu, \sigma^2)$

invariant: All directions are treated equal, only the distance from H_0 decides about the strength of evidence against H_0.

4.2.2 Spherical tests

In clinical trials, however, research questions are often "one-sided": We would usually like to know whether a new treatment is better than the standard-of-care. This desire has prompted a lot of research on directional multivariate tests as generalizations of the one-sided univariate test. A class of such tests, called spherical tests, has been suggested by Läuter (1996) and Läuter, Glimm, and Kropf (1996, 1998). To simplify the discussion, we will consider spherical two-sample tests of $H_0 : \boldsymbol{\mu}_T - \boldsymbol{\mu}_C = \mathbf{0}$ in the model $\bar{\mathbf{y}}_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), i = T, C$ from chapter 2. Spherical tests are based on a low-dimensional score derived from the original p-dimensional data. If this score is one-dimensional, it takes the form $\mathbf{d}'(\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C)$. Läuter, Glimm and Kropf (1996) show that this can be treated like the mean from a univariate normal distribution, i.e. tested by a standard t-test. The test is exact, that is, under H_0, the type I error is exactly α (although the variance of $\mathbf{d}'(\bar{\mathbf{y}}_T - \bar{\mathbf{y}}_C)$ is estimated), if the calculation of $\mathbf{d}$ is based on the total sums-of-products matrix $\mathbf{G} = (\mathbf{Y} - \mathbf{1}_n\bar{\mathbf{y}}')(\mathbf{Y} - \mathbf{1}_n\bar{\mathbf{y}}')'$ derived from the observed data, where $\mathbf{Y}$ is the $n \times p$-matrix of observations $\mathbf{y}_i'$ and $\bar{\mathbf{y}}$ is the vector of overall means per variable, calculated across treatment groups. Different ways of calculating $\mathbf{d}$ allow to tailor the test for power against specific alternatives. Two important variants are the standardized-sum test which uses $\mathbf{d} = Diag(\mathbf{G})^{-1/2}\mathbf{1}_p$, another the principal-components (PC) test where $\mathbf{d}$ is the eigenvector pertaining to the largest eigenvalue λ of $\mathbf{G}$. The latter test motivated by the assumption of a *one-factor model* (Läuter, 1992). In this model, the variables are differently scaled representations of an unobserved latent variable which can be imagined as the unobserved common cause of the measurements on the individual variables. Glimm and Läuter (2002) have given a proof of admissibility of the principal-components test in the class of spherical tests. The latter paper is essentially a summary of the most important results from my PhD thesis (Glimm, 1999).

The concept leads to a very flexible class of tests useful in many situations. Läuter, Glimm and Kropf (1996) discuss the one- and the two-sample test of means, one-way ANOVA and tests for multiple correlation. The range of possibilities is exhaustively explored in a more theoretical paper by Läuter, Glimm and Kropf (1998).

4.2.3 Spherical tests in multivariate linear models with mixed effects (publication 6 in Appendix)

Glimm (2000) describes how spherical tests can be constructed in complex multivariate linear models with mixed (i.e. fixed and random) effects. The paper uses a general representation of multivariate analysis-of-variance (MANOVA) models with fixed effects describing the influence of fixed covariates (e.g. treatment group or sex) on the population-average response and random effects describing correlations between the measurements from different patients - in addition to the correlation that exists between the measurements of different variables from the same patient. Such random effects can, for example, be used to describe the impact of a random sample of centers who participate in a large clinical trial, and where patients who are treated in the same center show a similar response because they are treated by the same physicians or with the same equipment.

The paper shows how spherical tests for the variance components (i.e. the variation in the responses which is due to the random effects) can be constructed from generalizations of Cochran's theorem (see e.g. Searle, 1971, for Cochran's theorem). The paper illustrates the construction of such tests with a generic example of a clinical trial where a multivariate response per patient is observed repeatedly in time. Tests for time trend, for correlation due to random effects and for compound-symmetric correlation in time are derived as examples for applying the principle. Furthermore, the paper contains a short set of simulations that compare several spherical tests with each other and with Wilks' Lambda-test. In the simulation, the spherical tests turn out to be vastly superior to Wilks' Lambda test. This, however, could have been easily predicted from the simulation setup in this case. The key objective of the paper was to present the general principle for constructing spherical tests in complex MANOVA models, not an exhaustive investigation of their performance in different scenarios.

4.2.4 A stable multivariate test with high power for the positive orthant (publication 7 in Appendix)

In spherical test theory, score weights $\mathbf{d}$ can be derived in many different ways depending on needs of a practical application. Glimm, Srivastava, and Läuter (2002) discuss how a stable multivariate test can be derived for the situation where high power is intended against the alternative that all components δ_i of $\boldsymbol{\delta} = \boldsymbol{\mu}_T - \boldsymbol{\mu}_C$ are larger or equal to 0 with at least one

equality, but no preference can be given to any specific direction inside the positive orthant $\delta_i \geq 0$ for all $i = 1, \ldots, p$.

For ease of exposition, the authors investigate the one-sample case, i.e. the test of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ from data $\boldsymbol{X} = \left(\boldsymbol{x}_j'\right)_{j=1,\ldots,n}$ with $\boldsymbol{x}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stochastically independent. This is no real restriction, as the derivations in the two-sample case are almost identical. The suggested tests use test statistics $\bar{u}^2 = \sum_{i=1}^{p} \max(0, u_i)^2$ based on the components of $\boldsymbol{u} = (u_i) = \sqrt{n}\boldsymbol{A}\bar{\boldsymbol{x}}$ where $\boldsymbol{A}$ is a matrix root of $\boldsymbol{G}^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, i.e. some matrix which fulfills $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{G}^{-1}$, and $\bar{\boldsymbol{x}}$ is the multivariate mean of the sample. Tang et al. (1989) had derived similar tests, but assuming that $\boldsymbol{\Sigma}$ is known, whereas this paper deals with the situation where $\boldsymbol{\Sigma}$ must be estimated. The paper derives the null distribution of $\bar{u}^2$ (which is independent of the concrete choice of $\boldsymbol{A}$) and discusses the choice of $\boldsymbol{A}$ so as to render $\bar{u}^2$ unique and attain properties like scale- and order-invariance. The suggested tests perform well in simulations compared to the conditional likelihood-ratio test by Wang and McDermott (1998) and a bootstrap-based method called M-test by Srivastava, Hirotsu, Aoki, and Glimm (2001). □

It is important to emphasize that directional multivariate tests like the ones just discussed have high power for specific alternatives and low power for others, but ultimately, they are tests for the point null hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$, not for composite hypotheses comprising sets of values of $\boldsymbol{\mu}$. They are constructed as tests of this H_0 against specific alternatives A. If the true parameter value is outside of both H_0 *and* A, then these tests allow no conclusion. For example, if a directional multivariate test of $H_0 : \mu_i = 0$ versus $A : \mu_i \geq 0$ (with at least one inequality) rejects H_0, then we cannot conclude that there is also sufficient evidence against $\mu_i \leq 0$ for all i. In general, almost all directional multivariate tests as well as the very closely related tests for restricted alternatives (Schaafsma and Smid, 1966; Perlman, 1969, Wang and McDermott, 1998; Sen and Silvapulle, 2005) share this property which does not occur in the univariate case. Most theoretical research regarded this as a curiosity (e.g. Silvapulle, 1997). In many biological applications focussed on finding differences between gene expressions in microbial communities, say, it does not matter, as the direction of differences is of minor interest, and power considerations are decisive for the choice of a directional test. In many clinical trials applications, however, this lack of interpretability is viewed as a serious disadvantage.

4.2.5 Multivariate tests keeping the type I error for null hypotheses covering entire regions of the parameter space (publication 8 in Appendix)

Glimm and Läuter (2010) address this topic and derive multivariate directional statistical tests that allow conclusions beyond the rejection of $\boldsymbol{\mu} = \mathbf{0}$. They discuss one directional test based on Hotelling's T^2-test and two variants of spherical tests. All three suggestions test $H_0 : \boldsymbol{\mu} \leq \mathbf{0}$ (i.e. all components μ_i of $\boldsymbol{\mu}$ are smaller than 0) versus $A : \mu_i > 0$ for at least one component μ_i of $\boldsymbol{\mu}$. Hence, no "blind spots" are left in the parameter space of $\boldsymbol{\mu}$.

Again, it is easiest to describe the methodology for the one-sample case with data $\boldsymbol{X} = \left(\boldsymbol{x}_j'\right)_{j=1,\ldots,n}$, $\boldsymbol{x}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. (Section 4 of the paper gives the analogous results for the two-sample comparison.) The directional Hotelling test rejects H_0 if no point $\boldsymbol{\mu}_0 = (\mu_{0i})_i$ with $\mu_{0i} \leq 0$ is inside the region

$$C_{1-2\alpha}(\overline{\mathbf{x}}, \mathbf{G}) \cup \{\boldsymbol{\mu}_0 \text{ with } \sum_{i=1}^{p} \frac{\mu_{0i} - \overline{x}_i}{\sqrt{g_{ii}}} > 0\},$$

where

$$C_{1-2\alpha}(\overline{\mathbf{x}}, \mathbf{G}) = \{\boldsymbol{\mu}_0 \text{ with } \frac{(n-p)n}{p}(\boldsymbol{\mu}_0 - \overline{\mathbf{x}})'\mathbf{G}^{-1}(\boldsymbol{\mu}_0 - \overline{\mathbf{x}}) < F_{1-2\alpha}(p, n-p)\}$$

is a confidence ellipsoid around the observed multivariate average responses $\overline{\mathbf{x}} = (\bar{x}_i)_{i=1,\ldots,p}$, $\frac{1}{n-1}\mathbf{G} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})'$ is the estimate of the covariance matrix $\boldsymbol{\Sigma}$ and g_{ij} are its elements. Glimm and Läuter give an algorithm to check this condition.

The other two suggestions are modifications of the standardized sum test which in this context rejects $H_0 : \boldsymbol{\mu} = \mathbf{0}$ if

$$t^0 = \sqrt{n-1}\frac{\sqrt{n}\overline{\mathbf{x}}'\mathbf{d}^0}{\sqrt{\mathbf{d}^{0\prime}\mathbf{G}\mathbf{d}^0}} \geq t_{1-\alpha}(n-1) \tag{4.1}$$

where $t_{1-\alpha}(n-1)$ is the $(1-\alpha)$ quantile of the t-distribution with $n-1$ degrees of freedom and $\mathbf{d}^0 = \left(\frac{1}{\sqrt{g_{ii}+n\overline{x}_i^2}}\right)_{i=1,\ldots,p}$. The paper describes how the weight vectors $\boldsymbol{d}$ and the matrix $\boldsymbol{G}$ that play a role in this test can be modified to convert it into tests of the more general hypothesis $H_0 : \boldsymbol{\mu} \leq \mathbf{0}$.

Glimm and Läuter (2010) give an application to a trial for an osteoporosis treatment against placebo with three endpoints: change in joint space width, a functional and a pain score. In this case, rejection of H_0 allows to conclude that the treatment is different from placebo and that it improves the patients' condition with respect to at least one of the three endpoints.

The suggested tests are scale-invariant, i.e. the test decision is not influenced by the measurement scale chosen for the variables. Glimm and Läuter conclude that, in comparison with multiple testing approaches, these tests have most power if the treatment effect is roughly equally strong in all variables, for example, if all variables are subject to an underlying common treatment effect. If the treatment effect is not ”evenly spread” across all variables in this way, but rather there is a single variable with a strong treatment effect, then multiple testing procedures are superior. □

In summary, the publications investigated in this chapter so far discuss ways of increasing power for directional alternatives while maintaining type I error control for a global test of H_0. Interpretational difficulties following rejection of H_0 remain an issue. Due to this, multivariate tests (irrespective of whether they are the classical affine-invariant ones, spherical tests, or test with restricted alternatives) are rarely applied in clinical trials of the later phases II-IV. In these phases, there are few relevant endpoints, usually we are interested in statements about single endpoints, and even if we are not, the power of multiple test procedures for the rejection of the global H_0 is usually competitive with that of multivariate tests. Thus, FWER-controlling methods serve the requirements of these trials well.

The picture changes in case of many variables. Here, FWER control becomes a very severe burden on the power of the statistical tests. This problem may sometimes be exacerbated by inefficient adjustment (e.g. the use of Bonferroni adjustment), but in essence, it is the FWER criterion (3.1) itself that becomes almost impossible to fulfill if many hypotheses are investigated simultaneously.

Luckily, as already discussed in chapter 1 the criterion usually gets less and less compelling as the number of variables increases. As an example that was already discussed, in gene expression analysis there is little point in investigating the expression levels from DNA snippets in isolation. Instead, one would like to know if, for example, diseased and healthy tissue displays any differences at the genetic level at all. If this is the case, researchers further hope to identify groups of relevant genes that behave similarly or antagonistic.

Regarding the first of these two research questions, the spherical tests from section 4.2.2 are useful for the global hypothesis H_0 of no differences between multivariate measurements from different populations. Glimm, Heuer, Engelen and Smalla (1997) describe an application to the comparison of the catabolic profiles from microbial communities in different types of soil.

In this application, each soil sample is characterized by its utilization of 95 sole-carbon sources of the BIOLOG identification system. Kropf et al. (2004), Eszlinger et al. (2005), Smalla et al. (2007) and Ding et al. (2012) give further applications in microbiology. Kropf (2000) and Schuster, Kropf and Roeder (2004) give some medical applications.

The second question regarding the identification of differentially expressed groups of variables is discussed informally in these publications. To this end, the contribution of the variables to the score weights $\boldsymbol{d}$ of the spherical tests is investigated.

4.2.6 Multivariate tests for high-dimensional genomics data (publication 9 in Appendix)

Mathematically more rigorous approaches are considered by Läuter, Glimm and Eszlinger (2005). They describe methods that first group genes into sets with correlated expression levels, rank these sets and then test for significant differences between diseased and healthy tissue. Läuter, Glimm and Eszlinger consider both methods that are based on the theory of spherical tests and resampling methods. More precisely, assume that the gene expression data is collected in a $n \times p$-matrix $\boldsymbol{X} = \left(\boldsymbol{x}_j'\right)_{j=1,\ldots,n}$ and we want to test the p-dimensional mean $\boldsymbol{\mu} = \boldsymbol{0}$, corresponding to no gene expression. The tests are based on the test statistic

$$B = \frac{n\boldsymbol{d}'\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}'\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{d}}. \tag{4.2}$$

If $\boldsymbol{X}$ is multivariate normally distributed and $\boldsymbol{\mu} = \boldsymbol{0}$, then $B \sim Beta(\frac{1}{2}, \frac{n-1}{2})$. The weight vector $\boldsymbol{d}$ can be determined by one of the methods described in Läuter, Glimm and Kropf (1998). To create sets of genes, the correlation of each gene with each other gene is calculated. Each gene i in turn acts as a "pivot". If the correlation is above a threshold c, the gene k is put into the pivot set $S(i)$ of gene j. At the end of the process, the pivot sets are sorted by importance based on a criterion that includes the number of genes in the set and their "total correlation" $\boldsymbol{x}_i'\boldsymbol{x}_k$, where $\boldsymbol{x}_i$ is the n-dimensional vector of expression levels from gene i. The sets $S(i)$ are then tested in order of importance with the test statistic B applied to all genes in the set. The process is continued until the first non-significant result is attained. Läuter, Glimm and Eszlinger show that this approach keeps the FWER in the following sense: If a set does not contain any differentially expressed gene, then this set will be identified as significant with probability α at most. In this sense, the strategy provides a compromise between the global tests originally

suggested by Läuter et al. (1996, 1998) and multiple test procedures that keep the FWER on the level of single genes. Note also that differential expression is the focus of interest, so on the level of a single gene, a two-sided decision is desired.

The paper also discusses methods that do not necessarily require the assumption of a multivariate normal distribution. To this end, a variant of the resampling procedures discussed in chapter 2.2 is suggested, using the sum $\sum_{k\in S(i)} B_i$ of Beta-statistics $B_i = \frac{n\bar{x}_i^2}{\boldsymbol{x}_i'\boldsymbol{x}_i}$ as test statistics for the sets $S(i)$ with $\bar{x}_i$ being the average of the expression level observations from gene i. Finally, the paper introduces a rotation test strategy that can be used in cases where sample sizes are too small for the resampling-based methods to work reliably. These latter methods require that the data has a left-spherical distribution under H_0 which means that the distribution of $\boldsymbol{C}'(\boldsymbol{X} - \mathbf{1}_n\boldsymbol{\mu}')$ must be the same as that of $\boldsymbol{X} - \mathbf{1}_n\boldsymbol{\mu}'$ for all orthogonal $p \times p$-matrices $\boldsymbol{C}$ (Fang and Zhang, 1990). The multivariate normal distribution of $\boldsymbol{X}$ with independent rows $\boldsymbol{x}_j \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ is a special case of a left-spherical distribution.

The methods are applied to data from 14 patients with thyroid diseases. The dataset contains expression patterns from tissue samples of cold thyroid nodules and the normal surroundings. The differences between the logarithmic expression values of nodular and surrounding tissue are analyzed in 148 of the originally screened 12 625 genes. Both methods identify a few highly correlated groups of genes with a total of 8 and 9 genes, respectively, partly represented repeatedly in the identified subsets.

The method of grouping, sorting and testing genes in this way is refined by Läuter, Horn, Rosolowski, and Glimm (2009). In this paper, the identification of candidate gene sets is done on 6374 genes from 108 patients suffering from Burkitt's lymphoma or diffuse large B-cell lymphomas (DLBCL), whereas the testing of the identified genes is performed on data from 10 biological replicates of c-myc overexpressing cells and 10 control cultures. The selection algorithm yields 99 significant maximum sets of genes, 68 of which follow the direction of c-myc expression changes in the cell experiments while 31 respond reciprocally. The paper then focusses on graphical displays for characterizing the gene expression from the 108 patients in 8 non-overlapping sets with a total of 240 identified genes. It turns out that the sets allow a clear distinction between Burkitt and non-Burkitt patients. The biological interpretation of the identified gene expressions is also discussed.□

It should be acknowledged here that these methods of testing variables after having ar-

ranged them into groups by their correlations are less popular than the *false discovery rate* (FDR) controlling procedures of Benjamini and Hochberg (1995). The latter have produced a vast amount of follow-up literature (Benjamini and Yekutieli, 2001; Storey, 2002; Meinshausen and Bühlmann, 2006; Speed et al., 2003; Efron, 2010; van der Laan and Dudoit, 2007). These methods originated from research on multiple testing. The FDR criterion is a relaxation of the FWER criterion: Rather than requiring that no true hypothesis must be rejected, it only requires that the expected fraction of true hypotheses among the rejected hypothesis must be limited to α, where α is usually chosen to be 5%. Thus, a concession is made to the huge number of hypotheses that typically occur in genomics. Like FWER, FDR is a concept and thus may be implemented in many different ways. Almost all of the suggested procedures, including Benjamini and Hochberg(1995)'s original approach, treat the hypotheses in isolation; they do not attempt to exploit relations between different variables.

Chapter 5

Estimation

5.1 Foundations

The previous chapters concentrated on statistical testing in the presence of multiple endpoints. However, point and interval estimation of treatment effects are also affected by multiplicity. For example, suppose that an investigator tests several doses of an experimental drug against a control treatment and then reports only the result from the dose group which performed best. It is obvious that the reported result has a stochastic tendency to look "better than it should". However, it also turns out that this informal statement is surprisingly hard to formalize mathematically. To illustrate this claim, let us assume that $x_i \sim N(\mu_i, \sigma_i^2)$, $i = 1 \ldots, p$ represent the stochastically independent average responses to p different doses of an experimental treatment. A large response corresponds to a favorable treatment effect. First of all, the distribution of the response from dose k with $x_k := \max(x_i)$ depends on all parameters $\mu_i, \sigma_i^2, i = 1, ..., p$: If $\mu_k = \mu_k - \mu_i \to \infty$ for all $i \neq k$, the distribution of x_k converges to $N(\mu_k, \sigma_k^2)$, whereas for $\mu_i/\sigma_i = \mu_1/\sigma_1$ for all $i = 1 \ldots, p$, the distribution is neither normal nor even symmetric. Hence, if in truth one of the doses yields a much better treatment effect than all other doses, there is very little bias in the reported best result, whereas the bias is large if in reality all doses have the same effect. Furthermore, it has been shown by Putter and Rubinstein (1968) (see Cohen and Sackrowitz, 1989) that no unbiased estimate of $\max(\mu_i)$ exists if $\mu_1, \ldots, \mu_p$ are unknown parameters that have to be estimated from the data.

Unbiased estimation of the effect of a selected treatment is possible, if additional responses to the selected treatment are obtained in a second step of sampling. A typical example would be a clinical trial where participating patients are randomly allocated to one of several doses

of the experimental treatment (or a control treatment) in stage 1. At the end of stage 1, the average response to the experimental treatment is calculated and one of the doses is selected. In stage 2 of the trial, additional patients are randomly allocated to the selected dose and the control treatment and finally, the responses to the selected dose and the control are compared. Clinical trials performed in this manner are called *adaptive* clinical trials. Such trials have been discussed extensively in the statistical literature (Bauer, 1989; Bauer and Köhne, 1994; Proschan and Hunsberger, 1995; Lehmacher and Wassmer, 1999; Müller and Schäfer, 2001, 2004; Bretz, König, Brannath, Glimm, and Posch, 2009 [1]). These publications also describe other design modifications (like sample-size re-estimation or subpopulation selection). They are primarily focussed on the impact of design modifications on error rates of statistical tests, less on the impact on point and interval estimation.

5.2 Contributions

5.2.1 Unbiased estimation of the effect of a selected treatment in a two-stage clinical trial (publication 10 in Appendix)

Based on work by Cohen and Sackrowitz (1989), Bowden and Glimm (2008) derived a uniformly minimum variance conditionally unbiased estimate (UMVCUE) of the selected treatment effect in a two-stage adaptive design with a treatment arm selection at interim. In this situation, the maximum-likelihood estimate (MLE) of the selected treatment arm "ignores the selection", and thus will usually be biased (see e.g. Bretz, König, Brannath, Glimm, and Posch, 2009, section 6). The UMVCUE is defined as an estimator of the effect of the selected treatment which is unbiased conditional on the order of treatments, as determined by the mean effect estimates from stage 1. Among all estimators of this kind, it has minimum variance. Since conditional unbiasedness is a stronger requirement than unconditional unbiasedness, the UMVCUE is also unconditionally unbiased. It is given by

$$\hat{\theta}_{(i)} = \frac{\sigma_2^2 X_{(i)} + \sigma_1^2 Y}{\sigma_1^2 + \sigma_2^2} - \frac{\sigma_2^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \frac{\phi\left(W_{i,i+1}\right) - \phi\left(W_{i,i-1}\right)}{\Phi\left(W_{i,i+1}\right) - \Phi\left(W_{i,i-1}\right)} \tag{5.1}$$

[1] Regarding the last of these publications, I was primarily responsible for section 6 which discusses estimation of the treatment effect after treatment arm selection or sample size re-estimation in an adaptive trial.

where $W_{s,i} = \frac{1}{\sigma_1^2}\left(\frac{\sigma_2^2 X_{(s)}+\sigma_1^2 Y}{\sqrt{\sigma_1^2+\sigma_2^2}} - \sqrt{\sigma_1^2+\sigma_2^2}X_{(i)}\right)$, $X_{(0)} := \infty$ and $X_{(k+1)} := -\infty$ and k is the number of treatments at the start of the trial. Here, $X_{(j)}$ denotes the effect estimate of the treatment with the j-th largest observed effect after the interim analysis ($X_i \sim N(\theta_i, \sigma_1^2)$ for all $i = 1, \ldots, k$). $Y \sim N(\theta_{(i)}, \sigma_2^2)$ is the stage-2-effect estimate of the selected treatment and $\phi(.)$ and $\Phi(.)$ are the density and the distribution function of the standard normal distribution, respectively. The formula gives a conditionally unbiased estimate for all explicitly selected treatments, i.e. it can be applied for rules that specify to carry the best, best two etc. treatments into stage 2. It can also be applied if the treatment with the largest effect is not among those continued into stage 2. If only the best treatment is selected at interim and the variance remains constant (i.e. $\sigma_1^2 = \sigma_2^2$), the formula simplifies to

$$\hat{\theta}_{(1)} = Z/2 - \frac{1}{\sqrt{2}}\frac{\phi(W)}{\Phi(W)} \tag{5.2}$$

with $Z = X_{(1)} + Y$ and $W = \sqrt{2}(Z/2 - X_{(2)})$. This shows that the UMVCUE consists of the MLE $Z/2$ minus a correction term for the bias from treating $X_{(1)}$ as if it had been selected at random (and not due to its high value).

The paper presents these results in general for random variables X_i and Y. As a special case, X_i and Y could be the observed mean responses from several patients. Then σ_1^2 would be the variance of the observed stage-1-means, i.e. $\sigma_1^2 = \frac{\sigma_\epsilon^2}{n_1}$, in case of equal stage 1 sample sizes n_1 per treatment group and a common error variance σ_ϵ^2 across treatments and stages. Similarly, $\sigma_2^2 = \frac{\sigma_{1\epsilon}^2}{n_2}$ for stage 2. However, the result would (asymptotically) also hold for any other estimate selected from asymptotically normally distributed treatment effect estimates, e.g. if this is obtained from an ANOVA model with several groups and possibly other covariates.

Bowden and Glimm also compare the mean-squared error (MSE) of the UMVCUE and the MLE and suggest a method to derive confidence intervals. It turns out that the UMVCUE's unbiasedness comes at a high price: In terms of the MSE, the MLE is superior to the UMVCUE. The choice of method thus depends to a large extent on the importance researchers are willing to place on unbiasedness.

Bowden and Glimm (2013) generalized the two-stage UMVCUE to multistage adaptive trials that start with k treatment arms and drop some of them at each stage until only one arm is tested at the final stage.

5.2.2 Shrinkage estimation of the selected treatment (publication 11 in Appendix)

Bowden, Glimm, and Brannath (2013) discuss an alternative Bayesian approach to estimating the effect of a selected treatment. This latter approach leads to an estimate which is no longer unbiased, but superior in terms of mean-squared error to the UMVCUE, the maximum-likelihood estimate and several other estimates that have been suggested in the literature (e.g. the conditional maximum-likelihood estimate of Bebu, Luta and Dragalin, 2010).

5.2.3 A confidence region for the mean of multidimensional data (publication 12 in Appendix)

Regarding interval estimation, Läuter and Glimm (2005) have used data compression methods very similar to those used in the PC-test (see section 4.2.2) to calculate confidence regions for the mean vector $\boldsymbol{\mu}$ of the multivariate normal model $\boldsymbol{X} = \left(\boldsymbol{x}_j'\right)_{j=1,\ldots,n}$, $\boldsymbol{x}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\bar{\boldsymbol{x}} = \sum_{j=1}^{n} \boldsymbol{x}_j$. A weight vector $\boldsymbol{d}$ is calculated as the eigenvector corresponding to the largest eigenvalue of the sums-of-products matrix $\boldsymbol{G} = \sum_{j=1}^{n} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$. Läuter and Glimm show that the set of values $\boldsymbol{\mu}_0$ which fulfill

$$\sqrt{n}\bar{\boldsymbol{x}}'\boldsymbol{d} - \sqrt{\frac{\lambda F_{1-\alpha}(1, n-1)}{n-1}} < \sqrt{n}\boldsymbol{\mu}_0'\boldsymbol{d} < \sqrt{n}\bar{\boldsymbol{x}}'\boldsymbol{d} + \sqrt{\frac{\lambda F_{1-\alpha}(1, n-1)}{n-1}}$$

provide a conservative $(1-\alpha)$-confidence region for $\boldsymbol{\mu}$. This region is a hyperdisc in the space of p variables. This means that there is no limit on the individual components μ_i of $\boldsymbol{\mu}$. Rather, the region is limited in the directions where most of the variation in the data occurs. The method is also not scale-invariant. As a consequence, this method of deriving confidence regions should only be used if individual variables are not of interest in isolation, and if all variables are measured on the same scale. The paper also presents results for testing multivariate data with hypotheses having $\boldsymbol{\mu} \neq \boldsymbol{0}$, generalizations of the weight vector $\boldsymbol{d}$ to a weight matrix $\boldsymbol{D}$, and application of principal components inference in combination with the selection of variables.

Chapter 6

Discussion

The papers presented here all deal with continuous multivariate data obtained from experiments in the life sciences. The task of summarizing the evidence on a feature of interest (for example the *treatment effect* in a clinical trial or *differences between microbial communities* in an agricultural experiment) when this feature manifests itself in many dimensions (several variables/endpoints/biomarkers, repeated measures in time or space) is a challenge to statistical inference.

In clinical trials assessing a treatment effect in multiple endpoints, it seems natural to focus on endpoints where the treatment effect is largest. This is not an illegitimate or unreasonable strategy per se. However, we must remain aware that reporting the maximum response is a special case of a *summary of multivariate data*. As illustrated in example 1, its stochastic behavior is affected by the other endpoints in the experiment. In the example, seemingly clear evidence in favor of a treatment effect in a selected channel turned out to be far less convincing when the result was adjusted for that selection. Chapter 3 discusses such methods for multiplicity adjustment in the context of statistical tests. We note that the maximum response from several endpoints is rarely consciously perceived as a summary of multivariate data by scientists attempting to interpret the results from a clinical trial. However, as has been demonstrated, proper statistical adjustments have to be made when the multiple endpoints correspond to several research questions which are tackled simultaneously in a clinical trial. This is particularly important if there is no clear prioritization of objectives and if each research question is of individual importance.

Of course, the summary of multivariate responses does not necessarily have to be via the selection of a single endpoint. Especially in case of many, possibly closely related endpoints,

like gene expressions or EEG channels, they may not represent distinct research questions of interest. In such cases, it is advantageous to look for other ways of summarizing the data, e.g. by comparing linear combinations of the results. Again, if the weights of the linear combinations are derived from the data, a multiplicity issue arises. Furthermore, efficient ways of summarizing the multivariate data (e.g. statistical tests having a high power for detecting certain relevant deviations from a hypothesis of no treatment effect) require multivariate statistical techniques. The papers in chapters 4 all deal with this.

In clinical trials, the focus is usually on finding or refuting evidence for a significant treatment effect. Therefore, the papers in chapters 3 and 4 are concerned with statistical tests. However, as already mentioned in chapter 1, point and interval estimation are also affected by multiplicity. This is the topic in the papers of chapter 5.

In the foreword, I had made an appeal for a well-stacked toolbox. I hope that the papers compiled in this work have added some tools for the proper handling of multiplicity to the biostatistician's toolbox.

Acknowledgments

I would like to thank all my co-authors for the collaboration, scientific discussions and friendship. With Jack Bowden, Werner Brannanth, Michael Branson, Lilla Di Scala, Florian Klinglmüller, Franz König, Martin Posch, Kornelius Rohmeyer, Muni Srivastava, Ting-Li Su and John Whitehead, there were many intense discussions via e-mail exchanges, phone calls and personal meetings about the topics of our papers.

Special thanks are due to my PhD supervisor, Jürgen Läuter. We have kept regular contact since the completion of my PhD in 1999 and continued working on the development of statistical methods for multivariate data. Without his support, his ideas and his passion for our profession, this work would not have been completed.

Furthermore, I also owe much to Siegfried Kropf, my former colleague from the Institute of Biometry and Medical Informatics at the Otto-von-Guericke-University. We also stayed in touch regularly and I benefitted a lot from many discussions about the statistical analyses of medical data and the design of clinical trials. Siegfried involved me in several very fruitful collaborations like the MÄQNU-project funded by the Deutsche Forschungsgesellschaft (DFG), but also smaller, more informal projects. His support in the submission of this thesis was invaluable and I am very grateful to him.

Likewise, my thanks are due to the other colleagues from the Institute of Biometry and Medical Informatics at the Otto-von-Guericke-University, in particular Friedhelm Röhl und Silke Ribal.

My colleagues in the statistical methodology group at Novartis Pharma AG, Basel, have also been very supportive of this work. This is particularly true for the global group head, Frank Bretz, but also for Willi Maurer (with both of whom I have worked very closely in recent years on multiple comparison procedures), for Mouna Akacha and Björn Bornkamp.

Zusammenfassung

Diese kumulative Habilitationsschrift umfasst 12 wissenschaftliche Publikationen, welche die statistische Analyse multipler Endpunkte in den Biowissenschaften, insbesondere in kontrollierten klinischen Studien, zum Gegenstand haben.

Multiple Endpunkte treten in einer klinischen Studie auf, wenn die Wirkung einer Behandlung durch mehrere Messgrößen beschrieben wird, welche simultan, ohne klare Prioritisierung analysiert werden sollen. Diese Situation ist in der Medizinstatistik allgegenwärtig. Eine wissenschaftlich korrekte Analyse multipler Endpunkte muss der *Multiplizitätsproblematik* Rechnung tragen, welche dadurch zum Ausdruck kommt, dass z.B. der geschätzte Behandlungseffekt des ”besten” Endpunktes eine Verzerrung aufweist, dass simultane statistische Tests das nominelle Fehlerniveau nicht einhalten usw., sofern keine Adjustierung der statistischen Analysemethodik vorgenommen wird.

Der zusammenfassende Text stellt zunächst eine Reihe von Situationen aus realen klinischen Studien vor, in welchen die Multiplizitätsproblematik behandelt werden muss. Ein Schwerpunkt liegt hierbei auf den konfirmatorischen statistischen Tests, welche für die zulassungsrelevanten Phase III-Studien von zentraler Bedeutung sind. Es werden jedoch auch Beispiele aus anderen Phasen der klinischen Entwicklung diskutiert.

Auf ein Kapitel zu gemeinsamen statistischen Grundlagen folgen die zwei zentralen Kapitel der Arbeit. Kapitel 3 umfasst Arbeiten zu *multiplen Vergleichsmethoden*. Diese Methoden sind geeignet für die Analyse einiger weniger Endpunkte, welche alle von individueller Bedeutung sind, wie z.B. verschiedene klinische Endpunkte einer Phase III-Studie. Kapitel 4 behandelt Verfahren aus der *multivariaten Statistik*. Solche Verfahren können auch im Falle einer großen Anzahl von Endpunkten eingesetzt werden. Sie sind allerdings weniger geeignet, wenn Aussagen über einzelne Endpunkte gewünscht werden. Ihre Bedeutung liegt daher eher im Bereich der exploratorischen klinischen Studien in Phasen I und II (z.B. der simultanen Analyse vieler Laborparameter), aber auch im Bereich der Genexpressionsanalyse.

Während die Arbeiten in Kapiteln 3 und 4 statistische Tests behandeln, werden in Kapitel 5 zwei Arbeiten zur Adjustierung von multiplen Effektschätzungen bzw. simultanen Konfidenzintervallen diskutiert.

References

Ahrens, H. and Läuter, J. (1974, 1981): *Mehrdimensionale Varianzanalyse*. Akademie Verlag, Berlin.

Anderson, T.W. (1958, 1984, 2003): *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Bauer P. (1989): Multistage testing with adaptive designs (with Discussion). *Biometrie und Informatik in Medizin und Biologie* **20**, 130148.

Bauer P. and Köhne K. (1994): Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 10291041. Correction: *Biometrics* **52**,380.

Bebu I., Luta G. and Dragalin V. (2010): Likelihood inference for a two-stage design with treatment selection. *Biometrical Journal* **52**, 811-822.

Bellman, R.E. (1957): *Dynamic programming*. Princeton University Press.

Benjamini, Y. and Hochberg, Y. (1995): Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289300.

Benjamini, Y. and Yekutieli, D. (2001): The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165-1188.

Bowden J., Brannath, W. and Glimm E. (2013): Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. Accepted by *Statistics in Medicine*. Published online on 21 July 2013. Early View: DOI: 10.1002/sim.5920.

Bowden J. and Glimm E. (2008): Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* **50**, 515527.

Bowden J. and Glimm E. (2013): Conditionally unbiased and near unbiased estimation for multi-stage drop-the-losers designs. *Accepted by Biometrical Journal*.

Box, G. E. P. (1954): Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* **25**, 290302.

Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009) : A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **28**: 586-604.

Bretz, F., König, F., Brannath, W., Glimm, E. and Posch, M. (2009): Adaptive designs for confirmatory clinical trials. *Statistics in Medicine* **28**: 1181-1217.

Bretz F., Posch M., Glimm E., Klinglmueller F., Maurer W. and Rohmeyer K. (2011): Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. *Biometrical Journal* **53**, 894-913.

Cohen A. and Sackrowitz H. (1989): Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* **8**, 273278.

Collett, D. (1994): *Modelling Survival Data in Medical Research*. Chapman& Hall, London.

Dempster, A.P. (1958): A high dimensional two sample significance test. *Annals of Mathematical Statistics* **29**, 9951010.

Dempster, A.P. (1960). A significance test for the separation of two highly multivariate two samples. *Biometrics* **16**, 4150.

Ding, G.-C., Smalla, K., Heuer, H. and Kropf, S. (2012): A new proposal for a principal component-based test for high-dimensional data applied to the analysis of PhyloChip data. *Biometrical Journal* **54**: 94-107.

Di Scala, L. and Glimm, E. (2011): Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* **30**, 3067–3081. Erratum: *Statistics in Medicine* **32**.

Dmitrienko A., Bretz F., Westfall P.H., Troendle J., Wiens B.L., Tamhane A.C., Hsu J.C. (2009): Multiple testing methodology. In *Multiple testing problems in pharmaceutical statistics*, Dmitrienko A., Tamhane A.C., Bretz F. (eds). Chapman & Hall/CRC, Boca Raton.

Efron, B. (2010): *Large-Scale Inference*. Cambridge University Press.

Efron, B., and Tibshirani, R. J. (1993): *An introduction to the bootstrap*. Chapman & Hall, New York.

Eszlinger, M., Krohn, K., Berger, K., Läuter, J.,Kropf, S., Beck, M., Führer, M. and Paschke, R. (2005): Gene expression analysis reveals evidence for increased expression of cell cycle-associated genes and Gq-protein-protein kinase C signaling in cold thyroid nodules. *Journal of Clinical Endocrinology & Metabolism* **90**, 1163-1170.

Fang, K.-T. and Zhang, Y.-T. (1990): *Generalized Multivariate Analysis*. Springer, Berlin.

Gelman, A., Carlin, J.B., Stern, H. S. and Rubin, D.B. (2004): *Bayesian Data Analysis*. 2nd edition. Chapman & Hall / CRC, Boca Raton.

Glimm, E. (1999): Güte- und Optimalitätseigenschaften stabiler multivariater Verfahren. PhD Dissertation, University of Magdeburg.

Glimm, E. (2000): Spherical tests in balanced multivariate mixed models. *Biometrical Journal* **42**, 937-950.

Glimm, E., Heuer, H., Engelen, B., Smalla, K. and Backhaus, H. (1997): Statistical comparisons of community catabolic profiles. *Journal of Microbiological Methods* **30**, 71-80.

Glimm, E. and Läuter, J. (2002): On the admissibility of stable multivariate tests. *Journal of Multivariate Analysis* **86**, 254-265.

Glimm, E. and Läuter, J. (2010): Directional multivariate tests rejecting null and negative effects in all variables. *Biometrical Journal* **52**, 757-770.

Glimm, E., Maurer, W., and Bretz, F. (2010): Hierarchical testing of multiple endpoints in group sequential trials. *Statistics in Medicine* **29**, 219–228.

Glimm, E., Srivastava, M.S. and Läuter, J. (2002): Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics- Simulation and Computation* **31**, 589-604.

Hochberg Y., Tamhane A.C. (1987): *Multiple comparison procedures*. Wiley, New York.

Holm S (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Jour-*

nal of Statistics **6**, 65–70.

Hommel G., Bretz F. and Maurer W. (2007): Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* **26**, 4063–4073.

Hotelling, H. (1931): The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**, 360-378.

Hsu, J. (1996): *Multiple Comparisons: Theory and Methods*. Chapman and Hall/CRC, Boca Raton.

Hung H.M.J., Wang S.J. and O'Neill R. (2007): Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics* **17**, 1201–1210.

Jennison, C. and Turnbull, B.W. (2000): *Group-sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, Boca Raton.

König F., Brannath W., Bretz F. and Posch, M. (2008): Adaptive Dunnett tests for treatment selection, *Statistics in Medicine* **27**, 1612-1625.

Kropf, S. (2000): *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*. Shaker Verlag, Aachen.

Kropf, S., Heuer, H., Grüning, M. and Smalla, K. (2004): Significance test for comparing complex microbial community fingerprints using pairwise similarity measures. *Journal of Microbiological Methods* **57**: 187–195.

Läuter, J. (1992): *Stabile Multivariate Verfahren*. Akademie Verlag, Berlin.

Läuter J. (1996). Exact *t* and *F* tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964–970.

Läuter, J. and Glimm, E. (2005): A theorem on the principal components inference. *Statistics* **39**, 207-219.

Läuter, J., Glimm, E. and Eszlinger, M. (2005): Search for relevant sets of variables in a high-dimensional setup keeping the FWE criterion. *Statistica Neerlandica* **59**, 298-312.

Läuter J., Glimm E., and Kropf S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal* **38**, 5–23. Erratum: *Biometrical Journal* **40**, 1015.

Läuter J., Glimm E., and Kropf S. (1998): Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26**, 1972–1988. Correction: *Annals of Statistics* **27**, 1441.

Läuter, J., Horn, F., Rosolowski, M., and Glimm, E. (2009): High-dimensional data analysis: selection of variables, data compression and graphics-application to gene expression. *Biometrical Journal* **51**, 235-251.

Lehmacher W. and Wassmer G. (1999): Adaptive sample size calculations in group sequential trials. *Biometrics* **55**, 12861290.

Lehmann, E.L. and Romano, J.P. (2005): *Testing Statistical Hypotheses.* 3rd edition. Springer, New York.

Marcus, R., Peritz, E., Gabriel, K.R. (1976): On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655-660.

Maurer, W., Glimm, E. and Bretz, F. (2011): Multiple and repeated testing of primary, coprimary and secondary hypotheses. *Statistics in Biopharmaceutical Research* **3**, 336-352.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979): *Multivariate Analysis*. Academic Press, London.

Meinshausen, N. and Bühlmann, P. (2006): High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** 1436-1462.

Müller H.H. and Schäfer H. (2001): Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**, 886891.

Müller H.H. and Schäfer H. (2004): A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* **23**, 24972508.

O'Brien P.C. and Fleming T.R. (1979): A multiple testing procedure for clinical trials. *Biometrics* **5**, 549-556.

Pawitan, Y. (2001): *In all likelihood: Statistical Modeling and Inference and Using Likelihood*. Clarendon Press, Oxford.

Perlman, M.D. (1969): One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549-567.

Pocock S.J. (1977): Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191–199.

Proschan M.A. and Hunsberger S.A. (1995): Designed extension of studies based on conditional power. *Biometrics 1995* **51**, 13151324.

Puri, M.L. and Sen, P.K. (1971): *Nonparametric Methods in Multivariate Analysis*. Wiley, New York.

Putter, J. and Rubenstein, D. (1968): *Technical report tr165: On estimating the mean of a selected population*. University of Wisconsin, Statistics Department, Wisconsin.

Romano, J.P. and Wolf, M. (2007): Control of generalized error rates in multiple testing. *Annals of Statistics* **35**, 1378–1408.

Schaafsma, W. and Smid, L.J. (1966): Most stringent somewhere most powerful tests against alternatives restricted by a number of inequalities. *Annals of Mathematical Statistics* **37**, 1161-1172.

Schuster, E., Kropf, S. and Roeder, I. (2004): Micro Array Based Gene Expression Analysis using Parametric Multivariate Tests per Gene A Generalized Application of Multiple Procedures with Data-driven Order of Hypotheses. *Biometrical Journal* **46**, 687-698.

Searle, S.R. (1971): *Linear Models*. Wiley, New York.

Silvapulle, M.J. (1997): A curious example involving the likelihood-ratio test against one-sided hypotheses. *The American Statistician* **51**, 178-180.

Silvapulle, M.J. and Sen, P.K. (2005): *Constrained Statistical Inference*. Wiley, New York.

Smalla, K., Oros-Sichler, M., Milling, A., Heuer, H., Baumgarte, S., Becker, R., Neuber, G., Kropf, S. Ulrich, A. and Tebbe, C.C. (2007): Bacterial diversity of soils assessed by DGGE, T-RFLP and SSCP fingerprints of PCR-amplified 16S rRNA gene fragments: do the different methods provide similar results? *Journal of Microbiological Methods* **69**:470-479.

Speed, T.S. (ed.) (2003): *Statistical analysis of gene expression data*, Chapman & Hall, Boca Raton.

Srivastava, M.S. and Khatri, C.G. (1979): *An Introduction to Multivariate Statistics*. Elsevier,

New York.

Srivastava, M.S. (2002): *Methods of Multivariate Statistics*. Wiley, New York.

Srivastava, M.S. (2007): Multivariate theory for Analyzing High Dimensional Data. *Journal of the Japan Statistical Society* **37**, 53-86.

Srivastava, M.S. (2009): A Review of Multivariate Theory for High Dimensional Data with Fewer Observations. In: *Advances in Multivariate Statistical Methods*. (A. SenGupta, ed.), World Scientific Publishing, Singapore, 25-52.

Srivastava, M.S. and Du, M. (2008): A test for the Mean Vector with Fewer Observations than the Dimension. *Journal of Multivariate Analysis* **99**, 386-402.

Srivastava, M.S., Hirotsu, C., Aoki, S. and Glimm, E. (2001): One-sided tests. In: *Data Analysis from Statistical Foundations - A Festschrift in Honour of the 75th Birthday of D.A.S. Fraser*. Nova Science Publishers, 387-401.

Storey, J.D. (2002): A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B (Methodological)* **64**, 479-498.

Su, T.-L., Glimm, E., Whitehead, J. and Branson, M. (2012): An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial. *Biopharmaceutical Statistics* **11**, 107-117.

Tamhane A.C., Mehta C.R. and Liu L.(2010): Testing a primary and a secondary endpoint in a group sequential design. *Biometrics* **66**, 1174–1184.

Tang, D.I., Gnecco, C. and Geller, N. (1989): An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* **76**, 577-583.

van der Laan, M. J. and Dudoit, S. (2007): *Multiple Testing Procedures with Applications to Genomics*. New York: Springer.

Wang, Y. and McDermott, M. P. (1998): Conditional likelihood ratio test for a nonnegative normal mean vector. *Journal of the American Statistical Association* **93**, 380-386.

Westfall, P.H. and Young, S.S. (1993): *Resampling-based multiple testing*. John Wiley & Sons, New York.

Appendix: List of submitted publications

1. Maurer, W., Glimm, E. and Bretz, F. (2011): Multiple and repeated testing of primary, coprimary and secondary hypotheses. *Statistics in Biopharmaceutical Research* **3**, 336-352.

2. Bretz F., Posch M., Glimm E., Klinglmueller F., Maurer W. and Rohmeyer K. (2011): Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes or parametric tests. *Biometrical Journal* **53**, 894-913.

3. Glimm, E., Maurer, W., and Bretz, F. (2010): Hierarchical testing of multiple endpoints in group sequential trials. *Statistics in Medicine* **29**, 219–228.

4. Di Scala, L. and Glimm, E. (2011): Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* **30**, 3067–3081. Erratum: *Statistics in Medicine* **32**.

5. Su, T.-L., Glimm, E., Whitehead, J. and Branson, M. (2012): An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial. *Biopharmaceutical Statistics* **11**, 107-117.

6. Glimm, E. (2000): Spherical tests in balanced multivariate mixed models. *Biometrical Journal* **42**, 937-950.

7. Glimm, E., Srivastava, M.S. and Läuter, J. (2002): Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics- Simulation and Computation* **31**, 589-604.

8. Glimm, E. and Läuter, J. (2010): Directional multivariate tests rejecting null and negative effects in all variables. *Biometrical Journal* **52**, 757-770.

9. Läuter, J., Glimm, E. and Eszlinger, M. (2005): Search for relevant sets of variables in a high-dimensional setup keeping the FWE criterion. *Statistica Neerlandica* **59**, 298-312.

10. Bowden J. and Glimm E. (2008): Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* **50**, 515527.

11. Bowden J., Brannath, W. and Glimm E. (2013): Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach. *Accepted by Statistics in Medicine*.

12. Läuter, J. and Glimm, E. (2005): A theorem on the principal components inference. *Statistics* **39**, 207-219.

Multiple and Repeated Testing of Primary, Coprimary, and Secondary Hypotheses

Willi MAURER, Ekkehard GLIMM, and Frank BRETZ

In confirmatory clinical trials the Type I error rate must be controlled for claims forming the basis for approval and labeling of a new drug. Strong control of the familywise error rate is usually needed for hypotheses related to the primary endpoint(s). For hypotheses related to secondary endpoint(s) which are only of interest if the corresponding "parent" primary null hypotheses have been rejected, less strict error rate control might be sufficient. We review and extend procedures for families of primary and secondary hypotheses when either at least one of the primary hypotheses or all coprimary hypotheses must be rejected to claim success for the trial. Such families of hypotheses arise naturally from comparing several treatments with a control, combined noninferiority and superiority testing for primary and secondary variables, the presence of multiple primary or secondary endpoints or any combination thereof. We show that many of the procedures proposed in the literature follow a common underlying principle and in some cases can be improved. In addition we present some general results on Type I error rates for the different families and subfamilies of hypotheses and their relation to group-sequential testing of multiple hypotheses.

Key Words: Closed testing; Confirmatory trial; Familywise error rate; Gatekeeping procedure; Graphical approach; Simes' test.

1. Introduction

In confirmatory clinical trials it is common practice to control the Type I error rate for claims forming the basis for approval and labeling of a new drug. Regulatory guidance and practice ask for a strong control of the familywise error rate (FWER), at least for the hypotheses related to the primary endpoints. That is, given a family of n null hypotheses $\mathcal{F} = \{H_1, \ldots, H_n\}$, the probability to falsely reject at least one true null hypothesis should be bounded by α under any configuration of true and false null hypotheses. Hypotheses related to the secondary endpoint(s) serving only for the qualification of an established primary effect are mainly of interest if the corresponding "parent" primary null hypotheses have been rejected before. Efficient control of Type I error rates can be challenging when families of primary and secondary hypotheses are to be considered or when there are coprimary endpoints which must all be significant for a successful claim. Such families of hypotheses arise naturally when additional multiplicity stems from comparing several treatments with a control, combined noninferiority and superiority testing for primary and secondary variables, the presence of multiple primary or secondary endpoints or any combination thereof. Depending on the context, hypotheses may be arranged in a hierarchy with some hypotheses being equally important and others being formally tested only conditional on the rejection of more important ones. Some of the arising issues have been discussed by Hung and Wang (2009, 2010) and Bretz, Maurer, and Gallo (2009a).

Statistics in Biopharmaceutical Research
2011, Vol. 3, No. 2
DOI: 10.1198/sbr.2010.10010

An example of a procedure that takes into account such partial orderings is given by first testing the family of primary hypotheses with a method controlling the FWER at level α, followed by testing only those secondary hypotheses for which the "parent" hypotheses have been rejected before. Bretz et al. (2009a) mentioned that such procedures control the Type I error at level α neither for the union of primary and secondary hypotheses nor for the secondary hypotheses alone, even if the testing procedure conditionally controls the FWER at level α for the secondary hypotheses. They also give upper bounds of Type I error rates (without proof) for the special case of one descendant secondary hypothesis per primary hypothesis. For this and similar cases, it may be sensible to control the FWER separately for the families of primary and secondary hypotheses as well as for their union or other Boolean functions of them. In this article, we derive upper bounds for the resulting error rates for a more general situation than considered by Bretz et al. (2009a) and discuss how the tightness of these bounds depends on distributional assumptions and logical interdependencies of the hypotheses. We also show that the scenarios for which the upper bounds are attained are closely related to those arising in repeated testing problems of primary and secondary hypotheses in group sequential trials (Glimm, Maurer and Bretz 2010; Tamhane, Mehta, and Liu 2010).

Testing procedures that control the FWER for the family of all hypotheses under consideration and account for partial ordering or logical dependencies are increasingly discussed in the literature. Many of them, in particular so-called gatekeeping procedures, are based on closed testing procedures using Bonferroni tests for the intersection hypotheses. For some of the cases where the underlying logical dependence structure is that of a family of primary hypotheses where each of its members is the "parent" of logically associated secondary hypotheses gatekeeping test procedures have been proposed by several authors, for example, Dmitrienko, Offen, and Westfall (2003); Hommel, Bretz, and Maurer (2007). Bretz, Maurer, Brannath, and Posch (2009b) proposed sequentially rejective graphical approaches that are flexible and easy to communicate. It was shown by Bretz, Maurer, and Hommel (2010) how this approach can be applied to the situation of multiple tests with primary and secondary endpoints. In this article we give a formal definition of desirable properties (succession and consistency) of test procedures for the logical dependence of primary hypotheses with descendant secondary hypotheses and show how they can be generated by means of the graphical approach. In extension of proposals by Quan, Capizzi, and Zhang (2009) we also show how some of the arising test procedures can be further improved by exploiting the correlation of the test statistics or a high likelihood of similar effects. The special case of multiple coprimary endpoints that all must show a statistically significant effect for the trial to be successful in a multiple and/or repeated hypotheses testing situation will be considered as well and a novel test procedure based on the Simes test is presented.

2. Families of Primary and Secondary Hypotheses

In the following we consider testing a family of hypotheses $\mathcal{F}$ that can be partitioned into a primary family $\mathcal{F}_p$ and a secondary family $\mathcal{F}_s$, that is, $\mathcal{F} = \mathcal{F}_p \cup \mathcal{F}_s$, and $\mathcal{F}_p \cap \mathcal{F}_s = \emptyset$. Note that $\mathcal{F}_p$ and $\mathcal{F}_s$ are simply sets of the corresponding labels (or names) of the hypotheses. For example, if $H_{p,1} : \mu_1 = 0$ and $H_{s,1} : \mu_2 = 0$ happen to coincide in a concrete model, for example, because the correlation between the respective test statistics is 1, then they are still treated as two distinct elements of $\mathcal{F}$.

We assume that for each hypothesis $H_{p,i} \in \mathcal{F}_p$ there is a unique set of "descendant" secondary hypotheses $\mathcal{F}_{s(i)}$, $i \in N$, where $N = \{1, \ldots, n\}$ denotes the index set of primary hypotheses. We denote by n the number of distinct primary hypotheses, irrespective if this relates to number of treatments, number of endpoints or other categories, depending on the application. Further, we assume that each secondary hypothesis in $\mathcal{F}_s$ must have at least one "parent" primary hypothesis, that is, $\mathcal{F}_s = \bigcup_{i=1}^{n} \mathcal{F}_{s(i)}$. A secondary hypothesis $H_{s(i),j} \in \mathcal{F}_{s(i)}$ is only of interest (in a confirmatory sense) if one of the respective "parent" primary hypothesis $H_{p,i}$ is rejected. The erroneous rejection rate of the primary null hypotheses needs to be controlled at level α_p ($= \alpha$, say). In the sequel, we describe several configurations of primary and secondary hypotheses satisfying the above conditions. Each case is illustrated by a design option satisfying different objectives for the following common example of a trial to be planned in ophthalmology.

Example: A 6-month multicenter, parallel group, double-masked phase III study is planned in patients with quiescent, noninfectious uveitis for comparing an experimental compound administered on top of standard of care treatment versus standard of care (immunosuppressive and anti-inflammatory treatment) for maintaining uveitis suppression when reducing systemic immunosuppression. The following endpoints are considered primary or secondary depending on the objectives and design of the trial:

P_1**:** Recurrence of active uveitis within the treatment period (binary composite endpoint based on the occurrence of S_2 or S_3).

P_2: Change in best corrected visual acuity from baseline.

S_1: Change in immunosuppressive medication score from baseline to 6 month.

S_2: Decrease in best corrected visual acuity of more than 10 ETDRS letters in either eye (binary).

S_3: An increase in vitreous haze of at least two steps (binary).

1. *One primary and one or more secondary endpoints in multiple treatments.* Assume that $n \geq 2$ treatments (e.g., n doses) are tested for superiority against control for a single primary endpoint and $m \geq 1$ secondary endpoints. The one-sided primary null hypotheses $H_{p,i}, i = 1, \ldots, n$, state that the effect of treatment i on the primary variable is not larger than that of control. The rejection of at least one of these hypotheses is assumed to be a prerequisite for a "positive" study. The family $\mathcal{F}_{s(i)}$ of descendant secondary null hypotheses of $H_{p,i}$ consists of the one-sided null hypotheses of nonsuperiority of treatment i in the m secondary variables. Hence, $\mathcal{F}_{s(i)} \bigcap \mathcal{F}_{s(i')} = \emptyset$ for $i \neq i'$. We assume that rejection of at least one of the primary and associated secondary hypotheses is necessary for a positive study that may allow additional label claims. A similar test situation arises if the multiple comparisons refer to different, possibly overlapping subgroups of the analysis population.

   Example: The effect of three dose regimens is compared to standard of care with respect to the only primary endpoint P_1. Key-secondary endpoints are S_1 and S_2.

2. *Multiple primary and multiple secondary endpoints.* Here, we assume n multiple primary endpoints (usually $n = 2$) where at least one of them needs to be significant for a positive study. Each primary endpoint has its own set of descendant secondary hypotheses. This situation may occur, for example, when the primary endpoints are composites of different events and the descendant secondary endpoints are the components of the parent primary composite endpoint. Different primary variables may have a common secondary variable as descendant such that $\mathcal{F}_{s(i)} \bigcap \mathcal{F}_{s(i')} \neq \emptyset$ for some $i \neq i'$.

   Example: Only one dose regimen is compared to standard of care. However, two primary endpoints P_1 and P_2 with descendant secondary endpoint(s) each are considered. The trial is successful if for at least one of them superiority over standard of care can be shown. The descendant secondary endpoints of P_1 are its components S_2 and S_3. The only descendant secondary endpoint of P_2 is S_2.

3. *Primary and secondary comparisons of multiple endpoints.* As in case 2 there are n multiple primary endpoints (usually $n = 2$) where at least one of them needs to be significant for a positive study. The hierarchy of primary and secondary hypotheses, however, is given by a clear order of importance in the comparison of two treatment arms vs. one control or of one test treatment vs. two controls (e.g., placebo and an active control). As in case 1, then $\mathcal{F}_{s(i)} \bigcap \mathcal{F}_{s(i')} = \emptyset$ for $i \neq i'$. The particular situation of two doses vs. control where the comparison of the higher dose with control on the n endpoints is considered as the primary one was discussed, for example, by Quan et al. (2009).

   Example: In addition to a dose of primary interest a lower dose is tested in the trial. The lower dose might be recommended for use if it is efficacious and causes less safety problems than the higher dose. As in Case 2 for each dose (the primary higher dose and the secondary lower dose) two primary endpoints, P_1 and P_2, are assessed.

4. *Noninferiority and superiority testing of multiple endpoints and/or multiple treatment arms.* The hierarchy of primary and secondary hypotheses arises naturally from noninferiority being a prerequisite for showing superiority over the same control. Given there are hypotheses related to multiple (primary) endpoints or multiple treatment arms to be tested, the family of primary hypotheses consists of those related to noninferiority and the secondary hypotheses are those related to superiority. Generally there is at most one descendant (secondary) hypothesis to each parent (primary) hypothesis and vice versa; see Hung and Wang (2010) for further discussion.

   Example: The experimental treatment is not administered on top of standard of care but the three doses mentioned in Case 1 alone are compared to standard of care (immunosuppressant plus steroids) on one primary variable P_1 with the aim to show noninferiority (primary hypotheses) or even superiority (secondary hypotheses) with respect to standard of care.

5. *Repeated testing of a primary and a secondary hypothesis.* One primary hypothesis H_p and one secondary hypothesis H_s are tested repeatedly at $n-1$

interim and a final analysis in a group sequential trial. Formally one can consider the up to n tests of the two hypotheses as being related to different hypotheses $H_{p,i}$ and $H_{s,i}$, $i = 1, \ldots, n$. Practically, one is only interested in the "overall" hypotheses $H_p = \bigcap_{i=1}^{n} H_{p,i}$ and $H_s = \bigcap_{i=1}^{n} H_{s,i}$, that is, the rejection of one of the hypotheses $H_{p,i}$ at any of the n analyses leads to the rejection of H_p. Similarly, the rejection of $H_{s,i}$ at any of the n analysis leads to the rejection of H_s. Since the Type I error rate for an overall (primary or secondary) hypothesis is bounded by the FWER of the respective family, the general results on error rate control for various test procedures presented in the next section apply also to the situation of repeated testing of primary and secondary hypotheses. Since in this case up to n tests at the n interim analyses are done in a strictly sequential manner starting with $i = 1$ and an overall hypothesis is rejected as soon as a rejection is possible at one of the interim analyses with a potential stop of the entire trial, the general results can be refined, taking also into account the particular correlation structure between the test statistics at different information fractions. This case was extensively discussed by Glimm et al. (2010) and Tamhane et al. (2010).

Example: One dose of the experimental drug is compared to standard of care on a primary and a secondary endpoint (P_1 and S_1) in a group sequential design with one interim analysis.

3. Consistent Tests and Error Rates for Families of Primary and Secondary Hypotheses

Care has to be taken when applying gatekeeping procedures to avoid properties which are unnecessary or seem to be "against common sense," see, among others, Hung and Wang (2009) for examples. To avoid procedures that suffer from such deficiencies we require the following properties:

(a) a secondary hypothesis can only be rejected if at least one of its parent primary hypotheses is rejected, and

(b) the retention of a secondary hypothesis cannot preclude the rejection of a primary hypothesis.

We call procedures that satisfy conditions (a) and (b) *consistent* and procedures that satisfy condition (a) *successive*. In Section 4 we investigate successive procedures in more detail. In the following we introduce three test procedures $S_j, j = 1,2,3$. S_1 and S_2 are consistent, S_3 has only property (b). Afterwards we give boundaries of Type I error rates for the various families of hypotheses. All three procedures have in common that the primary family $\mathcal{F}_p$ is tested with a testing procedure controlling the FWER at level α_p. We focus on *closed* test procedures because of their desirable properties (Bauer 1991) but this is not a prerequisite for the results we are presenting in the sequel.

Assume that trial results are available and the test procedure on the primary hypotheses is performed. Let $r\mathcal{F}_p \subseteq \mathcal{F}_p$ be the set of primary hypotheses that have been *r*ejected and $c\mathcal{F}_s = \bigcup(\mathcal{F}_{s(i)} : H_{p,i} \in r\mathcal{F}_p)$ denote the set of secondary hypotheses with rejected "parent" primary hypotheses (i.e., the *c*andidate set for rejection).

In the sequel we denote by $H_{p,i}$ a primary hypothesis, $H_{s,j}$ a secondary hypothesis and by $H_{s(i),j}$ a secondary hypothesis that is a descendant of $H_{p,i}$.

S_1: Test all secondary hypotheses $H_{s,j} \in c\mathcal{F}_s$ using a closed test procedure on $c\mathcal{F}_s$ at level α_s. Retain all secondary hypotheses with retained primary hypotheses.

S_2: Test each $H_{s,j} \in c\mathcal{F}_s$ at level α_s and retain secondary hypotheses with retained primary hypotheses.

S_3: Test each $H_{s,j} \in \mathcal{F}_s$ at level α_s, irrespective of whether or not a primary hypothesis is rejected.

Let $R_j(H)$ denote the event that a null hypothesis H is rejected with strategy $S_j, j = 1,2,3$, and $W_j(\mathcal{F})$ the event that a true null hypothesis from the family $\mathcal{F}$ is rejected with strategy S_j. If no null hypothesis in $\mathcal{F}$ is true, we set $W_j(\mathcal{F}) = \emptyset$. Let $P(W(\mathcal{F}))$ denote the probability that the event W occurs for a given scenario of true and false hypotheses in family $\mathcal{F}$. Note that $\max(P(W(\mathcal{F})))$ is the FWER for a hypotheses family $\mathcal{F}$, where the maximum is taken over all possible scenarios of true and false hypotheses.

All three test procedures above have the property that the test decisions of the secondary hypotheses do not influence those of the primary hypotheses. Hence,

$$P(W_j(\mathcal{F}_p)) \leq \alpha_p, j = 1,2,3, \tag{1}$$

that is, for the family of primary hypotheses the FWER is protected at level α_p. For a given scenario of true and false hypotheses and trial results, rejection of a hypothesis $H_{s,j}$ with strategy S_1 implies rejection with strategy S_2, which in turn implies rejection with S_3. Consequently,

$$\begin{aligned} R_1\left(H_{s(i),j}\right) &\subseteq R_2\left(H_{s(i),j}\right) \\ &\subseteq R_3\left(H_{s(i),j}\right), \quad i \in N, H_{s(i),j} \in \mathcal{F}_{s(i)} \end{aligned}$$

and hence

$$W_1(\mathcal{F}_s) \subseteq W_2(\mathcal{F}_s) \subseteq W_3(\mathcal{F}_s)$$

for any fixed scenario of true and false primary and secondary hypotheses. With $\mathcal{F} = \mathcal{F}_p \cup \mathcal{F}_s$ this implies

$$W_1(\mathcal{F}) \subseteq W_2(\mathcal{F}) \subseteq W_3(\mathcal{F}).$$

Let m denote the number of different hypotheses in $\mathcal{F}_s$. Then by the above implications, we have

$$P(W_1(\mathcal{F}_s)) \leq P(W_2(\mathcal{F}_s)) \leq P(W_3(\mathcal{F}_s)) \leq m\alpha_s. \quad (2)$$

Further, as $\mathcal{F} = \mathcal{F}_p \cup \mathcal{F}_s$,

$$P(W_3(\mathcal{F})) = P(W_3(\mathcal{F}_p) \cup W_3(\mathcal{F}_s)) \leq \alpha_p + m\alpha_s.$$

due to the Bonferroni inequality.

If each primary hypothesis has at least $h > 0$ descendant secondary hypotheses and every secondary hypothesis has only one parent primary hypothesis, this upper bound can be tightened for the more restrictive procedures S_1 and S_2 to

$$P(W_1(\mathcal{F})) \leq P(W_2(\mathcal{F})) \leq \max(m\alpha_s, \alpha_p + (m-h)\alpha_s), \quad (3)$$

see Appendix A.1 for a proof.

For $\alpha_p \leq \alpha_s$ the inequality (3) reduces to

$$P(W_1(\mathcal{F})) \leq P(W_2(\mathcal{F})) \leq m\alpha_s. \quad (4)$$

Intuitively one would expect that these inequalities are strict, that is, that the FWER for the family of secondary hypotheses is strictly smaller than that for the combined families of primary and secondary hypotheses and that the FWER for applying the more stringent procedure S_1 is strictly smaller than that induced by procedure S_2. However, this is not always the case. Assuming $m\alpha_s \leq 1$, it is always possible to construct examples where the boundaries are tight (i.e., (3) holds as an equality), even for strategy S_1. Such constructions are based on cases where all primary variables are linked in a non-stochastic way such that a significant test result in one of them implies that all other primary tests are non-significant with probability 1. Of course, situations like that are degenerate borderline cases of real testing problems and will never occur in practice.

For the more realistic case of (asymptotically) multivariate normal test statistics associated with $\mathcal{F}_p$ and $\mathcal{F}_s$, there are parameter constellations where boundary (3) is (asymptotically) tight for $n = m = 2$. For $m, n > 2$ and $\alpha_p \leq \alpha_s$, there is no parameter setting for which the upper bound $m\alpha_s$ is reached, but (3) is almost tight without additional restrictions on the parameters of the normal distribution and for small m. Appendix A.2 investigates some corresponding scenarios.

3.1 Testing Two Primary Hypotheses With One Descendant Secondary Hypothesis Each

We consider the situation of testing two primary hypotheses with one descendant secondary hypothesis each. It is also the most elementary situation where the inequalities above are not trivial and where the tightness of the boundaries can be investigated. In the sequel we assume $\alpha_p = \alpha_s = \alpha$. Hence the two primary hypotheses are tested with some closed test at multiple level α. Procedure S_1 tests the family $c\mathcal{F}_s$ with a closed test also at level α, whereas procedure S_2 allows testing each secondary hypothesis in $c\mathcal{F}_s$ at level α.

From Equation (1) it follows

$$P(W_j(\mathcal{F}_p)) \leq \alpha, j = 1,2, \quad (5)$$

and from (3) and (4)

$$P(W_1(\mathcal{F}_s)) \leq P(W_1(\mathcal{F})) \leq P(W_2(\mathcal{F})) \leq 2\alpha \quad (6)$$

for any scenario of true and false hypotheses in $\mathcal{F}$.

These boundaries are tight, that is, there exist scenarios for which $P(W_1(\mathcal{F}_s)) = 2\alpha$. Hence, together with the above inequalities,

$$\begin{aligned} \max P(W_1(\mathcal{F}_s)) &= \max P(W_1(\mathcal{F})) \\ &= \max P(W_2(\mathcal{F})) = 2\alpha, \end{aligned} \quad (7)$$

where the maximum is taken over all possible scenarios of true and false hypotheses and correlations between the test statistics. To show this, assume that the vector of test statistics $(T_{p,1}, T_{s,1}, T_{p,2}, T_{s,2})'$ for the primary and secondary hypotheses, respectively, follows a multivariate normal distribution

$$N\left(\begin{pmatrix} \mu_{p,1} \\ \mu_{s,1} \\ \mu_{p,2} \\ \mu_{s,2} \end{pmatrix}, \begin{pmatrix} 1 & \rho & \tau & \rho\tau \\ \rho & 1 & \rho\tau & \tau \\ \tau & \rho\tau & 1 & \rho \\ \rho\tau & \tau & \rho & 1 \end{pmatrix}\right), \quad (8)$$

where $\tau = \text{corr}(T_{p,1}, T_{p,2}) = \text{corr}(T_{s,1}, T_{s,2})$ denotes the correlation between the primary and the secondary test statistics, respectively, and $\rho = \text{corr}(T_{p,i}, T_{s,i}), i = 1,2$, denotes the correlation between the primary and the descendant secondary test statistic. Note that the assumption of $\text{corr}(T_{p,1}, T_{s,2}) = \text{corr}(T_{p,2}, T_{s,1}) = \rho\tau$ is equivalent to conditional independence of $T_{p,1}$ and $T_{s,2}$ given either $T_{p,2}$ or $T_{s,1}$.

Consider the specific scenario that the alternative hypotheses hold for the primary hypotheses (i.e., $\mu_{p,i} > 0, i = 1,2$), whereas for the secondary hypotheses both null hypotheses hold (i.e., $\mu_{s,i} = 0, i = 1,2$). As a representative for a procedure that has the properties S_1 we assume that the primary hypotheses are tested with a Bonferroni-Holm test (Holm 1979) at level α.

If only one primary hypothesis is rejected, the single descendant secondary hypothesis is tested at level α. Otherwise, if both primary hypotheses are rejected the two descendant hypotheses are tested with a Bonferroni-Holm test at level α.

If now the primary test statistics are completely negatively correlated (i.e., $\tau = -1$) and the secondary test statistics are completely positively correlated with their parent primary test statistics ($\rho = 1, i = 1,2$) then $P(W_1(\mathcal{F}_s)) = 2\alpha$ for all scenarios where $\mu_p = \mu_{p,1} = \mu_{p,2} > 0$ are in the interval $u_{\alpha/2} - u_\alpha \le \mu_p \le 2u_\alpha, i = 1,2$, where u_α denotes the $(1-\alpha)$-quantile of the standard normal distribution. For these means of the primary test statistics, the rejection probability $\pi_{p,i} = P(T_{p,i} \ge u_\alpha), i = 1,2$ of a primary test is between $\Phi(u_{\alpha/2} - 2u_\alpha)$ and $1-\alpha$. The lower bound of this interval is slightly below 2α for the usually used significance levels α. A proof of this is given in the Appendix A.2 together with a general formula to compute the FWER of the secondary hypotheses when the above test procedure is used. If the Bonferroni-Holm procedure is replaced by the simple Bonferroni procedure, $P(W_1(\mathcal{F}_S)) = 2\alpha$ even holds for $u_{\alpha/2} - u_\alpha \le \mu_p \le u_{\alpha/2} + u_\alpha$, showing that regarding the rejection probability for a *secondary* hypothesis, a Bonferroni-Holm adjustment of the *primary* tests does not yield a uniformly more powerful procedure than the simple Bonferroni method.

Clearly the assumption of completely negatively correlated primary test statistics is impossible for some of the configurations described in Section 2. For example, if the correlation τ between the primary test statistics is induced by the design and, as for the situation in case 1, determined by the relative size of the control group versus that of the active groups, then $\tau > 0$. Therefore the above result only shows that it is impossible to tighten the boundaries without taking into account the trial design and making further assumptions on the joint distributions of the endpoints and the associated test statistics.

The different cases we considered in Section 2 can be further detailed for the case of two primary and two secondary hypotheses together with values or ranges for the correlations between test statistics, see Table 1.

1. *One primary with one descendant secondary endpoint for two comparisons.* (a) Comparison of two treatment arms of size n_{trt} versus a control arm of size n_{ctrl}. (b) Comparison of active versus control in two subgroups with sample sizes n_1, n_2, respectively, and a common size n_{comm}. If the subgroups are disjoint, then $n_{\text{comm}} = 0$. If one of the subgroups (1, say) is a subset of the other, then $n_{\text{comm}} = n_1$. In this latter case, the correlations $n_{\text{comm}}/\sqrt{n_1 \cdot n_2}$ in Table 1 become $\sqrt{n_{\text{comm}}/n_2}$.

2. *Two primary endpoints with a secondary endpoint each.*

3. *A primary and secondary comparison of two equally important endpoints with correlation τ.* For the comparisons the same assumptions (a) and (b) are made as for case 1.

4. *Noninferiority (primary) and superiority (secondary) testing* for (a) two endpoints of equal importance with correlation τ, (b) two treatment arms vs. control, and (c) two subgroups. Here the correlation ρ between the noninferiority and the superiority test is 1 if the same analysis population is used for the two tests (e.g., the "full analysis set")

Table 1. Ranges for ρ and τ under the different cases detailed in the text. ρ = correlation between primary and descendant secondary test statistics and τ = correlation between primary and between secondary test statistics.

Case	ρ	τ
1a	$-1 \le \rho \le 1$	$n_{\text{ctrl}}/(n_{\text{trt}} + n_{\text{ctrl}})$
1b	$-1 \le \rho \le 1$	$n_{\text{comm}}/\sqrt{n_1 n_2}$
2	$-1 \le \rho_1, \rho_2 \le 1$	$-1 \le \tau_1, \tau_2 \le 1$
3a	$n_{\text{ctrl}}/(n_{\text{trt}} + n_{\text{ctrl}})$	$-1 \le \tau \le 1$
3b	$n_{\text{comm}}/\sqrt{n_1 n_2}$	$-1 \le \tau \le 1$
4a	1 or $0 \le \rho < 1$	$-1 \le \tau \le 1$
4b	1 or $0 \le \rho < 1$	$n_{\text{ctrl}}/(n_{\text{trt}} + n_{\text{ctrl}})$
4c	1 or $0 \le \rho < 1$	$n_{\text{comm}}/\sqrt{n_1 n_2}$
5a	$-1 \le \rho \le 1$	$\sqrt{I_1}$
5b	$\sqrt{n_{\text{comm}}/n_{\text{tot}}}$	$\sqrt{I_1}$
5c	1	$\sqrt{I_1}$

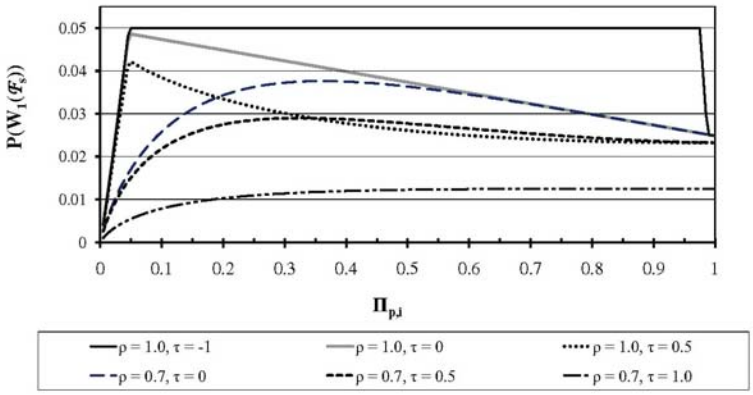


Figure 1. FWER of procedure S_1 on $\mathcal{F}_s$, that is, $P(W_1(\mathcal{F}_s))$ with $\alpha_s = \alpha_p = 0.025$, as a function of $\pi_{p,i}$ for $\alpha = 0.025$ and selected values of ρ and τ.

or $0 \leq \rho \leq 1$ if the set for noninferiority is a subpopulation of the population used for the superiority test (e.g., the per protocol population).

5. *Group sequential testing of a primary and a secondary hypothesis with one interim analysis.* The correlation τ between the test statistics depends on the information fraction I_1 available at the interim analysis. For example, if n_{tot} is the total sample size or number of events and n_{int} the respective number available at the interim analysis then $I_1 = n_{\text{int}}/n_{\text{tot}}$. For the primary/secondary hierarchy we can consider here (a) a primary and secondary endpoint, (b) the total population and a subpopulation, and (c) noninferiority and superiority testing in the same population.

The FWER can only reach the upper bound of 2α if $\tau = -1$. Unfortunately, a substantial lowering of this bound is usually not attainable in practice. For example, if strategy S_1 is applied using the Bonferroni-Holm procedure and if

(i) $\tau = 0$, that is, the two primary test statistics are independent,

(ii) the local power of the primary test at level $\frac{\alpha}{2}$ is α, that is, $P(T_{p,i} \geq u_{\alpha/2}) = \alpha$,

(iii) $\rho = 1$, that is, the primary and secondary variables are perfectly correlated,

then the general result (A.4) in Appendix A.2 implies that $P(W_1(\mathcal{F}_s)) \approx 2\alpha - 2\alpha^2$ for typically used test levels. Hence, $P(W_1(\mathcal{F}_s))$, $P(W_1(\mathcal{F}))$ and $P(W_2(\mathcal{F}))$ can be close to 2α in realistic situations for Cases 1b, 2, 4a, and 4c.

Figure 1 displays $P(W_1(\mathcal{F}_s))$, the actual FWER for the family of secondary hypotheses under procedure S_1 using Bonferroni-Holm tests at level $\alpha = 0.025$ for the primary hypotheses and for the family of secondary hypotheses with rejected primary hypotheses. Note that in this case $P(W_1(\mathcal{F}_s)) = P(W_1(\mathcal{F}))$ since the primary hypotheses are assumed to be false. The FWER is plotted in dependence of the rejection probability $\pi_{p,i}$ of the primary hypotheses (assuming $\alpha = 0.025$) and for selected correlations ρ and τ between and within primary and secondary endpoints. Note that the structural correlation $\tau = 0.5$ occurs, for example, in cases (1a) and (4b) if treatment arm and control arms are of the same size and for cases with subgroup analyses or interim analyses if $n_{\text{comm}}/\sqrt{n_1 n_2} = 0.5$ or $I_1 = 0.25$, respectively. The values of the FWER have been computed using expression (A.3) in the Appendix by means of numerical integration methods described in Genz and Bretz (2009) and implemented in many statistical software packages like R or SAS.

4. Partially Hierarchical Testing Procedures Protecting the FWER

The results presented in the preceding section show that there is no substantial reduction of the upper bound for the FWER if unconditional testing of the secondary hypotheses (procedure S_3) is replaced by testing only secondary hypotheses with rejected primary hypotheses. Nevertheless such consistent procedures can be of inter-

est if the main goal is to protect the FWER for the family of primary hypotheses at given level α_p because they allow us to perform an exhaustive closed test on the primary hypotheses independent of the secondary hypotheses and hence have a high power to reject all false primary null hypotheses. In this section we will explore procedures that protect the FWER for the family $\mathcal{F}$ of primary and secondary hypotheses and which are consistent as well. As all parallel gatekeeping procedures, they entail a small price to pay, because secondary hypotheses can be tested without prior rejection of all primary hypotheses. This slightly reduces the power to reject all false primary hypotheses. However, the power to reject at least one false primary hypothesis can be as high as for any of the procedures discussed in Section 3.

4.1 Consonant and Successive Bonferroni-based Closed Testing Procedures

Consider the problem of testing $k = n + m$ elementary primary and secondary hypotheses $H_1, \ldots, H_k$ and let $K = \{1, \ldots, k\}$ denote the associated index set.We assume that the elementary hypotheses satisfy the free combination condition, that is, for any subset $I \subseteq K$ the simultaneous truth of $H_i, i \in I$, and falsehood of the remaining hypotheses is a possible event. The only situations considered in Section 2 where this might not apply are Cases 4 and 5, for Case 4 given that noninferiority and superiority are tested on the same analysis population and hence superiority implies noninferiority. In this situation, however, the result below still apply since the nesting of primary hypotheses within secondary hypotheses is fully exploited when a procedure is consistent.

Applying the closure principle (Marcus, Peritz, and Gabriel 1976) leads to testing the intersection hypotheses $H_I = \bigcap_{i \in I} H_i$, $I \subseteq K$. For each intersection hypothesis H_I we assume a set of levels $\boldsymbol{\alpha}(I) = \{\alpha_i(I), i \in I\}$ such that $0 \leq \alpha_i(I) \leq \alpha$ and $\sum_{i \in I} \alpha_i(I) \leq \alpha$ where α is the intended significance level of the test for the combined family of primary and secondary hypotheses. In the sequel we assume that $\sum_{i \in I} \alpha_i(I) = \alpha, I \subseteq K$, that is, the intended level is exhausted for all intersection hypotheses. The relative weights $w_i(I) = \alpha_i(I)/\alpha, i \in I \subseteq K$ quantify the relative importance of the hypotheses H_i included in the intersection H_I.

We assume that each intersection hypothesis is tested with a weighted Bonferroni test, that is, we reject H_I if $p_i \leq \alpha_i(I)$ for at least one $i \in I$. This defines the class $\mathcal{B}$ of all closed testing procedures that use weighted Bonferroni tests for each intersection hypothesis. Many standard multiple testing procedures belong to the class $\mathcal{B}$, such as the weighted Bonferroni-Holm procedure (Holm 1979), fixed sequence tests (Maurer, Hothorn, and Lehmacher 1995; Westfall and Krishen 2001), fallback procedures (Wiens 2003; Wiens and Dmitrienko 2005), and Bonferroni-based gatekeeping procedures (Dmitrienko, Offen, and Westfall 2003; Hommel, Bretz, and Maurer 2007). In the sequel we will assume additionally that all tests for the intersection hypotheses are exhaustive.

A closed test is consonant if for any intersection hypothesis H_J that can be rejected, there is an index $j \in J$ such that for any $I \subseteq J$ with $j \in I$ the intersection hypothesis H_I can be rejected (Gabriel 1969). Hommel et al. (2007) showed that for weighted Bonferroni tests the consonance property is equivalent to the condition that for all $I \subseteq J \subseteq K$ the sets $\boldsymbol{\alpha}(I)$ obey the monotonicity condition $\alpha_i(I) \geq \alpha_i(J), i \in I$. Closed test procedures that use weighted Bonferroni tests for each intersection hypothesis and satisfy the above monotonicity condition lead to sequentially rejective test procedures of at most k steps, which can be performed as follows: Start testing the global intersection hypothesis $H_K, K = \{1, \ldots, k\}$. If it is rejected, there is an index $i \in K$ as described above such that H_i is rejected. At the next step, one continues testing the (reduced) global intersection $H_{K \setminus \{i\}}$ of the remaining, not yet rejected hypotheses, and so on, until the first nonrejection. In case the free combination condition does not hold, defining separate level-α tests for all intersection hypotheses (including those that are equal to others) and applying the closed testing principle results in a valid but potentially unnecessarily conservative procedure.

For a closed test procedure the succession property (property (a), Section 3) can be stated as follows: For a given index set $I \subseteq K = \{1, 2, \ldots, k\}$ let $I(p) = I \cap N$ be the index subset of the primary hypotheses $H_i, i \in I$, where as before $N = \{1, \ldots, n\}$ denotes the index set of all primary hypotheses. Further, let $I(s)$ denote the respective index subset of the secondary hypotheses. Hence, $I(p) \cup I(s) = I$ and $I(p) \cap I(s) = \emptyset$. In addition let $I(s') \subseteq I(s)$ denote the index set of secondary hypotheses that are descendants of primary hypotheses $H_i, i \in I(p)$. Then a closed test is successive if the level α test decisions for $H_I, I \subseteq K$, do not depend on hypotheses H_i, $i \in I(s')$. For the class $\mathcal{B}$ succession can be expressed as follows: The test procedure defined by $\boldsymbol{\alpha}(I)$ is successive if $\alpha_i(I) = 0$ for all $i \in I(s'), I \subseteq K$. These two statements are not difficult—but somewhat intricate—to prove. The idea of proof is the same as for proving via the closed testing principle that a strictly hierarchical (or serial gatekeeping) procedure controls the FWER at level α. Despite the restrictions imposed on the $2^k - 1$ level sets $\boldsymbol{\alpha}(I)$ by applying exhaustive, consonant, and successive Bonferroni-based tests, there is still a large degree of freedom for choosing the levels. If one is presented a procedure defined by $\boldsymbol{\alpha}(I)$, it is difficult to recognize the basic principle behind the choice of $\boldsymbol{\alpha}(I)$; on the other

hand if logical dependencies or orders of relative importance are given, it is not easy to construct a test procedure $\boldsymbol{\alpha}(I)$ that takes them into account.

One way to overcome that difficulty is using algorithmic, graph-based approaches to sequentially rejective Bonferroni procedures (Bretz et al. 2009b; Burman, Sonesson, and Guilbaud 2009). We will concentrate on the approach by Bretz et al. (2009b), since here the representation of the partition into primary and secondary hypotheses and successive tests turns out to be particularly intuitive and transparent. It will also allow a better understanding of the price to pay for the stringent control of the FWER for the combined families of primary and secondary hypotheses, compared to the more liberal procedures discussed in Section 3.

4.2 Sequentially Rejective Graphical Procedure for Successive Bonferroni-Based Closed Tests

Before we describe the specifics of the graphical approach applied to our situation of successive testing of primary and secondary hypotheses, we illustrate the method by means of a (simple) example with $k = 4$ hypotheses: H_1, H_2 are the primary and H_3, H_4 the secondary hypotheses, where H_3 is the only descendant of H_1, and H_4 is the only descendant of H_2. Since we do not want to reject a secondary hypothesis until its parent primary hypothesis is rejected we set the initial local levels $\alpha_3 = \alpha_4 = 0$. The four elementary hypotheses are represented by vertices of a graph with associated weights representing the local significance levels. In our case we have chosen $\alpha_1 = \alpha_2 = \alpha/2$, that is, the two primary hypotheses are considered as equally important. If one of them can be rejected at $\alpha/2$, this level is first shifted entirely to its descendant secondary hypothesis. If this can be rejected at level $\alpha/2$, its level is shifted and added to the level of the remaining, nonparent primary hypothesis which in turn now can be tested at full level α. After the potential rejection of this second primary hypothesis, its descendant secondary hypothesis then can be tested at level α. This particular gatekeeping procedure can be interpreted as a Bonferroni-Holm test applied to the hierarchical pairs of primary and secondary hypotheses (H_1, H_3) and (H_2, H_4).

This procedure can be fully described by the graph given in Figure 2(a) and an algorithm for sequentially updating it. The elementary hypotheses are represented in the graph by vertices with associated weights representing the local significance levels. The weight g_{ij} associated with a directed edge between any two vertices H_i and H_j indicates the fraction of the (local) significance level at the initial vertex (head) that is added to the significance level at the terminal vertex (tail) if the hypothesis H_i at the head is rejected. More formally, the initial local significance levels $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ are interpreted as weights of the k vertices with $\sum_{i=1}^{k} \alpha_i \leq \alpha$ and the $k \times k$ transition matrix $\mathbf{G} = (g_{ij})$ defines the weights of all directed edges. As a convention, edges with weight 0 are not drawn. In addition, regularity conditions hold for the initial matrix. Expressed in terms of the weights on the directed edges, the sum of the weights with tail on node H_i is restricted by 1 for $i \in K$ and there are no elementary loops (edges where head and tail coincide), that is, $g_{ii} = 0, i \in K$.

To illustrate the sequentially rejective graphical procedure, Figure 2 displays two graphs with example rejection sequences. Given the initial graphs in the left column, we have the significance levels $\boldsymbol{\alpha} = (\alpha/2, \alpha/2, 0, 0)$ for both graphs and the transition matrices

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{G} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

for the graphs (a) and (b), respectively.

Let p_i denote the unadjusted p-values for $H_i, i \in K$. Bretz et al. (2009b) have shown that the following algorithm determines a unique sequentially rejective (consonant) procedure with FWER controlled at level α.

0. Set $I = K$.
1. Select an i such that $p_i \leq \alpha_i$. If no such i exists stop, otherwise reject H_i. If $|I| = 1$ stop.
2. Update the graph:

$$I \to I \setminus \{i\}$$

$$\alpha_j \to \begin{cases} \alpha_j + \alpha_i g_{ij}, & j \in I, \\ 0, & \text{otherwise.} \end{cases}$$

If $|I| = 1$ go to Step 1.

$$g_{j\ell} \to \begin{cases} \frac{g_{j\ell} + g_{ji} g_{i\ell}}{1 - g_{ji} g_{ij}}, & j, \ell \in I, j \neq \ell, g_{ji} g_{ij} < 1 \\ 0, & \text{otherwise.} \end{cases}$$

3. If $|I| \geq 1$, go to Step 1; otherwise stop.

SAS/IML code that can be easily adapted to a particular initial graph is provided in Bretz et al. (2010).

As an example for the update steps, possible rejection sequences are given for the two initial graphs in Figure 2. They result from assuming the unadjusted p-values $(p_1, \ldots, p_4) = (0.01, 0.03, 0.02, 0.08)$ and $\alpha = 0.05$. In both cases H_1, H_2, and H_3 can be rejected, though in different sequences. Note that in case more than one hypothesis could be rejected at a particular step, the algorithm guarantees that the sequence of rejection has no influence on the test decision (Bretz et al. 2009b).

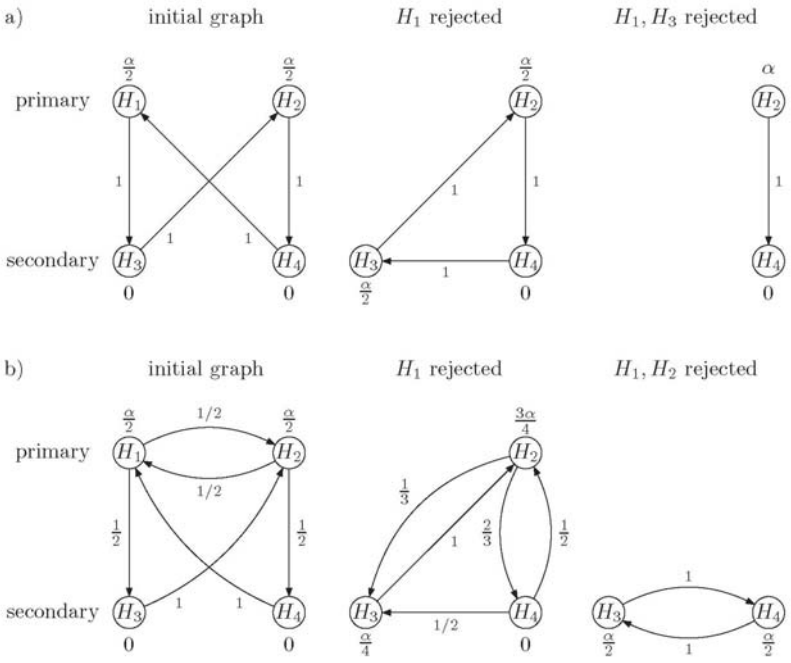


Figure 2. Graphs for two successive sequentially rejective procedure with example rejection sequences.

The succession property for families of primary hypotheses with descendant families of secondary hypotheses (as introduced in Section 4.1) can be generated easily by the following property of the initial graph (i.e., $I = K$): Let K_p be the index set of all primary hypotheses, K_s the index set of all secondary hypotheses, $S(i), i \in K_p$, the index set of descendants of H_i, and $P(j), j \in K_s$ the index set of (primary) parents of H_j. The consonant sequentially rejective procedure generated by a graph $\mathcal{G} = (\boldsymbol{\alpha}, \mathbf{G})$ is successive if

(i) $\alpha_j = 0, j \in K_s$ and

(ii) $g_{ij} = 0$ for any $j \in K_s$, with $i \in K_p$ and $j \notin S(i)$ or where $i \in K_s, i \neq j$, and $P(i) \neq P(j)$.

In other words: If a graph initially has weights 0 on all secondary hypotheses (vertices) and the only edges with positive weight leading into a secondary hypothesis (node) are those originating at its parent primary hypotheses and there are no edges leading from a secondary hypothesis to another secondary hypothesis that has not the same parents, then the rejection algorithm generates a successive procedure. A proof of this statement is given in Appendix A.3. An example of such a graph is given in Figure 3 with three primary hypotheses each with a descendant family of secondary hypotheses. The latter may be empty or have more than one member; different families also may contain identical hypotheses. Only one edge is drawn from a primary hypothesis to a descendant family in order not to overload the picture; for concrete cases the number of edges would be that of the members in the descendant family. Weights on the edges have been omitted; an edge drawn in Figure 3 means that it may have a positive weight, whereas edges not drawn must have weight 0.

It is instructive to verify the succession property in the two graphs of Figure 2. In order for a successive testing procedure to be also consistent according to definition (b) (Section 3) the retention of a secondary hypothesis cannot preclude the rejection of a primary hypothesis. This can be easily achieved in testing procedures defined by the graphical approach: The procedure is consistent if in the initial graph it is successive and for any primary hypothesis (vertex) H_i the initial weight α_i is either positive or there is a directed path with positive weights on the edges leading from another primary hypothesis (vertex) $H_{i'}$ with positive weight α'_i to H_i. The two example procedures given in Figure 2 hence are consistent.

For the case of two primary variables with one de-

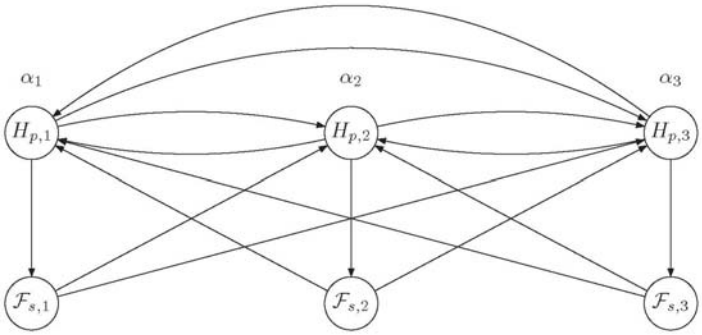


Figure 3. Concept of graph for partly hierarchical sequentially rejective procedure.

scendant secondary hypothesis each, another property can be verified in Figure 2: Irrespective of the rejection sequence at most two local significance levels can be positive at any rejection stage. This can be used to further improve the procedure by applying a generalization of the Simes (1986) test to the intersection hypotheses and also to construct powerful procedures in case all coprimary endpoints must be declared significant in order to achieve a success in a multiple comparison situation, as shown in the next section.

4.3 Improving Sequentially Rejective Procedures by Generalized Simes Tests

As mentioned under Case (3) in Section 2, Quan et al. (2009) considered the situation where the hierarchy of primary and secondary hypotheses is given by a clear order of importance in the comparison of two treatment arms vs. one control or of one test treatment vs. two controls (e.g., placebo and an active control). The n endpoints are considered as equally important. The procedure to test the $2n$ hypotheses they proposed, translated into our notation, is as follows:

(i) Reject all hypotheses $H_{p,i} \in \mathcal{F}_p$ and all $H_{s,i} \in \mathcal{F}_s$ if $p_{p,i} \leq \alpha$ and $p_{s,i} \leq \alpha, i = 1, \ldots, n$;

(ii) reject $H_{p,i}$ if $p_{p,i} \leq \alpha_i$ and reject $H_{s,i}$ if $p_{p,i} \leq \alpha_i$ and $p_{s,i} \leq \alpha_i, i = 1, \ldots, n$ where $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i = \alpha$.

Part (ii) applies the Bonferroni inequality to the partial hierarchical tests $H_{p,i} \rightarrow H_{s,i}$ where each pair is tested—starting with the primary hypothesis—at level α_i. The graphical representation of Part (ii) of this procedure is a graph with exactly one edge with weight 1, leading from each primary to the descendant secondary hypothesis, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ as initial weights on the (primary) vertices (see Figure 4 with $n = 3$). Hence, Part (ii) protects the FWER at level α. However, it can be uniformly improved by adding edges in the graph leading from the secondary hypotheses to primary hypotheses that are not their parents. Part (i) of the procedure exploits a generalized Simes inequality. The possibility to replace a weighted Bonferroni test in a gatekeeping procedure by a weighted Simes test has already been suggested by Dmitrienko, Offen, and Westfall (2003) and refined by Chen, Luo, and Capizzi (2005). To fully exploit the increased power of the Simes test, however, it would be necessary to check all $2^k - 1$ intersection hypotheses to determine which elementary hypotheses can be rejected. If only part of the Simes inequality is exploited one still can preserve the sequentially rejective nature of the consonant tests generated, for example, by the graphical approach, as we will show in the sequel.

As described by Quan et al. (2009) for a set of hypotheses $\mathcal{F}_I = \{H_i, i \in I\}, I \subseteq K$, the intersection hypothesis H_I can be rejected when for some $j \in I$

$$p_{(j)} \leq \sum_{i=1}^{j} \alpha_{(i)}, \tag{9}$$

where $p_{(j)}$ are the ordered p-values of the hypotheses in $\mathcal{F}_I$ and $\alpha_{(j)}$ are the corresponding local significance levels. Kling (2005) showed that this weighted Simes procedure controls the Type I error rate if the univariate test statistics are positive regression dependent which, for example, is the case if the test statistics follow a multivariate normal distribution with nonnegative correlations. Assuming that positive regression dependence of the univariate test statistics holds, the following generalization of the procedure of Quan et al. (2009) protects the FWER at level α. Let B be a closed and exhaustive weighted Bonferroni procedure of level α and $S(B)$ an extended procedure that rejects a hypothesis $H_i \in \mathcal{F}$ if B rejects

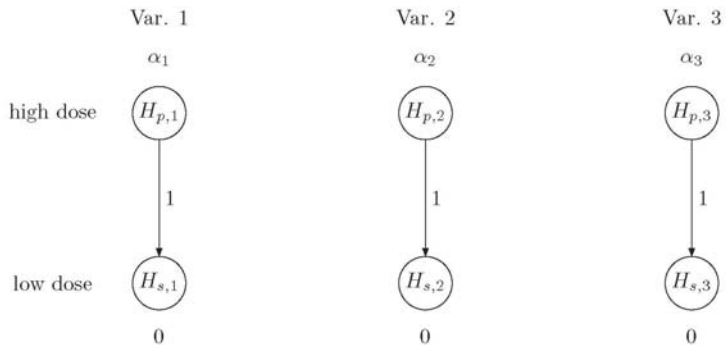


Figure 4. Graphical representation of Part (ii) in the procedure by Quan et al. (2009) for $n = 3$.

H_i, or if locally all hypotheses $H_i \in \mathcal{F}$ can be rejected at level α. Then $S(B)$ protects the FWER at level α as well and is uniformly more powerful than B. A proof of this statement is given in Appendix A.4.

This extension of weighted Bonferroni procedures can be applied in particular to sequentially rejective and successive procedures. The succession property remains preserved by the additional possibility to reject all hypotheses if they are locally significant at level α. We use the test procedure defined by the initial graph in Figure 2(a) to illustrate in which situations this extension allows to reject more hypotheses than the method proposed by Quan et al. (2009).

Assume that there are two primary endpoints to be tested in a multiple comparison of a high dose and a low dose against a control. It is sufficient that at least one of the two endpoints can be rejected and a (secondary) lower dose hypothesis on the two endpoints is rejected only if the respective high dose hypothesis could be rejected. Under the same assumptions for the initial levels as in the example from Section 4.2, where $\alpha_1 = \alpha_2 = \alpha/2 = 0.025$ for the two primary hypotheses with the assumed same outcome, that is, $(p_1, \ldots, p_4) = (0.01, 0.03, 0.02, 0.08)$, the method of Quan et al. (2009) allows to reject H_1 and H_3, whereas the graphical procedure allows to reject in addition H_2. In general the above procedure allows to reject any hypothesis that is rejected by the procedure of Quan et al. (2009).

For the above case with only four hypotheses the necessity that the test statistics are positive regression dependent can be relaxed. Brannath, Bretz, Maurer, and Sarkar (2009) have shown that for one-sided testing of two hypotheses a weighted trimmed Simes test can be used which protects the level α even if the correlations between the test statistics are negative. Let H_1 and H_2 be two hypotheses, $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = \alpha$ and p_1, p_2 respective univariate (one-sided) p-values. Then the trimmed version of the weighted Simes test allows to reject $H_1 \cap H_2$, irrespective of the correlation of the multivariate normal or t-distributed test statistics if

$$\begin{aligned} &\text{(a) } p_1 \leq \alpha_1 \text{ and } p_2 < 1 - \alpha_2 \\ &\text{or (b) } p_2 \leq \alpha_2 \text{ and } p_1 < 1 - \alpha_1 \\ &\text{or (c) } \max(p_1, p_2) \leq \alpha\,. \end{aligned} \tag{10}$$

There is a small price to be paid in order to use condition (10), namely that a hypothesis cannot be rejected if the observed effect in the other hypothesis is significant in the wrong direction. The motivation for trimming the Simes test is similar to the consistency requirement introduced by Alosh and Huque (2010). They considered the problem of testing a primary and a secondary endpoint and proposed a trimming strategy to avoid similar interpretation problems for the overall outcome of the trial when outcomes on different endpoints or doses are contradictory.

Despite the fact that the validity of the trimmed Simes test has only been derived for two hypotheses, it can be well applied to our situation with four hypotheses (two primary and two secondary hypotheses). The most general generating graph with the succession property for the situation where the initial α-levels for the two primary hypotheses are both positive is given in Figure 5; see also Bretz, Maurer, and Hommel (2010).

It is easy to check that for any rejection sequence at most two of the remaining hypotheses can have positive weight since if a hypothesis is rejected in the sequence, one hypothesis with positive weight is removed and its weight is shifted to at most one hypotheses with weight 0. Figure 2 illustrates this property. Hence the trimmed Simes procedure can be applied to all intersection hypotheses with more than one hypothesis in the intersection. To simplify the resulting closed test we suggest to

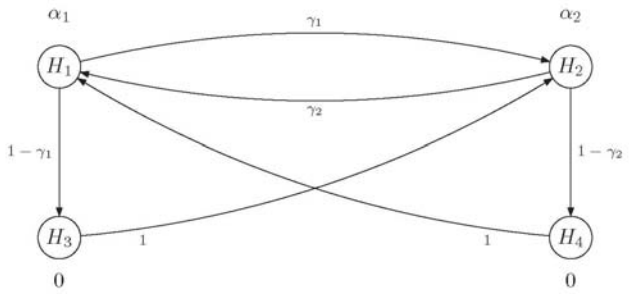


Figure 5. General graph for a successive sequentially rejective procedure for two primary hypotheses H_1, H_2 with positive initial weight and respective descendant secondary hypotheses H_3, H_4.

replace the condition for exclusion of contradictory results by the following—somewhat more conservative—condition $p_i < 1-\alpha, i = 1,\ldots,4$. Given two primary hypotheses with one distinct descendant hypothesis each, the resulting test procedure is then as follows:

(i) Retain all four hypotheses if $p_i \geq 1-\alpha$ for any $i \in \{1,\ldots,4\}$;

(ii) reject all four hypotheses if $p_i < \alpha$ for all $i \in \{1,\ldots,4\}$;

(iii) if neither (i) or (ii) applies, perform a closed successive weighted Bonferroni-test, for example, based on a successive graphical approach as given in Figure 5.

Condition (i) above is the price to pay for the additional possibility to reject all four hypotheses in Step (ii) irrespective of the correlations between the test statistics. Given that true opposing effects are unlikely and observed contradictory results would anyway prevent an overall claim of success, the small loss in power is more than compensated by the additional possibility (ii) to reject all four hypotheses. Power comparisons that qualitatively also apply to our case of four hypotheses can be found in Brannath et al. (2009).

4.4 Improving Multiple Comparison Procedures With Coprimary Endpoints by Generalized Simes Tests

In many indications regulatory guidance asks for statistically significant results in two endpoints for achieving a successful claim. Alzheimer's disease is a typical example where the respective CHMP guideline (2008) asks for significant differences in each of two primary variables which assess cognitive functions and activities of daily living, respectively. In these cases we have an intersection-union testing situation, that is, given two hypotheses H_i and H_j related to two different endpoints we have to reject the hypothesis $H_i \cup H_j$ by rejecting both individual hypotheses at level α. If the trial design has multiple treatments to be compared to a control, the testing situation is very similar to the one with primary and descendant secondary hypotheses. We arbitrarily consider one of the primary endpoints as primary and the other as secondary. In order to achieve statistical significance in both endpoints we do not lose anything when requesting that the designated primary hypothesis has to be rejected before the secondary one. If the two endpoints are equally important, however, the respective hypotheses should be sequentially tested at the same level. When using the graphical approach to define the Bonferroni part of the testing strategy, edges leading from a parent to a descendant hypothesis then should have full weight. As an example, consider the initial graph of Figure 2(a). The aim is to reject at least one of the pairs of hypotheses H_1, H_3 and H_2, H_4. We denote the two comparisons by A and B, respectively, and by $H_A = H_1 \cup H_3$ and $H_B = H_2 \cup H_4$ the two union-hypotheses to be tested and the initial α-levels by α_A and α_B instead of α_1 and α_2, respectively. Then we can construct a level α-test by applying the generalized Simes test described in Section 4.3 to the hypotheses $H_1,\ldots,H_4$ and reject H_A if both H_1 and H_3 are rejected and reject H_B if both H_2 and H_4 are rejected. This leads to the following test procedure which protects the multiple level $\alpha = \alpha_A + \alpha_B$:

(i) Retain H_A and H_B if $p_i \geq 1-\alpha$ for any $i \in \{1,\ldots,4\}$;

(ii) reject H_A and H_B if $p_i < \alpha$ for all $i \in \{1,\ldots,4\}$;

(iii) reject H_A if $\max(p_1, p_3) \leq \alpha_A$ or reject H_B if $\max(p_2, p_4) \leq \alpha_B$.

It is instructive to express this procedure in terms of the univariate test statistics $T_i, i = 1,\ldots,4$. Let $T_A =$

$\min(T_1, T_3)$ and $T_B = \min(T_2, T_4)$ be test statistics for the hypotheses H_A and H_B, respectively. Let u_α be $(1-\alpha)$-quantiles of the distribution of $T_i, i = 1, \ldots, 4$ under their respective null hypotheses. Then rejecting, for example, H_A if $T_A > u_\alpha$ is a valid level-α test. Given that the original test statistics T_i fulfill the conditions for a generalized Simes test mentioned in Section 4.3, then the above procedure implies that this is also the case for T_A and T_B. The test procedure is then as follows:

(i) Retain H_A and H_B if $\min(T_A, T_B) \leq -u_\alpha$;

(ii) reject H_A and H_B if $\min(T_A, T_B) > u_\alpha$;

(iii) reject H_A if $T_A > u_{\alpha_A}$, reject H_B if $T_B > u_{\alpha_B}$.

Note that if one is still interested in the elementary hypotheses H_i, the generalized Simes procedure on the graphical approach initially described allows a refinement of the test decisions. Given the example of test results in Section 4.2 on can reject H_1, H_2, and H_3, that is, H_A and a "part" of H_B. The above procedure would only allow to reject H_A. This refinement can be particularly helpful, if the two endpoints are not equally important.

5. Discussion

In confirmatory clinical trials we are increasingly often confronted with multiple testing situations involving several hypotheses of varying degree of importance. One reason is that in many indications effective treatments are available thus leading to trials with more than one comparator and/or more than one dose or treatment regimen in phase III. Often the multiplicity is further increased by the presence of more than one endpoint. In the last decade gatekeeping testing procedures have been developed that allow to take into account partially ordered hierarchy structures among the hypotheses. In this article we concentrate on a frequently occurring hierarchy structure where the hypotheses can be classified into primary (e.g., relevant for gaining approval) and secondary hypotheses (e.g., relevant only for additional label claims). Often the primary hypotheses are linked in a parent/descendant relation to one or more secondary hypotheses. We have listed five different situations where this type of partial ordering occurs naturally. We suggest two desirable properties for such testing procedures: succession and consistency.

Due to the different importance of primary and secondary hypotheses a control of the familywise Type I error rate at different levels α_p and α_s for the primary and secondary hypotheses families $\mathcal{F}_p$ and $\mathcal{F}_s$ might be of interest. In the first part of the article, we show that the unconditional FWER for $\mathcal{F}_s$ can be considerably larger than the conditional one when applying procedures with only conditional control of the FWER. In the second part of the article we investigate the consequences of controlling the FWER at a given significance level α, in particular when using consonant exhaustive Bonferroni-based closed test procedures. Conditions for such test procedures to be also consistent are given. We show that the graphical approach to sequentially rejective testing procedures by Bretz et al. (2009b) is an intuitively appealing and transparent way to define and compare such procedures. The succession and consistency properties can easily be translated into properties of the graphs that define the procedure.

It also can be shown that in practically relevant situations the Bonferroni procedures can further be improved by exploiting the generalized Simes inequality or the trimmed Simes inequality described by Brannath et al. (2009). This is discussed in the final part of the article where we show that such an extension in general is more powerful than the pure Bonferroni-based method. The succession properties of the underlying graphs lend itself to an application in testing situations where regulatory guidance asks for statistically significant results in two endpoints for achieving a successful claim. We show that the trimmed Simes procedure then can be used in multiple comparison situations where in at least one comparison for both endpoints statistically significant results have to be achieved.

Appendix

A.1 Proof of Equation (3) for Upper Boundary of FWER

Let $\mathcal{F}_p$ denote the family of n distinct primary hypotheses, of which each member has at least $h \geq 1$ distinct secondary hypotheses such that there are $m \geq n$ secondary hypotheses in family $\mathcal{F}_s$. Again $\mathcal{F}$ denotes the union of the two families with $n+m$ members which we denote by H_i, where $H_i = H_{p,i}, i \in N$ and $H_j = H_{s,j-n}, j = n+1, \ldots, n+m$. We further assume that a secondary hypothesis can have only one parent primary hypothesis. We need to show, that for procedure S_2 the inequality

$$P(W_2(\mathcal{F})) \leq \max(m\alpha_s, \alpha_p + (m-h)\alpha_s) \qquad \text{(A.1)}$$

holds.

Proof: We define a closed test procedure C with the properties that (i) C rejects a hypothesis $H \in \mathcal{F} = \mathcal{F}_p \cup \mathcal{F}_s$ whenever S_2 rejects and (ii) the probability $P(W_C(\mathcal{F}))$ to reject a true null hypotheses with procedure C for any given scenario of true and false hypotheses is controlled at level $\max(m\alpha_s, \alpha_p + (m-h)\alpha_s)$. We construct C by defining a test for the intersection hypotheses H_I of all subsets $\mathcal{F}_I \subseteq \mathcal{F}$, where I is the index set of the hypotheses $H_i \in \mathcal{F}_I$, that is, $H_I = \bigcap_{i \in I} H_i$. For

a given index set $I \subseteq \{1,2,\ldots,n+m\}$ let $I(p) = I \cap N$ be the index subset of the primary hypotheses in $\mathcal{F}_I$ and $I(s)$ be the respective index subset of the secondary hypotheses; hence $I(p) \cup I(s) = I$ and $I(p) \cap I(s) = \emptyset$. With $H_{I(p)} = \bigcap_{i \in I(p)} H_i$, let $S_{2,p}$ be the closed test at level α_p defined by S_2 on the primary hypotheses and $S_{2,I(p)}$ the respective test on the intersection hypotheses $H_{I(p)}$. With $\mathcal{F}_{I(s^*)}$ denoting the set of secondary hypotheses in $\mathcal{F}_{I(s)}$ which are not descendants of primary hypotheses in $\mathcal{F}_{I(p)}$, C is defined as follows: Reject H_I if $S_{2,p}$ rejects $H_{I(p)}$ or if any of the hypotheses in $\mathcal{F}_{I(s^*)}$ can be rejected at level α_s. If $I(p) = \emptyset$, then the probability to erroneously reject H_I is bounded by $m\alpha_s$. If there is at least one primary hypothesis in $\mathcal{F}_I$ there are at most $m-h$ secondary hypotheses in $\mathcal{F}_{I(s^*)}$ and hence the Type I error rate is bounded by $(\alpha_p + (m-h)\alpha_s)$, due to the Bonferroni inequality. Hence all of the intersection tests are bounded by $\max(m\alpha_s, \alpha_p + (m-h)\alpha_s)$ and therefore this is also the case for the FWER of procedure C.

It remains to be shown that C has also property (i) from the beginning of the proof. Assume that hypothesis H_i is rejected by procedure S_2. If H_i is a primary hypothesis, then all intersection hypotheses of primary hypotheses $H_{I(p)}$ where $H_{I(p)} \subseteq H_i$ are rejected by S_2 which implies the rejection of $H_{I(p)} \cap H_{I(s)} \subseteq H_i$ by C. Hence by the closed testing principle also procedure C rejects H_i. If H_i is a secondary hypothesis that is rejected by S_2 locally at level α_s then also one of the parent primary hypotheses of H_j, say $H_{p,j}$, must have been rejected by S_2. Hence all intersection hypotheses H_I where $H_I \subseteq H_i$ can be rejected by C, either because $H_{p,j}$ is also a member of H_I or, if this is not the case, because the rejection of the elementary secondary hypothesis H_i by S_2 allows the rejection of H_I by C.

A.2 FWER for Strategy S_1 for Two Primary and Secondary Hypotheses with Bonferroni-Holm as Closed Test

Let $R_{p,1} = \{T_{p,1} \geq u_{\alpha/2}\} \cap \{T_{p,2} \geq u_{\alpha/2}\}$, $R_{p,2} = \{u_{\alpha/2} > T_{p,1} \geq u_\alpha\} \cap \{T_{p,2} \geq u_{\alpha/2}\}$ and $R_{p,3} = \{T_{p,1} \geq u_{\alpha/2}\} \cap \{u_{\alpha/2} > T_{p,2} \geq u_\alpha\}$ be the three disjoint events that lead to the rejection of both primary hypotheses. Similarly $R_{s,1} = \{T_{s,1} \geq u_{\alpha/2}\}$ and $R_{s,2} = \{T_{s,1} < u_{\alpha/2}, T_{s,2} \geq u_{\alpha/2}\}$ are the two disjoint events that can lead to rejection of at least one secondary hypothesis. With strategy S_1 and using Bonferroni-Holm as the multiplicity-adjustment method, the probability $P(R_s) = P(R_{s,1} \cup R_{s,2})$ to reject at least one secondary hypothesis is then

$$\begin{aligned} P(R_s) = \quad & P\left(T_{p,1} \geq u_{\alpha/2}, T_{p,2} < u_\alpha, T_{s,1} \geq u_\alpha\right) \\ & + P\left(T_{p,1} < u_\alpha, T_{p,2} \geq u_{\alpha/2}, T_{s,2} \geq u_\alpha\right) \\ & + P\left(\{R_{p,1} \cup R_{p,2} \cup R_{p,3}\} \bigcap \{R_{s,1} \cup R_{s,2}\}\right). \end{aligned} \tag{A.2}$$

The first two expressions are the probabilities that exactly one primary and its descendant secondary hypotheses are rejected. The third expression can be written as the sum of probabilities of six mutually disjoint events that are intersections of events of the type $T \geq u$ or $T < u$ for which numerical evaluation is available.

If the test statistics are jointly normally distributed with $E(T_{i,1}) = E(T_{i,2}); i = p,s$, $\rho = \text{corr}(T_{p,j}, T_{s,j}), j = 1,2$, $\tau = \text{corr}(T_{i,1}, T_{i,2}); i = p,s$ and conditional independence of $T_{p,1}$ and $T_{s,2}$ given either $T_{p,2}$ or $T_{s,1}$, then this formula simplifies to

$$\begin{aligned} P(R_s) \; = \; & 2P\left(T_{p,1} \geq u_{\alpha/2}, T_{p,2} < u_\alpha, T_{s,1} \geq u_\alpha\right) \\ & + 2P\left(u_{\alpha/2} > T_{p,1} \geq u_\alpha, T_{p,2} \geq u_{\alpha/2}, T_{s,1} \geq u_{\alpha/2}\right) \\ & + 2P\left(u_{\alpha/2} > T_{p,1} \geq u_\alpha, T_{p,2} \geq u_{\alpha/2}, T_{s,1} \right. \\ & \left. < u_{\alpha/2}, T_{s,2} \geq u_{\alpha/2}\right) \\ & + P\left(T_{p,1} \geq u_{\alpha/2}, T_{p,2} \geq u_{\alpha/2}, T_{s,1} \geq u_{\alpha/2}\right) \\ & + P\left(T_{p,1} \geq u_{\alpha/2}, T_{p,2} \geq u_{\alpha/2}, T_{s,1} \right. \\ & \left. < u_{\alpha/2}, T_{s,2} \geq u_{\alpha/2}\right). \end{aligned} \tag{A.3}$$

For $\rho = 1$, $\tau = 0$, $E(T_{p,1}) = E(T_{p,2}) = u_{\alpha/2} - u_\alpha$, we have $P\left(T_{p,1} \geq u_{\alpha/2}\right) = P\left(T_{p,1} \geq u_{\alpha/2}, T_{s,1} \geq u_\alpha\right) = \alpha$. Hence under these assumptions, the probability to reject exactly one primary hypothesis at level $\alpha/2$ and its descendant secondary hypothesis at level α is $2\alpha(1-\alpha) = 2\alpha - 2\alpha^2$. This rough approximation does however not take into account the event that both primary hypotheses are rejected. More exactly, (A.3), then becomes

$$P(R_s) = \alpha + \alpha P\left(T_{s,1} < 2u_\alpha - u_{\alpha/2}\right) - \frac{\alpha^2}{4}. \tag{A.4}$$

To show the change from the above approximation, for $\alpha = 0.05$, this is $2\alpha - 2.09 \cdot \alpha^2$ and for $\alpha = 0.025$, it is $2\alpha - 2.11 \cdot \alpha^2$.

For $\rho = 1$, $\tau = -1$, $\mu_p = E(T_{p,1}) = E(T_{p,2}) \leq u_{\alpha/2} + u_\alpha$ and $E(T_{s,1}) = E(T_{s,2}) = 0$, we note that due to $\tau = -1$, it follows $T_{p,2} = 2\mu_p - T_{p,1}$, all terms in (A.3) except the first one being 0. Hence, (A.3) further simplifies to

$$P(R_s) = 2P(T_{s,1} > \max(u_{\alpha/2} - \mu_p, \mu_p - u_\alpha, u_\alpha)).$$

In the range $u_{\alpha/2} - u_\alpha \leq \mu_p \leq 2u_\alpha$, and for $\alpha < 0.303$ the maximum is u_α, such that (A.3) simplifies to $P(R_s) = 2P(T_{s,1} \geq u_\alpha) = 2\alpha$.

A.3 Graphical Procedures With Succession Property

(*Sketch of the proof of the following statement from Section 4.2*) The consonant sequentially rejective procedure generated by a graph $\mathcal{G} = (\boldsymbol{\alpha}, \mathbf{G})$ is successive if for any secondary hypothesis H_j (i) $\alpha_j = 0, j \in K_s$ and (ii) $g_{ij} = 0$ for any $j \in K_s$, with $i \in K_p$ and $j \notin S(i)$ or where $i \in K_s, i \neq j$, and $P(i) \neq P(j)$.

Proof: We have to show that an initial graph with this property generates (via the algorithm given in Section 4.2) a successive test procedure, that is, secondary hypotheses are rejected only if a parent primary hypothesis has been rejected. To this end we show that the test procedure is successive if at any stage in the rejection sequence the updated graph (without already rejected hypotheses) has the following property $\mathcal{S}$: With $I \subseteq K$ denoting the set of indices of not yet rejected hypotheses, I_p and I_s the index sets of not yet rejected primary and secondary hypotheses, respectively and $S(i), i \in I_p$, the index set of descendants of H_i, and $P(j), j \in I_s$ the parents of H_j. Additionally let $I_{s'} = \cup(S(i); i \in I_p) \subseteq I_s$ denote the index set of secondary hypotheses with (not yet rejected) parent primary hypotheses, and $(\boldsymbol{\alpha}(I), \mathbf{G}(I))$ the resulting graph on the remaining set of hypotheses, then

(i) $\alpha_j = 0, j \in I_{s'}$ and

(ii) $g_{ij} = 0$ for any $j \in I_{s'}$, with $i \in I_p$ and $j \notin S(i)$ or where $i \in I_s, i \neq j$, and $P(i) \neq P(j)$.

If at a certain step in the rejection sequence the graph has property $\mathcal{S}$, only secondary hypotheses with already rejected primary parent can be rejected due to the first condition (i.e., the procedure is successive at that stage). Assume a primary hypothesis is now rejected and I is reduced (by one index) to I^-. Then the first property $\alpha_j(I^-) = 0, j \in I^-_{s'}$ still holds because by condition (ii) the level of the rejected hypothesis cannot be shifted to a secondary hypothesis H_j with a (nonrejected) parent in I^-_p, that is, where $j \in I^-_{s'}$. If a secondary hypothesis is rejected then by the second part of condition (ii) the level of this hypothesis cannot be shifted to a secondary hypothesis H_j with not yet rejected primary parent hypotheses, since in this case $I^-(s') = I(s')$. Therefore, again $\alpha_j(I^-) = 0, j \in I^-_{s'}$ on the reduced set I^-. Similar reasoning also shows—by applying the update algorithm to the transition matrix G—that property $\mathcal{S}$ is invariant under sequential rejection. Clearly the initial graph has property $\mathcal{S}$ because $K_s = K_{s'}$ and hence by complete induction the statement is proven.

A.4 Generalization of the Procedure by Quan et al. (2009)

(*Short proof of the following statement from Section 4.3*) Let B be a closed and exhaustive weighted Bonferroni procedure of level α and $S(B)$ an extended procedure that rejects a hypothesis $H_i \in \mathcal{F}$ if B rejects H_i, or if locally all hypotheses $H_i \in \mathcal{F}$ can be rejected at level α. Then $S(B)$ protects the FWER at level α as well and is uniformly more powerful than B.

Proof: Let $I \subseteq K$ be a subset of size s of the index set of all hypotheses in $\mathcal{F}$ and H_I the respective intersection hypothesis. Let $\boldsymbol{\alpha}(I)$ be the set of local levels $\alpha_i(I), i \in I$. Then H_i can be rejected by procedure B if $p_k \leq \alpha_k(I)$ for at least one $k \in I$. If instead of the weighted Bonferroni test the weighted Simes test is used, then H_i can be rejected as well. To see this, assume that $p_k = p_{(j)}(I)$ for some $j \in \{1, \ldots, s\}$ and hence $p_{(j)}(I) \leq \alpha_k(I) = \alpha_{(j)}(I) \leq \sum_{i=1}^{j} \alpha_{(i)}(I)$. Therefore, by definition, also the weighted Simes test rejects H_I. With $p_{(s)} = \max(p_i), i \in I$, H_I can also be rejected if inequality (9) holds for $j = s$, that is, if $p_{(s)}(I) \leq \sum_{i=1}^{s} \alpha_{(i)}(I) = \sum_{i \in I} \alpha_i(I) = \alpha$, or—equivalently—if $p_i \leq \alpha$ for all $i \in I$. The above statement then follows by applying the closed test principle.

Acknowledgments

We thank the guest editors for inviting us to contribute to this Festschrift for Prof. Gary Koch. We also thank the editor and two anonymous referees for their thorough reviews that helped improving the quality of the article and are grateful to Nikos Sfikas for providing us the basis for the example cases from Section 2.

[Received February 2010. Revised June 2010.]

References

Alosh, M., and Huque, M. F. (2010), "A Consistency-Adjusted Alpha Adaptive Strategy for Sequential Testing," *Statistics in Medicine*, 29, 1559-1571. 347

Bauer, P (1991), "Multiple Testing in Clinical Trials," *Statistics in Medicine*, 10, 871–890. 339

Brannath, W., Bretz, F., Maurer, W., and Sarkar, S. (2009), "Trimmed Weighted Simes' Test for Two One-Sided Hypotheses With Arbitrarily Correlated Test Statistics," *Biometrical Journal*, 51, 885–898. 347, 348, 349

Bretz, F., Maurer, W., and Gallo, P. (2009a), Discussion of "Some controversial multiple testing problems in regulatory applications," by H. M. J. Hung and S.-J. Wang, *Journal of Biopharmaceutical Statistics*, 19, 25–34. 336, 337

Bretz, F., Maurer, W., Brannath, W., and Posch, M. (2009b), "A Graphical Approach to Sequentially Rejective Multiple Test Procedures," *Statistics in Medicine*, 28, 586–604. 337, 344, 349

Bretz, F., Maurer, W. and Hommel, G. (2010), "Test and Power Considerations for Multiple Endpoint Analyses using Sequentially Rejective Graphical Procedures," *Statistics in Medicine*, (in press). 337, 344, 347

Burman, C.-F., Sonesson C., and Guilbaud O.(2009), "A Recycling Framework for the Construction of Bonferroni-based Multiple Tests," *Statistics in Medicine*, 28, 739–761. 344

Chen, X., Luo, X., and Capizzi, T. (2005), " The Application of Enhanced Parallel Gatekeeping Strategies," *Statistics in Medicine*, 24, 1385-1397. 346

Committee for Medicinal Products for Human Use (2008), "Guideline on Medicinal Products for the Treatment of Alzheimer's Disease and other Dementias," *Doc. Ref. CPMP/EWP/553/95* Rev. 1. 348

Dmitrienko A., Offen W.W., and Westfall P.H. (2003), "Gatekeeping Strategies for Clinical Trials that do not Require all Primary Effects to be Significant," *Statistics in Medicine*, 22, 2387–2400. 337, 343, 346

Gabriel, K. R. (1969), "Simultaneous Test Procedures—Some Theory of Multiple Comparisons," *The Annals of Mathematical Statistics*, 40, 224–520. 343

Genz, A., and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Heidelberg: Springer Verlag. 342

Glimm, E., Maurer, W., and Bretz, F. (2010), "Hierarchical Testing of Multiple Endpoints in Group Sequential Trials," *Statistics in Medicine*, 29, 219–228. 337, 339

Holm, S. (1979), "A Simple Sequentally Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70. 340, 343

Hommel, G., Bretz, F., and Maurer, W. (2007), "Powerful Short-Cuts for Multiple Testing Procedures with Special Reference to Gatekeeping Strategies," *Statistics in Medicine*, 26, 4063–4073. 337, 343

Hung, H.M.J., and Wang, S.J. (2009), "Some Controversial Multiple Testing Problems in Regulatory Applications," *Journal of Biopharmaceutical Statistics*, 19, 1–11. 336, 339

Hung, H.M.J., and Wang, S.J. (2010), "Challenges to Multiple Testing in Clinical Trials," *Biometrical Journal*, (in press). 336, 338

Kling, Y. (2005), "Issues of Multiple Hypothesis Testing in Statistical Process Control," Thesis, The Neiman Library of Exact Sciences & Engineering, Tel-Aviv University. 346

Marcus, R., Peritz, E., and Gabriel, K.R., (1976), "On Closed Testing Procedures With Special Reference to Ordered Analysis of Variance," *Biometrika*, 63, 655–660. 343

Maurer, W., Hothorn, L., and Lehmacher, W. (1995), "Multiple Comparisons in Drug Clinical Trials and Preclinical Assays, A-priori Ordered Hypotheses," in *Biometrie in der Chemisch-Pharmazeutischen Industrie*, ed. J. Vollmar, Stuttgart: Fischer Verlag: pp. 3–18. 343

Quan, H., Capizzi, T., and Zhang, J. (2009), "Multiplicity Adjustment for Clinical Trials With Two Doses of an Active Treatment," *Statistics in Biopharmaceutical Research*, 1, 258–267. 337, 338, 346, 347, 351

Simes, R.J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754. 346

Tamhane, A.C., Mehta, C.R., and Liu, L. (2010), "Testing a Primary Endpoint and a Secondary Endpoint in a Group Sequential Design," *Biometrics*, (in press). 337, 339

Westfall, P.H., and Krishen, A., (2001), "Optimally Weighted, Fixed Sequence, and Gatekeeping Multiple Testing Procedures," *Journal of Statistical Planning and Inference*, 99, 25–40. 343

Wiens, B.L. (2003), "A Fixed Sequence Bonferroni Procedure for Testing Multiple Endpoints," *Pharmaceutical Statistics*, 2, 211–215. 343

Wiens, B.L., and Dmitrienko, A. (2005), "The Fallback Procedure for Evaluating a Single Family of Hypotheses," *Journal of Biopharmaceutical Statistics*, 15, 929–942. 343

About the Authors

Willi Maurer is Senior Biometrical Fellow, Ekkehard Glimm is Senior Expert Statistical Methodologist, and Frank Bretz is Global Statistical Methodology Head, Novartis Pharma AG, WSJ-027.1.028, Forum 1, Novartis Campus, CH-4056 Basel, Switzerland (E-mail for correspondence: *willi.maurer@novartis.com*).

Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests

Frank Bretz[*,**,1], **Martin Posch**[**,2], **Ekkehard Glimm**[1], **Florian Klinglmueller**[2], **Willi Maurer**[1], and **Kornelius Rohmeyer**[3]

[1] Statistical Methodology, Novartis Pharma AG, Basel, Switzerland
[2] Section of Medical Statistics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[3] Institute of Biostatistics, University of Hannover, Hannover, Germany

Received 26 November 2010, revised 15 May 2011, accepted 4 June 2011

The confirmatory analysis of pre-specified multiple hypotheses has become common in pivotal clinical trials. In the recent past multiple test procedures have been developed that reflect the relative importance of different study objectives, such as fixed sequence, fallback, and gatekeeping procedures. In addition, graphical approaches have been proposed that facilitate the visualization and communication of Bonferroni-based closed test procedures for common multiple test problems, such as comparing several treatments with a control, assessing the benefit of a new drug for more than one endpoint, combined non-inferiority and superiority testing, or testing a treatment at different dose levels in an overall and a subpopulation. In this paper, we focus on extended graphical approaches by dissociating the underlying weighting strategy from the employed test procedure. This allows one to first derive suitable weighting strategies that reflect the given study objectives and subsequently apply appropriate test procedures, such as weighted Bonferroni tests, weighted parametric tests accounting for the correlation between the test statistics, or weighted Simes tests. We illustrate the extended graphical approaches with several examples. In addition, we describe briefly the `gMCP` package in `R`, which implements some of the methods described in this paper.

Keywords: Dunnett test; Gatekeeping procedure; Min-*p* test; Non-inferiority; Truncated Holm.

1 Introduction

Multiple test procedures are often used in the analysis of clinical trials addressing multiple objectives, such as comparing several treatments with a control and assessing the benefit of a new drug for more than one endpoint. Several multiple test procedures have been developed in the recent past that allow one to map the relative importance of the different study objectives as well as their relation onto an appropriately tailored multiple test procedure.

A common strategy to reduce the degree of multiplicity is to group the hypotheses into primary and secondary objectives (O'Neill, 1997). Test procedures accounting for the inherent logical relationships include fixed sequence tests (Maurer et al., 1995; Westfall and Krishen, 2001), gatekeeping procedures (Bauer et al., 1998; Westfall and Krishen, 2001; Dmitrienko et al., 2003) and fallback procedures (Wiens, 2003; Huque and Alosh, 2008). Li and Mehrotra (2008) introduced a more general approach for adapting the significance level to test secondary hypotheses based on the

*Corresponding author: e-mail: frank.bretz@novartis.com, Phone: +41-61-324-4064, Fax: +41-61-324-3039
**These authors contributed equally to this work.

finding for the primary hypotheses. Alosh and Huque (2009) introduced the notion of consistency when testing for an effect in the overall population and in a specific subgroup. The authors extended this consistency concept to other situations (Alosh and Huque, 2010), including how to address multiplicity issues of a composite endpoint and its components in clinical trials (Huque et al., 2011). Hung and Wang (2009, 2010) considered some controversial multiple test problems, with emphasis on regulatory applications, and pointed out illogical problems that may arise with recently developed multiple test procedures.

In this paper, we focus on graphical approaches which have been introduced independently by Bretz et al. (2009) and Burman et al. (2009). The key idea is to express the resulting multiple test procedures by directed, weighted graphs, where each node corresponds to an elementary hypothesis, together with a simple algorithm to generate such graphs while sequentially testing the individual hypotheses. Using graphical approaches, one can explore different test strategies together with the clinical team and thus tailor the multiple test procedure to the given study objectives. So far, the description of these graphical approaches has focused on Bonferroni-based test procedures. In this paper, we investigate extensions of the original ideas. In particular, we discuss in Section 2 how a separation between the weighting strategy and the test procedure facilitates the application of a graphical approach beyond Bonferroni-based test procedures. In Section 3, we illustrate these ideas with different test procedures. We start with a brief review of Bonferroni-based test procedures and subsequently describe parametric graphical approaches that account for the correlation between the test statistics as well as graphical approaches using the Simes test. In Section 4, we describe the `gMCP` package in `R` which implements some of the methods discussed in this paper and illustrate it with a clinical trial example using a truncated Holm procedure. Concluding remarks are given in Section 5.

2 Graphical weighting strategies

Consider the problem of testing m elementary hypotheses $H_1,\ldots,H_m$, some of which could be more important than others, e.g. primary and secondary objectives. Let $I = \{1,\ldots,m\}$ denote the associated index set. The closure principle introduced by Marcus et al. (1976) is commonly used to construct powerful multiple test procedures. Accordingly, we consider all non-empty intersection hypotheses $H_J = \cap_{j\in J} H_j, J \subseteq I$. We further pre-specify an α-level test for each H_J. The resulting closed test procedure rejects $H_i, i \in I$, if all intersection hypotheses H_J with $i \in J \subseteq I$ are rejected by their corresponding α-level tests. By construction, closed test procedures control the familywise error rate (FWER) in the strong sense at level $\alpha \in (0,1)$. That is, the probability to reject at least one true null hypothesis is bounded by α under any configuration of true and false null hypotheses (Hochberg and Tamhane, 1987). In fact, closed test procedures have certain optimality properties whenever the FWER has to be controlled (Bauer, 1991). In what follows, we assume that the hypotheses $H_1,\ldots,H_m$ satisfy the free combination condition (Holm, 1979). If this condition is not satisfied, the methods in this paper still control the FWER at level α, although they can possibly be improved because of the reduced closure tree (Brannath and Bretz, 2010).

One important class of closed test procedures is obtained by applying weighted Bonferroni tests to each intersection hypothesis H_J. For each $J \subseteq I$ assume a collection of weights $w_j(J)$ such that $0 \le w_j(J) \le 1$ and $\sum_{j\in J} w_j(J) \le 1$. With the weighted Bonferroni test we reject H_J if $p_j \le \alpha_j(J) = w_j(J)\alpha$ for at least one $j \in J$, where p_j denotes the unadjusted p-value for H_j. Hommel et al. (2007) introduced a useful subclass of sequentially rejective Bonferroni-based closed test procedures. They showed that the monotonicity condition

$$w_j(J) \le w_j(J') \quad \text{for all } J' \subseteq J \subseteq I \quad \text{and} \quad j \in J' \tag{1}$$

ensures consonance, i.e. if an intersection hypothesis H_J is rejected, there is an index $j \in J$, such that the elementary hypothesis H_j can be rejected as well. This substantially simplifies the implementation and interpretation of related closed test procedures, as the closure tree of 2^m-1 intersection hypotheses is tested in only m steps. Many common multiple test procedures satisfy (1), see Hommel et al. (2007) for examples.

Bretz et al. (2009) and Burman et al. (2009) independently derived graphical representations and associated rejection algorithms for important subclasses of the Hommel et al. (2007) procedures. The graphical representations and rejection algorithms in these two articles are different, though underlying ideas are closely related; see Guilbaud and Karlsson (2011) for some comparative examples. Using the graphical approach of Bretz et al. (2009), the hypotheses $H_1,\ldots,H_m$ are represented by vertices with associated weights denoting the local significance levels $\alpha_1,\ldots,\alpha_m$. In addition, any two vertices H_i and H_j are connected through directed edges, where the associated weight g_{ij} indicates the fraction of the (local) significance level α_i that is propagated to H_j once H_i (the hypothesis at the tail of the edge) has been rejected. A weight $g_{ij} = 0$ indicates that no propagation of the significance level is foreseen and the edge is dropped for convenience. Figure 1 shows an example.

While the original graphical approaches were introduced based on weighted Bonferroni tests, we propose here to dissociate the underlying *weighting strategy* from the employed *test procedure*. The benefit of such an approach is the enhanced transparency by (i) first deriving suitable weighting strategies that reflect the given study objectives (and which can be communicated to the clinical team) and (ii) subsequently applying appropriate test procedures that do not necessarily have to be based on Bonferroni's inequality.

Graphical weighting strategies are conceptually similar to the graphs proposed by Bretz et al. (2009). They essentially summarize the complete set of $\sum_{i=1}^{m} i \binom{m}{i} = m2^{m-1}$ weights determining the full closure tree. A weighted multiple test can then be applied to each intersection hypothesis H_J, such as a weighted Bonferroni test, a weighted min-p test accounting for the correlation between the test statistics, or a weighted Simes test; see Section 3 for details. Weighting strategies are formally defined through the weights $w_i(I)$, $i \in I$, for the global null hypothesis H_I and the transition matrix $\mathbf{G} = (g_{ij})$, where $0 \le g_{ij} \le 1$, $g_{ii} = 0$, and $\sum_{j=1}^{m} g_{ij} \le 1$ for all $i, j \in I$. We additionally need to determine how the graph is updated once a vertex is removed. This can be achieved by tailoring Algorithm 1 in Bretz et al. (2009) to the graphical weighting strategies as follows. For a given index set $J \subseteq I$, let $J^c = I \setminus J$ denote the set of indices that are not contained in J. Then the following algorithm determines the weights $w_j(J)$, $j \in J$. This algorithm has to be repeated for each $J \subseteq I$ to generate the $m2^{m-1}$ weights for the full closure.

Algorithm 1 (Weighting Strategy)

(i) Select $j \in J^c$ and remove H_j

(ii) Update the graph:

$$I \to I\setminus\{j\}, J^c \to J^c\setminus\{j\}$$

$$w_\ell(I) \to \begin{cases} w_\ell(I) + w_j(I) g_{j\ell}, & \ell \in I \\ 0, & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ell \neq k, g_{\ell j} g_{j\ell} < 1 \\ 0, & \text{otherwise} \end{cases}$$

(iii) If $|J^c| \ge 1$, go to step (i); otherwise $w_\ell(J) = w_\ell(I), \ell \in J$, and stop.

As shown by Bretz et al. (2009), the weights $w_j(J)$, $j \in J$ are unique. In particular, they do not depend on the sequence in which hypotheses H_j, $j \in J^c$, are removed in step (i) of Algorithm 1. Note

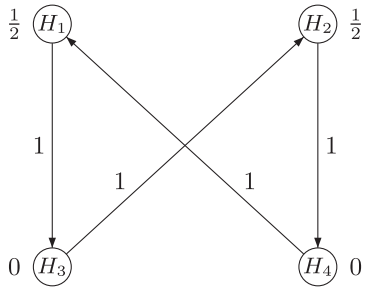


Figure 1 Weighting strategy for two hierarchically ordered endpoints and two dose levels.

that Algorithm 1 requires specifying the weights $w_j(I)$ for the global intersection hypothesis H_I and the elements of the transition matrix $\mathbf{G}$. This leads to the specification of $m+m(m-1) = m^2$ parameters if $\sum_{j\in I} w_j(I) \leq 1$ and $\sum_{j=1}^{m} g_{ij} \leq 1$ or $m-1+m(m-2) = m^2-m-1$ parameters if $\sum_{j\in I} w_j(I) = 1$ and $\sum_{j=1}^{m} g_{ij} = 1$, for all $i, j \in I$.

Example 1

As an example, assume a primary family of two hypotheses $\mathcal{F}_1 = \{H_1, H_2\}$ and a secondary family of two hypotheses $\mathcal{F}_2 = \{H_3, H_4\}$. The hypotheses H_1 and H_2 could denote, for example, the comparison of low and high dose with a control, for either a primary endpoint, a non-inferiority claim, or an overall population. Accordingly, the hypotheses H_3 and H_4 would then denote the comparison of the same two doses with a control, for either a secondary endpoint, a superiority claim, or a pre-specified subgroup. Figure 1 visualizes one possible weighting strategy. It is motivated by a strict hierarchy within dose: the secondary endpoint will only be assessed if efficacy was shown previously for the primary endpoint (so-called successiveness property; see Maurer et al., 2011). If for one of the doses efficacy can be shown for both the primary and the secondary endpoint, the associated weight is passed on to the other dose. Therefore we have $I = \{1, 2, 3, 4\}$, $w_1(I) = w_2(I) = 0.5$ for the primary hypotheses and $w_3(I) = w_4(I) = 0$ for the secondary hypotheses, which implies that no secondary hypothesis can be rejected until a primary hypothesis is rejected and propagates its weight. The associated transition matrix is

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The graph in Figure 1 together with Algorithm 1 from above fully specify the 32 weights of the full closure tree, as summarized in Table 1. This table parallels the weight tables introduced by Dmitrienko et al. (2003). Note that the weights $w_j(J)$, $j \in J^c$, are formally not defined and expressed by "–" in Table 1. Figure 2 displays the updated graphs resulting from Figure 1 after removing H_1, H_2, H_3, or H_4. The four updated graphs in Figure 2 correspond to the four rows in Table 1 containing the weights for the three-way intersection hypotheses. Removing any two hypotheses results in six possible two-way intersection hypotheses and the two vertexes are connected by two directed edges, each with weight 1 (graphical display omitted here). Note that Figure 2 displays the principle of recalculating the weights by updating the graphs. It is possible and also necessary to remove hypotheses with weight 0 (in this example H_3 and H_4 with $w_3(I) = w_4(I) = 0$) in order to compute the respective weights for the larger intersection hypotheses.

Note that Figure 1 displays only one possible weighting strategy. Many other weighting strategies are possible and perhaps more reasonable, depending on the given context. We refer to Bretz et al. (2011) for a generic discussion about testing two families $\mathcal{F}_1$ and $\mathcal{F}_2$ with two hypotheses each.

Table 1 Weights for the intersection hypotheses derived from Figure 1.

Intersection hypothesis	Weights			
	H_1	H_2	H_3	H_4
$H_1 \cap H_2 \cap H_3 \cap H_4$	0.5	0.5	0	0
$H_1 \cap H_2 \cap H_3$	0.5	0.5	0	–
$H_1 \cap H_2 \cap H_4$	0.5	0.5	–	0
$H_1 \cap H_2$	0.5	0.5	–	–
$H_1 \cap H_3 \cap H_4$	0.5	–	0	0.5
$H_1 \cap H_3$	1	–	0	–
$H_1 \cap H_4$	0.5	–	–	0.5
H_1	1	–	–	–
$H_2 \cap H_3 \cap H_4$	–	0.5	0.5	0
$H_2 \cap H_3$	–	0.5	0.5	–
$H_2 \cap H_4$	–	1	–	0
H_2	–	1	–	–
$H_3 \cap H_4$	–	–	0.5	0.5
H_3	–	–	1	–
H_4	–	–	–	1

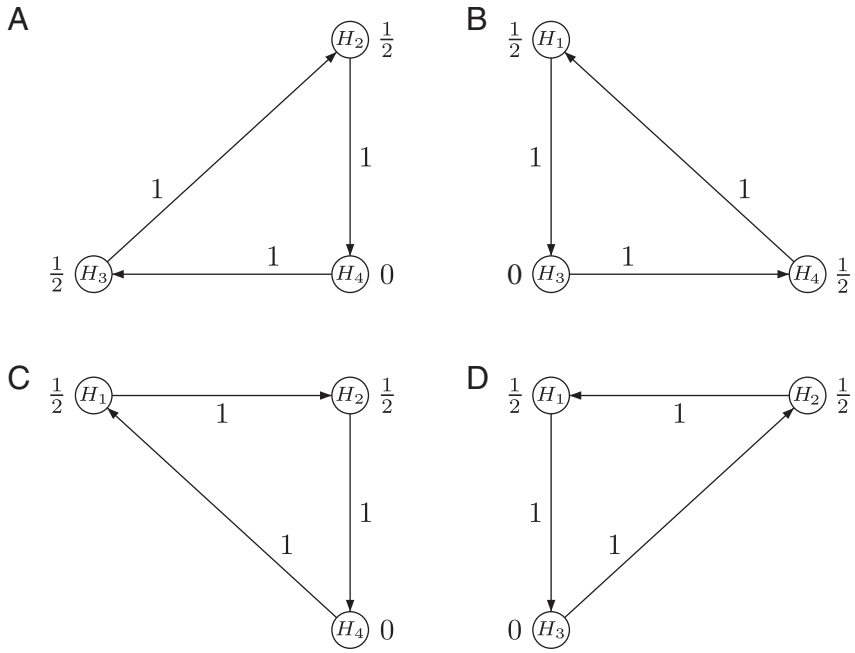


Figure 2 Updated graphs resulting from Figure 1 after removing (A) H_1, (B) H_2, (C) H_3, and (D) H_4.

3 Test procedures

In Section 2, we proposed to dissociate the underlying weighting strategy from the employed test procedure and gave a generic description of the former, illustrated with an example. In this section

we give details on different test procedures that could be employed to test the intersection hypotheses, including weighted Bonferroni tests, weighted min-p tests accounting for the correlation between the test statistics, and weighted Simes' tests.

3.1 Weighted Bonferroni tests

The weighted Bonferroni test introduced in Section 2 is the simplest applicable test procedure, leading to the original graphical approaches by Bretz et al. (2009). Applying the Bonferroni test leads to simple and transparent test procedures that are often easier to communicate than alternative, potentially more powerful approaches. As a matter of fact, the Bonferroni test is often perceived to provide credible trial outcomes in clinical practice. Most importantly in the context of the graphical weighting strategies considered here, applying the Bonferroni test leads to shortcut procedures as long as the monotonicity condition (1) is satisfied. That is, one can start with a graph as shown in Figure 1 and sequentially test the m hypotheses as long as individual null hypotheses H_i, $i \in I$, are rejected. Based on Algorithm 1 from Section 2, we give in the following a similar algorithm that accounts for the weighted Bonferroni tests, thus leading to the sequentially rejective multiple test procedures described in Bretz et al. (2009):

Algorithm 2 (Weighted Bonferroni Test)

(i) Select a $j \in I$ such that $p_j \leq w_j(I)\alpha$ and reject H_j; otherwise stop.
(ii) Update the graph:

$$I \to I \backslash \{j\}$$

$$w_\ell(I) \to \begin{cases} w_\ell(I) + w_j(I) g_{j\ell}, & \ell \in I \\ 0, & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ell \neq k, g_{\ell j} g_{j\ell} < 1 \\ 0, & \text{otherwise} \end{cases}$$

(iii) If $|I| \geq 1$, go to step (i); otherwise stop.

Similar to Algorithm 1, the results in Bretz et al. (2009) ensure that the decisions of the resulting sequentially rejective multiple test procedures remain unchanged regardless of the actual rejection sequence. That is, if in step (i) of Algorithm 2 more than one hypothesis could be rejected, it does not matter with which to proceed. Although Algorithms 1 and 2 have a similar update rule in step (ii), they differ in the way that the index sets are updated. While Algorithm 2 starts with the global index set I and reduces it sequentially as long as hypotheses are rejected, Algorithm 1 removes, for each $J \subseteq I$, consecutively all indices from I that are not contained in J until the set J is obtained. Note that performing a closed weighted Bonferroni test procedure using the weights from Algorithm 1 leads to exactly the same test decisions as performing a sequentially rejective multiple test procedure with Algorithm 2 based on the same starting weights.

Figure 3 gives an example of a Bonferroni-based sequentially rejective multiple test procedures for the weighting strategy proposed in Example 1. Assume, for example, the unadjusted p-values $p_1 = 0.01$, $p_2 = 0.005$, $p_3 = 0.1$, and $p_4 = 0.5$. Then we can reject both H_1 and H_2, but none of the other hypotheses. Figure 3 displays the initial graph together with a possible rejection sequence. As mentioned above, the final decisions on which hypotheses to reject do not depend on the particular rejection sequence. That is, with the initial graph from Figure 3 we would obtain the same decisions, regardless of whether we first reject H_2 and then H_1, or vice versa.

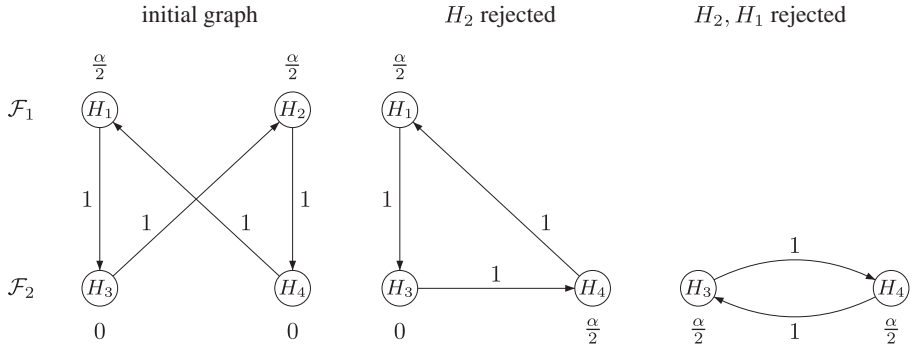


Figure 3 Graph for sequentially rejective procedure with example rejection sequence.

Many standard approaches from the literature can be visualized using Bonferroni-based graphical test procedures, including the weighted or unweighted Bonferroni–Holm procedure (Holm, 1979), fixed sequence tests (Maurer et. al, 1995; Westfall and Krishen, 2001), fallback procedures (Wiens, 2003), and gatekeeping procedures (Bauer et al., 1998; Westfall and Krishen, 2001; Dmitrienko et al., 2003). Adjusted p-values and simultaneous confidence intervals can be calculated as well, although the resulting simultaneous confidence intervals are known to be of limited practical use, as they are often non-informative; see Strassburger and Bretz (2008), Guilbaud (2008, 2009) and Bretz et al. (2009) for details. Bretz et al. (2011) provided `SAS/IML` code to perform the resulting Bonferroni-based sequentially rejective multiple test procedures. In Section 4, we describe the `gMCP` package in `R`, which offers a convenient graphical user interface (GUI) for these approaches.

One general disadvantage of Bonferroni-based approaches is a perceived power loss, motivating the use of weighted parametric tests that account for the correlation between the test statistics or the use of weighted Simes tests. We discuss these alternative test procedures in Sections 3.2 and 3.3, respectively.

3.2 Weighted parametric tests

If for the intersection hypotheses $H_J, J \subseteq I$, the joint distribution of the p-values $p_j, j \in J$, are known, a weighted min-p test can be defined (Westfall and Young, 1993; Westfall et al., 1998). This test rejects H_J if there exists a $j \in J$ such that $p_j \leq c_J w_j(J)\alpha$, where c_J is the largest constant satisfying

$$P_{H_J}\left(\bigcup_{j \in J}\{p_j \leq c_J w_j(J)\alpha\}\right) \leq \alpha. \tag{2}$$

If the p-values are continuously distributed, there is a c_J such that the rejection probability is exactly α. Determination of c_J requires knowledge of the joint null distribution of the p-values and computation of the corresponding multivariate cumulative distribution functions. If the test statistics are multivariate normal or t distributed under the null hypotheses, these probabilities can be calculated using, for example, the `mvtnorm` package in `R` (Genz and Bretz, 2009). Alternatively, resampling-based methods may be used to approximate the joint null distribution; see Westfall and Young (1993).

If $c_J = 1$ in (2), the weighted parametric test reduces to the weighted Bonferroni test. This fully exhausts the level if and only if the joint distribution of continuously distributed p-values with strictly positive density function over $(0,1)^m$ satisfies

$$P_{H_J}(\{p_j \leq c_J w_j(J)\alpha\} \cap \{p_i \leq c_J w_i(J)\alpha\}) = 0$$

for all $i \neq j \in J$, because then all events are pairwise disjoint and $P_{H_J}(\cup_{j\in J}\{p_j \leq c_J w_j(J)\alpha\}) = \sum_{j\in J} P_{H_J}(p_j \leq c_J w_j(J)\alpha)$. Otherwise, $c_J > 1$ and the weighted parametric test gives a uniform improvement over the weighted Bonferroni test from Section 3.1.

If not all, but some of the multivariate distributions of the p-values are known, it is possible to derive conservative upper bounds of the rejection probability that still give an improvement over the Bonferroni test. Assume that I can be partitioned into l sets I_h such that $I = \cup_{h=1}^{l} I_h$ and $I_i \cap I_h = \emptyset$ for $i \neq h = 1,\ldots,l$. We assume that for each $h = 1,\ldots,l$ the joint distribution of the p-values $p_i, i \in I_h$, is known, but the joint distribution of p-values belonging to different I_h is not necessarily known. Now, let $J \subseteq I$ and choose the maximal critical value c_J such that

$$\sum_{h=1}^{l} P_{H_J}\left(\bigcup_{k\in I_h\cap J}\{p_k \leq c_J w_k(J)\alpha\}\right) \leq \alpha. \tag{3}$$

By the Bonferroni inequality, the left-hand side in (2), which cannot be computed if the full joint distribution is unknown, is bounded from above by the left-hand side in (3), whose computation requires only the knowledge of the joint distribution of the p-values in $I_h \cap J$, separately for each $h = 1,\ldots,l$. Thus, any c_J satisfying (3) will also satisfy (2), leading to a conservative test for the intersection hypothesis H_J.

It follows immediately from Eq. (1) that these parametric approaches are consonant if

$$c_J w_j(J) \leq c_{J'} w_j(J') \quad \text{for all } J' \subseteq J \subseteq I \quad \text{and} \quad j \in J'. \tag{4}$$

For p-values following a joint continuous distribution with strictly positive density function over $(0,1)^m$ this is also a necessary consonance condition. This condition is often violated by the weighted parametric tests above. Consider, for example, the Sidak (1967) test for three hypotheses with initial weights 1/3. Assume that for the test of the intersection of any two hypotheses the weights are 1/3 and 2/3. For $\alpha = 0.05$, the critical value $c_J w_j(J)\alpha = 0.01695$ for all three hypotheses in the first step. For all J' with $|J'| = 2$, we have $c_{J'} w_j(J')\alpha = 0.01686$ for the hypothesis H_j with the weight 1/3 in the second step, violating (4). This phenomenon is even more pronounced for positive correlations. If in the previous example the correlations are all 0.5 (corresponding to a Dunnett test in a balanced one-way layout with known variance), we have $c_J w_j(J)\alpha = 0.0196$ and $c_{J'} w_j(J')\alpha = 0.0182$.

If the consonance condition (4) is met, a sequentially rejective test procedure similar to the Bonferroni-based graphical tests from Section 3.1 can be defined.

Algorithm 3 (Weighted Parametric Test)

(i) Choose the maximal constant c_I that satisfies either (2) or (3) for $J = I$.
(ii) Select a $j \in I$ such that $p_j \leq c_I w_j(I)\alpha$ and reject H_j; otherwise stop.
(iii) Update the graph:

$$I \to I\backslash\{j\}$$

$$w_\ell(I) \to \begin{cases} w_\ell(I) + w_j(I) g_{j\ell}, & \ell \in I \\ 0, & \text{otherwise} \end{cases}$$

$$g_{\ell k} \to \begin{cases} \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}, & \ell, k \in I, \ell \neq k, g_{\ell j} g_{j\ell} < 1 \\ 0, & \text{otherwise} \end{cases}$$

(iv) If $|I| \geq 1$, go to step (i); otherwise stop.

For any specific multiple test procedure defined by a given graph, the consonance condition can be checked. If the consonance condition is not met, the weighting strategies introduced in Section 2

remain applicable, although the connection to a corresponding sequentially rejective test procedure is lost. In this case, Algorithm 3 no longer applies and one has to go through the entire closed test procedure. For a given weighting strategy, this procedure is uniformly more powerful than the associated Bonferroni-based procedure from Section 3.1. Note that adjusted p-values for each hypothesis H_i can be obtained by computing p-values for each intersection hypothesis H_J with $i \in J$ (given by the lowest local level for which the respective intersection hypothesis can be rejected) and then taking the maximum over them.

Before illustrating Algorithm 3 with two examples, we notice that Eq. (2) does not provide the only possible definition of a weighted parametric test. Instead of using $c_J w_j(J)\alpha$ as the critical values for p_j, $j \in J$, we could also use some other function $f_J(w_j(J),\alpha)$ fulfilling $f_J(w_j(J),\alpha) \geq w_j(J)\alpha$ for all $j \in J$ and all dependence structures of the p-values. For example, if $T_j = \Phi^{-1}(1-p_j)$ is the test statistic corresponding to the p-value of a z-test for H_j, then finding an ε_J such that

$$1 - P_{H_J}\left(\bigcup_{j \in J}\{T_j \leq \Phi^{-1}(1 - w_j(J)\alpha) - w_j(J)\varepsilon_J\}\right) = \alpha$$

would also define a test which is uniformly more powerful than the corresponding weighted Bonferroni test. A related approach to account for correlations in weighted multiple testing procedures defined by the graphical approach was considered in Millen and Dmitrienko (2011).

Example 2

We revisit the weighting strategy from Example 1. Assume that the joint null distribution of the p-values p_1, p_2 for the two primary dose-control comparisons as well as the joint null distribution of the p-values p_3, p_4 for the two secondary comparisons are known. Applying the standard analysis-of-variance assumptions with a known common variance, we have a bivariate normal distribution, where the correlation is determined only by the relative group sample sizes. In practice, the correlation between primary and secondary endpoints is typically unknown and thus the joint distributions of the pairs $(p_i, p_j), i = 1, 2,\ j = 3, 4$ are also unknown. Therefore, (2) cannot be computed and c_J cannot be determined directly. Setting $I_1 = \{1, 2\}$ and $I_2 = \{3, 4\}$, the joint null distribution of the test statistics for the hypotheses in I_1 and I_2 is known and the constants c_J can be determined by (3). Note that c_J depends on α and on the weights. Table 2 shows the local significance levels for both (A) the closed weighted Bonferroni test procedure and (B) the closed weighted parametric test procedure, assuming $\alpha = 0.025$ and equal group sample sizes.

Using, for example, the `mvtnorm` package in `R`, one can call

```
> myfct <- function(x, a, w, sig) {
+   1 - a - pmvnorm(lower = -Inf, upper = qnorm(1-x*w*a), sigma = sig)
+ }
> sig <- diag(2)*0.5 + 0.5
> uniroot(myfct, lower = 1, upper = 9, a = 0.025, w = rep(0.5, 2),
+   sig = sig)$root
[1] 1.078306
```

to compute $c_J = 1.0783$ for $J = \{3, 4\}$ as well as for all $J \supseteq \{1, 2\}$ and $c_J = 1$ otherwise. In other words, $H_3 \cap H_4$ and all intersection hypotheses that include H_1 and H_2 are tested with unweighted Dunnett z tests. However, intersection hypotheses containing $H_1 \cap H_4$ or $H_2 \cap H_3$ are tested with an unweighted Bonferroni test. As a consequence, the resulting family of tests is not consonant. For example, $c_{\{1,2,3,4\}} w_1(\{1, 2, 3, 4\})\alpha = 0.0135 > 0.0125 = c_{\{1,4\}} w_1(\{1, 4\})\alpha$, violating condition (4). Nevertheless, for a given weighting strategy, the closed test procedure based on parametric weighted tests dominates the associated procedure based on weighted Bonferroni tests. For example, if $p_1 = 0.0131, p_2 = 0.1, p_3 = 0.012$, and $p_4 = 0.01$, the

Table 2 Local significance levels (in %) of A: weighted Bonferroni (B: parametric, C: consonant parametric with $\delta = 0.0783$) test for the example from Figure 1 and $\alpha = 0.025$.

Intersection hypothesis	Local significance levels (in %)			
	H_1	H_2	H_3	H_4
$H_1 \cap H_2 \cap H_3 \cap H_4$	1.25 (1.35,1.35)	1.25 (1.35,1.35)	0 (0,0)	0 (0,0)
$H_1 \cap H_2 \cap H_3$	1.25 (1.35,1.35)	1.25 (1.35,1.35)	0 (0,0)	–
$H_1 \cap H_2 \cap H_4$	1.25 (1.35,1.35)	1.25 (1.35,1.35)	–	0 (0,0)
$H_1 \cap H_2$	1.25 (1.35,1.35)	1.25 (1.35,1.35)	–	–
$H_1 \cap H_3 \cap H_4$	1.25 (1.25,1.35)	–	0 (0,0)	1.25 (1.25,1.15)
$H_1 \cap H_3$	2.50 (2.50,2.50)	–	0 (0,0)	–
$H_1 \cap H_4$	1.25 (1.25,1.35)	–	–	1.25 (1.25,1.15)
H_1	2.50 (2.50,2.50)	–	–	–
$H_2 \cap H_3 \cap H_4$	–	1.25 (1.25,1.35)	1.25 (1.25,1.15)	0 (0,0)
$H_2 \cap H_3$	–	1.25 (1.25,1.35)	1.25 (1.25,1.15)	–
$H_2 \cap H_4$	–	2.50 (2.50,2.50)	–	0 (0,0)
H_2	–	2.50 (2.50,2.50)	–	–
$H_3 \cap H_4$	–	–	1.25 (1.35,1.35)	1.25 (1.35,1.35)
H_3	–	–	2.50 (2.50,2.50)	–
H_4	–	–	–	2.50 (2.50,2.50)

weighted parametric test procedure rejects H_1 and H_3, whereas the Bonferroni test rejects none. In Section 4, we revisit this numerical example and describe the `gMCP` package in `R`, which implements the closed weighted parametric test procedure (B). Related gatekeeping procedures addressing the problem of comparing several doses with a control for multiple hierarchical endpoints were described, among others, by Dmitrienko et al. (2006), Liu and Hsu (2009), and Xu et al. (2009).

Continuing with the example, one can enforce consonance via an appropriate modification of the weighting strategy from Figure 1. To achieve consonance, we introduce additional edges with weight δ (see Figure 4) such that the weight for H_1 (resp. H_2) is sufficiently increased to satisfy the monotonicity condition (4) when testing the intersection hypotheses $H_1 \cap H_4$ and $H_1 \cap H_3 \cap H_4$ (resp. $H_2 \cap H_3$ and $H_2 \cap H_3 \cap H_4$). If $\delta \geq \delta^* := c_{\{1,2,3,4\}} - 1$ the resulting closed test procedure is consonant and Algorithm 3 can be used to perform the test. In the above example with $\alpha = 0.025$, the lower bound is $\delta^* = 0.0783$. Setting $\delta = \delta^*$, we obtain the local significance levels for procedure (C) in Table 2. Note that because of the special weighting strategy employed in this example, these local significance levels are obtained with the regular Dunnett and univariate z tests.

The lower bound δ^* depends on the correlation between the test statistics for H_1 and H_2. Because $c_{\{1,2,3,4\}}$ increases with the correlation, this also holds for δ^*. In the limiting case that the sample size ratios of the dose groups and the control group tend to infinity, the correlation tends to 1. Consequently, $c_{\{1,2,3,4\}} = 2$, such that $\delta^* = 1$ and the graph is degenerated for all $\alpha > 0$. On the other hand, if the above sample size ratios tend to 0, the correlation tends to 0 and $\delta^* = 2(1 - (1 - \alpha)^{1/2})/\alpha - 1$ in limit.

Note that by enforcing consonance, the resulting multiple test procedure based on weighted parametric tests is no longer uniformly better than the associated Bonferroni-based test procedure which does not account for the correlations. That is, for a given weighting strategy, the closed test procedure based on parametric weighted tests may fail to reject certain hypotheses that otherwise are rejected by the associated procedure based on weighted Bonferroni tests. For example, if $p_1 = 0.01, p_2 = 0.1, p_3 = 0.012$, and $p_4 = 0.01$, the initial graph from Figure 3 rejects H_1 and H_3, whereas the consonant weighted parametric test procedure from Figure 4 with $\delta = 0.0783$ rejects only H_1.

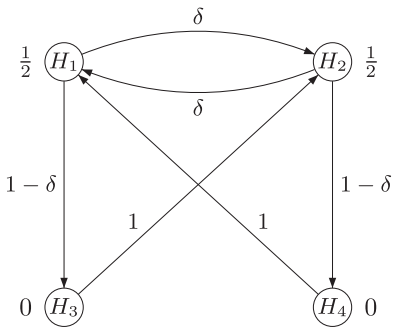


Figure 4 Graphical display of weighting strategy for a consonant weighted parametric test procedure.

Example 3

Consider again Example 1, but assume that H_1,H_2 are two non-inferiority hypotheses (say, for low and high dose against control) and H_3,H_4 are two superiority hypotheses (for the same two doses). We again make the standard analysis-of-variance assumptions with a known common variance and let $\alpha = 0.025$. Bonferroni-based graphical approaches for combined non-inferiority and superiority testing were described in Hung and Wang (2010) and Lawrence (2011). In the following, we exploit the fact that the correlations between the four test statistics are known. Therefore, the complete joint distribution is known and we can apply (2). Note that if $w_j(J) = 0$ for some $j \in J$, the joint distribution degenerates. In our example it thus suffices to calculate bivariate or univariate normal probabilities.

Assume first that the same population is used for all four tests. For simplicity, assume further that the group sample sizes are equal. Then the correlation between the non-inferiority and superiority tests within a same dose is 1; all other correlations are 0.5. Therefore, $c_J = 1.0783$ for $J = \{1,2\},\{1,4\},\{2,3\}$, and $\{3,4\}$. Otherwise, $c_J = 1$ and condition (4) is trivially satisfied. That is, consonance is ensured and one can apply Algorithm 3. This leads to a sequentially rejective multiple test procedure, where at each step either bivariate Dunnett z tests or individual z tests are used. This conclusion remains true if the common variance is unknown and Dunnett t tests or individual t tests are used.

To illustrate the procedure, let $\alpha = 0.025$ and assume the unadjusted p-values $p_1 = 0.01$, $p_2 = 0.02$, $p_3 = 0.005$, and $p_4 = 0.5$. Following Algorithm 3, we have $p_1 \leq c_I w_1(I)\alpha = 0.0135$ and can reject H_1. The update step then leads to the weights in Figure 2(A). Next, $p_3 \leq 0.0135$ and we can reject H_3. This leaves us with H_2, H_4 and the weights $w_2(\{2,4\}) = 1$, $w_4(\{2,4\}) = 0$. Therefore, H_2 is now tested at full level α. Because $p_2 \leq \alpha$, we reject H_2 and the procedure stops.

We now consider the situation that two different populations are used. Assume that the per-protocol population (PP) is used for non-inferiority testing and the intention-to-treat population (ITT) for superiority testing, where PP is a subpopulation of ITT. Let n_i denote the ITT sample size for group i, where $i = 0$ (1,2) denotes placebo (low dose, high dose). Let further $n_i^* \leq n_i$ denote the PP sample size for group i. Finally, let T_i denote the test statistic for $H_i, i = 1,\ldots,4$, and $\rho(T_i,T_j)$ the correlation between T_i and T_j. With this notation,

$$\rho(T_1, T_2) = \rho(T_3, T_4) = \left(\frac{n_1}{n_0+n_1}\right)^{1/2}\left(\frac{n_2}{n_0+n_2}\right)^{1/2}$$

which reduces to 0.5 if $n_0 = n_1 = n_2$. Further,

$$\rho(T_1, T_3) = \left(\frac{n_0+n_1}{n_0 n_1}\right)^{1/2}\left(\frac{n_0^* n_1^*}{n_0^*+n_1^*}\right)^{1/2} \quad \text{and} \quad \rho(T_2, T_4) = \left(\frac{n_0+n_2}{n_0 n_2}\right)^{1/2}\left(\frac{n_0^* n_2^*}{n_0^*+n_2^*}\right)^{1/2},$$

which both reduce to $(n_0^*/n_0)^{1/2}$ for $n_0 = n_i$ and $n_0^* = n_i^*$, $i = 1,2$. Finally, $\rho(T_1, T_4) = \rho(T_1, T_3)\rho(T_3, T_4)$ and $\rho(T_2, T_3) = \rho(T_2, T_4)\rho(T_3, T_4)$, which both reduce to $1/2(n_0^*/n_0)^{1/2}$ for $n_0 = n_1 = n_2$ and $n_0^* = n_1^* = n_2^*$. In this simplest case of equal group sample sizes within PP and ITT we thus have, assuming $n_0^*/n_0 = 0.9$ as an example

$$c_J = \begin{cases} 1 & \text{for } J = \{1,3\},\ J = \{2,4\} \text{ and } J = \{i\},\ i = 1,\ldots,4 \\ 1.0783 & \text{for } J = \{3,4\} \text{ and for all } J \supseteq \{1,2\} \\ 1.0706 & \text{otherwise} \end{cases}$$

As a consequence, the resulting family of tests is no longer consonant, although the differences in the resulting local significance levels are small. For example, $c_{\{1,2,3,4\}} w_1(\{1,2,3,4\})\alpha = 0.0135 > 0.0134 = c_{\{1,4\}} w_1(\{1,4\})\alpha$, violating condition (4). Similar to Example 2, we can enforce consonance by applying the graphical test procedure from Figure 4 with $\delta = 0.0071$.

Finally, we note that this multiple test procedure is immediately applicable to testing for a treatment effect at two different dose levels in an overall population and, if at least one dose is significant, continue testing in a pre-specified subpopulation. This could apply to testing, for example, in the global study population and a regional subpopulation or in the enrolled full population and a targeted genetic subpopulation.

3.3 Weighted Simes tests

Generalization of the original Bonferroni-based graphs from Section 3.1 also apply when the correlations between the test statistics are not exactly known, but certain restriction on them are assumed. A typical case in practice is to assume (or show) that the test statistics have a joint multivariate normal distribution with non-negative correlations. In this case, the Simes test is a popular test. Here, we discuss the use of a weighted version of the Simes test for the intersection hypotheses $H_J, J \subseteq I$.

The unweighted Simes test, as originally proposed by Simes (1986), rejects H_I if there exists a $j \in I$ such that $p_{(j)} \leq j/m\alpha$, where $p_{(1)} \leq \cdots \leq p_{(m)}$ denote the ordered p-values for the hypotheses $H_i, i \in I$. The Type 1 error rate is exactly α if the test statistics are independent and it is bounded by α if positive regression dependence holds. This follows from Benjamini and Yekutieli (2001), who showed false discovery rate control for a related step-up procedure under positive regression dependence on the test statistics. Note that this condition is not always easy to verify or even justify in practice.

The weighted Simes test introduced by Benjamini and Hochberg (1997) rejects H_I if for some $j \in I$ $p_{(j)} \leq \sum_{i=1}^{j} \alpha_{(i)}$, where $\alpha_{(i)} = w_{(i)}\alpha$ and $w_{(i)}$ denotes the weight associated with $p_{(i)}$. An equivalent condition is to reject H_I if for some $j \in I$

$$p_j \leq \sum_{i \in I_j} \alpha_i = \alpha \sum_{i \in I_j} w_i \tag{5}$$

where $I_j = \{k \in I; p_k \leq p_j\}$. This weighted Simes test reduces to the original (unweighted) Simes test if $w_i = 1/m, i \in I$. Kling (2005) showed that the weighted test is conservative if the univariate test statistics are positive regression dependent for any number of hypotheses. This, for example, is the case if the test statistics follow a multivariate normal distribution with non-negative correlations and the tests are one-sided (Benjamini and Heller, 2007).

For given weights $w_j(J)$, $J \subseteq I$, and assuming positive regression dependence among the univariate test statistics for all m hypotheses H_i, $i \in I$, the weighted Simes test can be applied to all intersection hypotheses H_J, $J \subseteq I$. By means of the closure principle the resulting multiple test procedure rejects

H_i, $i \in I$, at level α if for each $J \subseteq I$ with $i \in J$, there exists an index $j \in J$ such that

$$p_j \leq \alpha \sum_{k \in J_j} w_k(J) \tag{6}$$

where $J_j = \{k \in J; p_k \leq p_j\}$. This follows from the application of condition (5) to all subsets $J \subseteq I$, and the fact that any subset of m positive regression dependent test statistics is also positive regression dependent. Related gatekeeping procedures based on the Simes tests were described, among others, by Dmitrienko et al. (2003) and Chen et al. (2005).

If all weights are equal, the above procedure reduces to the procedure by Hommel (1988), which is known not to be consonant. In case of unequal weights, a corresponding sequentially rejective test procedure is not available and one may have to go through the entire closed test procedure using weighted Simes tests for each intersection hypotheses. Nevertheless, for a given weighting strategy, this procedure is uniformly more powerful than an associated Bonferroni-based procedure from Section 3.1. This follows from the fact that any hypothesis rejected by the closed weighted Bonferroni test procedure can also be rejected by the corresponding closed weighted Simes test procedure; see, for example, the Appendix in Maurer et al. (2011).

Although full consonance is generally not available for Simes-based closed test procedures, we can still derive a partially sequentially rejective test procedure which leads to the same test decision as the closed test procedure defined in (6). In the following, we assume that the weights are exhaustive, i.e. $\sum_{k \in J} w_k(J) = 1$ for all subsets $J \in I$.

Algorithm 4 (Weighted Simes Test)

(i) If $p_i > \alpha$ for all $i \in$ I, stop and retain all m hypotheses.
(ii) If $p_i \leq \alpha$ for all $i \in I$, stop and reject all hypotheses.
(iii) Perform the Bonferroni-based graphical test procedure from Section 3.1. Let I_r denote the index set of rejected hypotheses and I_r^c its complement in I. If $|I_r^c| < 3$, stop and retain the remaining hypotheses.
(iv) If $|I_r^c| \geq 3$ consider the weights $w_i(I_r^c), i \in I_r^c$, and the transition matrix $\mathbf{G}$ defined on I_r^c as the new initial graph for the remaining hypotheses. Compute the weights $w_k(J)$ for all $J \subseteq I_r^c$ with Algorithm 1.
(v) Reject $H_i, i \in I_r^c$, if for each $J \subseteq I_r^c$ with $i \in J$, there exists an index $j \in J$ such that

$$p_j \leq \alpha \sum_{k \in J_j} w_k(J). \tag{7}$$

With step (ii), all hypotheses H_i, $i \in I$ can be rejected if $p_j \leq \alpha$ for all $j \in I$. This follows from the fact that for each J there is always a largest p_j, $j \in J$, such that $J_j = J$ and therefore $\alpha \sum_{k \in J_j} w_k(J) = \alpha \sum_{k \in J} w_k(J) = \alpha$. Hence condition (6) holds for all $J \subseteq I$ and therefore for all H_i, $i \in I$. Note that if the weights are not exhaustive, step (ii) may no longer be valid and should be skipped.

The stopping condition in step (iii), $|I_r^c| < 3$, is explained as follows. Assume first that $|I_r^c| = 1$, i.e. one hypothesis is left, say H_i. If $p_i < \alpha$, one would have rejected already all hypotheses in step (ii) and stopped the procedure because for all other hypotheses than H_i necessarily $p_j \leq \alpha$. Therefore, $p_i > \alpha$ and one cannot reject H_i. Similarly, if $|I_r^c| = 2$, the respective p-values cannot be both smaller than α. Also if only one of them, say p_i, is smaller and the other is larger than α, then $p_i > w_i(I_r^c)\alpha$, since otherwise the Bonferroni test in step (iii) would have rejected H_i. In that case the Simes test cannot reject H_i either and hence both remaining hypotheses must be retained.

Algorithm 4 is essentially looking first for outcomes that are easy to verify (steps (i) and (ii)) or where sequential rejection of the hypotheses is possible (step (iii)). Only then one needs to compute for all remaining hypotheses and their subsets the weights and apply the closed weighted Simes

procedure as given in (6). It can happen though that no hypotheses can be rejected in the first three steps and that one has to perform step (iv) with the full set of all m hypotheses. Note that one could, of course, start immediately with step (iv) on the full hypotheses set. The resulting decisions are identical to those obtained with Algorithm 4, because for any given weighting strategy, any hypothesis rejected by the closed weighted Bonferroni test procedure is also rejected by the associated closed weighted Simes test procedure.

Similar to the case that knowledge about the joint distribution of the p-values is partially missing (as discussed in Section 3.2), we consider now the case that positive regression dependence cannot be assumed between all m test statistics. Let $I_h, h = 1, \ldots, l \leq m$, be a a partition (i.e., $I = \cup_{h=1}^{l} I_h$ and $I_h \cap I_i = \emptyset$ for $h \neq i$) such that for each family of hypotheses $H_i, i \in I_h$, positive regression dependence between the respective test statistics holds. Then we can reject $H_J, J \subseteq I$, if for some j and h with $j \in J_h = I_h \cap J$

$$p_j \leq \alpha \sum_{k \in J_{h,j}} w_k(J) \tag{8}$$

where $J_{h,j} = \{k \in J_h; p_k \leq p_j\}$. This procedure controls the Type I error rate at level α for any intersection hypothesis H_J. This is seen as follows. The weighted Simes test is applied separately to each of the partition sets J_h of J. With the definitions for J_h and $H_{h,j}$ above, for a fixed $h \in \{1, \ldots, l\}$, the probability of the event that there exists a $j \in J_h$ such that $p_j \leq \alpha \sum_{k \in J_{h,j}} w_k(J)$, is less than or equal to $\alpha \sum_{k \in J_h} w_k(J)$ by the weighted Simes test. Hence the probability that this happens in any of the partitions J_h is less than $\sum_{h=1}^{l} \alpha \sum_{k \in J_h} w_k(J) = \alpha \sum_{k \in J} w_k(J) = \alpha$ by means of the Bonferroni inequality. For a given partition $I_h, h = 1, \ldots, l$, with "local" regression dependence within the disjunct subsets of associated test statistics, condition (7) in the algorithm hence can be replaced by (8).

We conclude this section with an example. For the weighting strategy from Example 1, the resulting closed weighted Simes test will reject more hypotheses than the related closed weighted Bonferroni test only if all four p-values are less than or equal to α (Maurer et al., 2011). The latter is not the case for the numerical example in Section 3.1, because, for example, $p_3 = 0.1 > 0.025 = \alpha$ and hence no further hypothesis can be rejected. However, if we had instead, for example, $p_3 = 0.015$ and $p_4 = 0.022$, the closed weighted Simes test would reject all four hypotheses, two more than with the closed weighted Bonferroni test. Generally speaking, the weighted Simes test has power advantages over alternative weighted test procedures if the effect sizes are of similar magnitude.

4 gMCP package in R

The `gMCP` package (Rohmeyer and Klinglmueller, 2011) in `R` (R Development Core Team, 2011) currently implements the Bonferroni-based graphical approach from Section 3.1 and the closed weighted parametric tests from Section 3.2. `R` is a language and environment for statistical computing and graphics (Ihaka and Gentleman, 1996). It provides a wide variety of statistical and graphical techniques, and is highly extensible. The latest version of `gMCP` is available at the Comprehensive `R` Archive Network (CRAN) and can be accessed from `http://cran.r-project.org/package=gMCP/`. In the following, we give only a brief illustration of the `gMCP` package. We refer to the installation instructions at `http://cran.r-project.org/web/packages/gMCP/INSTALL` and the accompanying vignette for a description of the full functionality (Rohmeyer and Klinglmueller, 2011).

4.1 Weighted Bonferroni tests with gMCP

We consider the cardiovascular clinical trial example from Dmitrienko and Tamhane (2009) to illustrate the implementation of the Bonferroni-based graphical approach from Section 3.1 in the `gMCP` package. The trial compared a new compound with placebo for two primary and two

secondary endpoints. Consequently, we have two families of hypotheses $\mathcal{F}_1 = \{H_1, H_2\}$ and $\mathcal{F}_2 = \{H_3, H_4\}$.

Dmitrienko and Tamhane (2009) used this example to illustrate the truncated Holm procedure described in Dmitrienko et al. (2008) and Strassburger and Bretz (2008). Given multiple families of hypotheses in a pre-specified hierarchical order, the key idea of truncated tests is to avoid propagating the complete significance level within a family until all its hypotheses are rejected in order to proceed testing the next family in the hierarchy. Instead, once at least one hypothesis is rejected in a given family, a fraction of the significance level is reserved to test subsequent families of hypotheses. In principle, truncation can be applied to any of the test procedures discussed in Section 3.

In the cardiovascular study example, the hypotheses in $\mathcal{F}_2$ are only tested, if at least one of the hypotheses in $\mathcal{F}_1$ are rejected. We assume that $\mathcal{F}_1$ is tested using the truncated Holm procedure with truncation parameter $\gamma \in [0,1]$. Let $p_{(1)} < p_{(2)}$ denote the ordered p-values with associated hypotheses $H_{(1)}$ and $H_{(2)}$. Consequently, $H_{(1)}$ is tested at level $\alpha/2$. If $H_{(1)}$ is rejected, $H_{(2)}$ is tested at level $\alpha/2+\gamma(\alpha/2)$. The family $\mathcal{F}_2$ is then tested with the regular Holm procedure either at level $(1-\gamma)\alpha/2$ or at level α, depending on whether only one or both hypotheses in $\mathcal{F}_1$ are rejected, respectively.

The `gMCP` package offers a GUI to conveniently create and perform Bonferroni-based graphical test procedures, such as the one for the test procedure above. To this end, we invoke in `R` the `gMCP` package and subsequently call the GUI with

```
> library(gMCP)
> graphGUI()
```

Different buttons are available in the icon panel of the GUI to create a new graph. The main functionality includes the possibility of adding new nodes as well as new edges connecting any two selected nodes. In many cases, the edges will have to be dragged manually in order to improve the readability of the graphs. The associated labels, weights, and significant levels can be edited directly in the graph. Alternatively, the numerical information can be entered into the transition matrix and other fields on the right-hand side of the GUI. Figure 5 displays the complete test procedure for the cardiovascular study example using the `gMCP` package: The truncated Holm procedure for $\mathcal{F}_1$ with truncation parameter γ and the regular Holm procedure for $\mathcal{F}_2$. Note that we can immediately improve that test procedure by connecting the secondary hypotheses H_3 and H_4 with the primary hypotheses H_1 and H_2 through the ε-edges introduced in Bretz et al. (2009). We refer to the vignette of the `gMCP` package for a description of how to construct ε-edges with the GUI (Rohmeyer and Klinglmueller, 2011).

The GUI offers the possibility to perform sequentially Bonferroni-based test procedures defined through a graph like the one displayed in Figure 5 and in addition to calculate adjusted p-values as well as simultaneous confidence intervals. To illustrate this functionality, we consider Scenario 1 from Dmitrienko and Tamhane (2009) and assume the unadjusted p-values $p_1 = 0.0121$, $p_2 = 0.0337$, $p_3 = 0.0084$, and $p_4 = 0.0160$, which are entered directly into the GUI. By clicking on the corresponding button in the icon panel and and specifying $\gamma = 0.5$, one obtains in this example the adjusted p-values 0.024, 0.045, 0.045, and 0.045 for the four hypotheses H_1, H_2, H_3, and H_4, respectively. These adjusted p-values are identical to those reported in Dmitrienko and Tamhane (2009). Accordingly, one can reject all four hypotheses at level $\alpha = 0.05$. Simultaneous confidence intervals can be obtained as well from the GUI after entering additional information on effect estimates and standard errors. Finally, the user may perform the sequential test procedure by clicking on the green triangle in the icon bar. By doing so, the "Reject" buttons in the lower right become activated and one can step through the graph as long as significances occur.

4.2 Weighted parametric tests with `gMCP`

The `gMCP` package provides also a convenient interface to perform graphical test procedures without the GUI using the `R` command line. We illustrate this with the closed weighted parametric

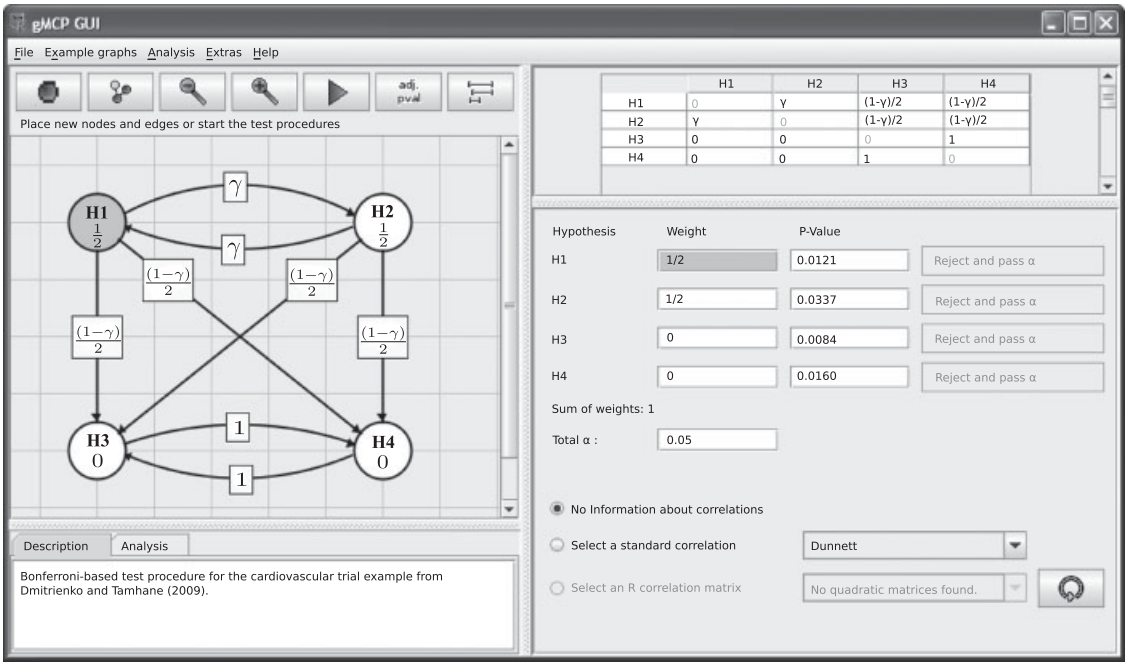


Figure 5 Screenshot of the GUI from the `gMCP` package. Left: Display of the graphical Bonferroni-based test procedure for the cardiovascular trial. Right: Transition matrix, initial weights and unadjusted *p*-values.

tests from Section 3.2 and revisit Example 2. We first define the related transition matrix **G** and the weights $w_i(I)$, $i \in I$, through

```
> G      <- matrix(0, nr=4, nc = 4)
> G[1,3] <- G[2,4] <- G[3,2] <- G[4,1] <- 1
>w       <- c(1/2, 1/2, 0, 0)
```

The function `matrix2graph` then converts the matrix `G` and the vector `w` into an object of type `graphMCP`

```
> graph <- matrix2graph(G, w)
> graph
A graphMCP graph
Overall alpha: 1
H1 (not rejected, alpha=0.5)
H2 (not rejected, alpha=0.5)
H3 (not rejected, alpha=0)
H4 (not rejected, alpha=0)
Edges:
H1 -(1)-> H3
H2 -(1)-> H4
H3 -(1)-> H2
H4 -(1)-> H1
```


The gMCP function takes objects of the type graphMCP as its input together with a vector of p-values and performs the specified multiple test procedure. In particular, one can specify a correlation matrix with the effect that a closed weighted parametric multiple test procedure is performed under the standard analysis-of-variance assumptions with known common variance.

In Example 2 we assumed normally distributed test statistics with a block-diagonal correlation matrix of the form

$$\begin{pmatrix} 1 & 0.5 & \text{NA} & \text{NA} \\ 0.5 & 1 & \text{NA} & \text{NA} \\ \text{NA} & \text{NA} & 1 & 0.5 \\ \text{NA} & \text{NA} & 0.5 & 1 \end{pmatrix},$$

where NA reflects the fact that the correlation between the primary and secondary endpoints is unknown. Accordingly, we let

```
> cr       <- matrix(NA, nr = 4, nc = 4)
> diag(cr) <- 1
> cr[1,2]  <- cr[2,1] <- cr[3,4] <- cr[4,3] <- 1/2
```

and define the unadjusted p-values

```
> p <- c(0.0131, 0.1, 0.012, 0.01)
```

Finally, we perform the closed weighted parametric test at a specified significance level $\alpha = 0.025$, say, by calling

```
> res <- gMCP(graph, p, corr = cr, alpha = 0.025)
```

This returns an object of class gMCPResult providing information on which hypotheses are rejected

```
> res@rejected
   H1    H2    H3    H4
TRUE FALSE TRUE FALSE
```

We conclude from the output that both H_1 and H_3 can be rejected. We come to the same conclusions, if we report the adjusted p-values and compare them with $\alpha = 0.025$

```
> res@adjPValues
          H1          H2          H3          H4
 0.02431856 0.10000000 0.02431856 0.10000000
```

Alternatively, one can use a sequentially rejective Bonferroni-based test procedure from Section 3.2 by omitting the corr argument

```
> gMCP(graph, p, alpha = 0.025)@rejected
    H1    H2    H3    H4
FALSE FALSE FALSE FALSE
```

As seen from the output, none of the null hypotheses can be rejected, which coincides with our conclusions from Section 3.2.

5 Discussion

This paper shows that the graphical approach introduced by Bretz et al. (2009) can be used to create and visualize tailored strategies for common multiple test problems. By dissociating the underlying weighting strategy from the employed test procedure, it is seen that the graphical approach is not restricted to Bonferroni-based tests. Similarly, the graphs introduced by Burman et al. (2009) define weights for all intersection hypotheses and the procedures discussed in this paper can be applied using these weights. Extended graphical approaches include weighted Simes tests and weighted min-p tests in the sense of Westfall and Young (1993). The latter take into account all or some of the joint multivariate distributions of p-values. Consonance and the corresponding shortcuts may be lost, but for any concrete multiple test strategy, consonance can be checked prior to a clinical study. As shown in this paper, consonance can be enforced and related sequentially rejective graphs established at least in some simple situations.

Many proposed multiple test procedures in the literature can be expressed with the methods described in this paper. On the other hand, the methods in this paper also allow one to investigate alternative procedures that go beyond the published results. But even if the closure principle is very common in practice, it does not necessarily lead to consonant multiple test procedures. We gave monotonicity conditions for ensuring consonant graphical weighting strategies, but it is not always clear when these conditions are met if weighted parametric or Simes tests are used. In principle, one could enforce consonance following, for example, the approach of Romano et al. (2011), although the computation of the rejection regions could become tedious. We leave this topic for further research.

Acknowledgements This paper is based on an invited presentation given at the BfArM Symposium on "Multiplicity Issues in Clinical Trials". The authors thank Dr. Norbert Benda (BfArM) for organizing and chairing this symposium. They are also grateful to three referees and the editor for their helpful comments. Part of this research was funded by the Austrian Science Fund (FWF): P23167.

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

References

Alosh, M. and Huque, M. F. (2009). A flexible strategy for testing subgroups and overall population. *Statistics in Medicine* **28**, 3–23.

Alosh, M. and Huque, M. F. (2010). A consistency-adjusted alpha-adaptive strategy for sequential testing. *Statistics in Medicine* **29**, 1559–1571.

Bauer, P. (1991). Multiple testing in clinical trials. *Statistics in Medicine* **10**, 871–890.

Bauer, P., Röhmel, J., Maurer, W. and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

Benjamini, Y. and Heller, R. (2007). False discovery rates for spatial signals. *Journal of the American Statistical Association* **102**, 1272–1281.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Brannath, W. and Bretz, F. (2010). Shortcuts for locally consonant closed test procedures. *Journal of the American Statistical Association* **105**, 660–669.

Bretz, F., Maurer, W., Brannath, W. and Posch, M. (2009). A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* **28**, 586–604.

Bretz, F., Maurer, W. and Hommel, G. (2011). Test and power considerations for multiple endpoint analyses using sequentially rejective graphical procedures. *Statistics in Medicine* **30**, 1489–1501.

Burman, C. F., Sonesson, C. and Guilbaud, O. (2009). A recycling framework for the construction of Bonferroni-based multiple tests. *Statistics in Medicine* **28**, 739–761.

Chen, X., Luo, X. and Capizzi, T. (2005). The application of enhanced parallel gatekeeping strategies. *Statistics in Medicine* **24**, 1385–1397.

Dmitrienko, A., Offen, W., Wang, O. and Xiao, D. (2006). Gatekeeping procedures in dose-response clinical trials based on the Dunnett test. *Pharmaceutical Statistics* **5**, 19–28.

Dmitrienko, A., Offen, W. W. and Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* **22**, 2387–2400.

Dmitrienko, A. and Tamhane, A. C. (2009). Gatekeeping procedures in clinical trials. In *Multiple Testing Problems in Pharmaceutical Statistics,* Dmitrienko, A., Tamhane, A. C. and Bretz, F. (Eds.). Chapman & Hall/CRC Biostatistics Series, Boca Raton.

Dmitrienko, A., Tamhane, A. and Wiens, B. (2008). General multi-stage gatekeeping procedures. *Biometrical Journal* **50**, 667–677.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Springer, Heidelberg.

Guilbaud, O. (2008). Simultaneous confidence regions corresponding to Holm's stepdown procedure and other closed-testing procedures. *Biometrical Journal* **50**, 678–692.

Guilbaud, O. (2009). Alternative confidence regions for Bonferroni-based closed-testing procedures that are not alpha-exhaustive. *Biometrical Journal* **51**, 721–735.

Guilbaud, O. and Karlsson, P. (2011). Confidence regions for Bonferroni-based closed tests extended to more general closed tests. *Journal of Biopharmaceutical Statistics* **21**, 682–707.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika* **75**, 383–386.

Hommel, G., Bretz, F. and Maurer, W. (2007). Powerful short-cuts for multiple testing procedures with special reference to gatekeeping strategies. *Statistics in Medicine* **26**, 4063–4073.

Hung, H. M. J. and Wang, S. J. (2009). Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* **19**, 1–11.

Hung, H. M. J. and Wang, S. J. (2010). Challenges to multiple testing in clinical trials. *Biometrical Journal* **52**, 747–756.

Huque, M. F. and Alosh, M. (2008). A flexible fixed-sequence testing method for hierarchically ordered correlated multiple endpoints in clinical trials. *Journal of Statistical Planning and Inference* **138**, 321–335.

Huque, M. F., Alosh, M. and Bhore, R. (2011). Addressing multiplicity issues of a composite endpoint and its components in clinical trials. *Journal of Biopharmaceutical Statistics* **21**, 610–634.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

Kling, Y. (2005). *Issues of Multiple Hypothesis Testing in Statistical Process Control*. Thesis, The Neiman Library of Exact Sciences & Engineering, Tel-Aviv University.

Lawrence, J. (2011). Testing non-inferiority and superiority for two endpoints for several treatments with a control. *Pharmaceutical Statistics*. DOI: 10.1002/pst.468.

Li, J. and Mehrotra, D. (2008). An efficient method for accommodating potentially underpowered primary endpoints. *Statistics in Medicine* **27**, 5377–5391.

Liu, Y. and Hsu, J. (2009). Testing for efficacy in primary and secondary endpoints by partitioning decision paths. *Journal of the American Statistical Association* **104**, 1661–1670.

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Maurer, W., Glimm, E. and Bretz, F. (2011). Multiple and repeated testing of primary, co-primary and secondary hypotheses. *Statistics in Biopharmaceutical Research* **3**, 336–352.

Maurer, W., Hothorn, L. and Lehmacher, W. (1995). Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In: *Biometrie in der Chemisch-Pharmazeutischen Industrie*, Vollmar, J. (Ed.). Fischer Verlag, Stuttgart 3–18.

Millen, B. A. and Dmitrienko, A. (2011). Chain procedures: A class of flexible closed testing procedures with clinical trial applications. *Statistics in Biopharmaceutical Research* **3**, 14–30.

O'Neill, R. T. (1997). Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* **18**, 550–556.

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org`

Rohmeyer, K. and Klinglmueller, F. (2011). gMCP: A graphical approach to sequentially rejective multiple test procedures. R package version 0.6-5. `http://cran.r-project.org/package=gMCP`

Romano, J. R., Shaikh, A. and Wolf, M. (2011). Consonance and the closure method in multiple testing. *The International Journal of Biostatistics* **7**, Article 12.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* **62**, 626–633.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

Strassburger, K. and Bretz, F. (2008). Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni based closed tests. *Statistics in Medicine* **27**, 4914–4927.

Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

Westfall, P. H., Krishen, A. and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine* **17**, 2107–2119.

Westfall, P. H., Young, S. S. (1993). *Resampling-based Multiple Testing*. Wiley, New York.

Wiens, B. L. (2003). A fixed sequence Bonferroni procedure for testing multiple endpoints. *Pharmaceutical Statistics* **2**, 211–215.

Xu, H. Y., Nuamah, I., Liu, J. Y., Lim, P. and Sampson, A. (2009). A Dunnett–Bonferroni-based parallel gatekeeping procedure for dose-response clinical trials with multiple endpoints. *Pharmaceutical Statistics* **8**, 301–316.

Research Article

Statistics in Medicine

Received 6 June 2009, Accepted 28 August 2009 Published online 13 October 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3748

Hierarchical testing of multiple endpoints in group-sequential trials

Ekkehard Glimm,*† Willi Maurer and Frank Bretz

We consider the situation of testing hierarchically a (key) secondary endpoint in a group-sequential clinical trial that is mainly driven by a primary endpoint. By 'mainly driven', we mean that the interim analyses are planned at points in time where a certain number of patients or events have accrued on the primary endpoint, and the trial will run either until statistical significance of the primary endpoint is achieved at one of the interim analyses or to the final analysis. We consider both the situation where the trial is stopped as soon as the primary endpoint is significant as well as the situation where it is continued after primary endpoint significance to further investigate the secondary endpoint. In addition, we investigate how to achieve strong control of the familywise error rate (FWER) at a pre-specified significance level α for both the primary and the secondary hypotheses. We systematically explore various multiplicity adjustment methods. Starting point is a naive strategy of testing the secondary endpoint at level α whenever the primary endpoint is significant. Hung *et al.* (*J. Biopharm. Stat.* 2007; 17:1201–1210) have already shown that this naive strategy does not maintain the FWER at level α. We derive a sharp upper bound for the rejection probability of the secondary endpoint in the naive strategy. This suggests a number of multiple test strategies and also provides a benchmark for deciding whether a method is conservative or might be improved while maintaining the FWER at α. We use a numerical example based on a real case study to illustrate the results of different hierarchical test strategies.

Keywords: closed test procedure; error spending approach; primary endpoint; secondary endpoint

1. Introduction

Confirmatory clinical trials typically include hundreds or thousands of patients and may last for several years. Interim analyses are often conducted in such large trials because of ethical and economical reasons: (i) patients should not be treated with a new therapy if the ongoing trial gives no indication for a potential benefit; (ii) clinical trials should not be continued (and decisions postponed) if a clear tendency favoring a particular treatment evolves. Thus, clinical trial designs that include the possibility for early decisions may help in reducing the overall costs and timelines of the development program for a new therapy. Repeatedly looking at the data may inflate the Type I error rate because the primary null hypothesis is tested anew at each interim analysis. Group-sequential design methodology is commonly used to account for the repeated data analyses and decision making: the trial can be stopped at any pre-specified interim analysis for either futility or efficacy while controlling the Type I error rate at a pre-specified significance level α. Most literature on group-sequential designs, however, focuses on two-armed trials and a single endpoint [1–4]. Extensions of these methods to multi-armed clinical trials and the control of the familywise Type I error rate (FWER) were investigated, among others, in [5–8].

A related topic, which has been considered much less in the literature, is the application of group-sequential designs to multiple endpoints. In this paper we consider group-sequential trial designs for testing hierarchically two (or more) endpoints. That is, the endpoints are tested in a fixed sequence, each at level α, until the first non-significance [9, 10]. Consider, for example, the simple situation of a two-armed trial with one interim analysis and two endpoints (say, one primary and one secondary endpoint). Assume that some group-sequential method (for example, [1, 2, 11]) is used to test the primary endpoint at level α. The question arises at which significance level the secondary endpoint can be tested once the primary endpoint has been declared significant, at either the interim or final analysis, while controlling the FWER across both endpoints.

Motivated by the fixed sequence test procedure for classical trial designs without interim analyses, a *naive* strategy is to test the secondary endpoint at level α whenever the primary endpoint is significant. However, Hung *et al.* [12] have already shown that this naive strategy does not maintain the FWER at level α. Using analytical methods, we demonstrate in this paper that the

Novartis Pharma AG, Statistical Methodology, Novartis Campus, CH-4056 Basel, Switzerland
**Correspondence to: Ekkehard Glimm, Novartis Pharma AG, Statistical Methodology, Novartis Campus, CH-4056 Basel, Switzerland.*
†E-mail: ekkehard.glimm@novartis.com

size inflation can be substantial. For example, if the interim analysis is performed after 50 per cent of the patients, the actual Type I error can be as large as 0.08 for $\alpha=0.05$.

An alternative strategy, which follows from the closed sequential test procedure described by Tang and Geller [6], is to apply an individual set of group-sequential boundaries separately to each endpoint. That is, if in the example above the primary endpoint is declared significant using its group-sequential boundary (Pocock, O'Brien-Fleming, ...), the secondary endpoint can be tested at its own group-sequential boundary (Pocock, O'Brien-Fleming, ...) matching to the timepoint, at which the the primary endpoint was declared significant. In essence, Tang and Geller [6] have shown that if a group-sequential test procedure at level α can be defined for each intersection hypothesis in a multiple hypothesis testing setting, then the application of the closure principle [13] leads to a sequential test procedure that protects the FWER at level α. In our case the group-sequential design for the primary hypothesis is also the test for the intersection hypothesis of the primary and the secondary variables; by construction, the closure principle results in the hierarchical approach described above.

In this paper we consider the situation of testing a (key) secondary endpoint in a group-sequential clinical trial which is mainly driven by a primary endpoint. By 'mainly driven', we mean that the interim analyses are planned at points in time where a certain number of patients or events have accrued on the primary endpoint and the trial will run either until statistical significance on the primary endpoint is achieved at one of the interim analyses or to the final analysis. We consider both the situation where the trial is stopped as soon as the primary endpoint is significant as well as the situation where it is continued after primary endpoint significance to further investigate the secondary endpoint. In addition, we investigate the properties of various multiplicity adjustment methods with respect to power and strong control of the FWER at a pre-specified significance level α for both primary and secondary hypotheses. Starting point is the derivation of a sharp upper bound for the rejection probability of the secondary endpoint in the naive strategy mentioned above. This suggests a number of multiple test strategies and also provides a benchmark for deciding whether a method is conservative or might be improved while maintaining the overall significance level α.

Different error spending function approaches with local control of the Type I error rate are available for each of the two endpoints. The reasons for choosing one of them depend on the hierarchical approach chosen. The choice for the primary endpoint is not primarily guided by statistical power considerations but rather by the wish to stop the trial early only if the results are so convincing that it could be regarded as unethical to continue the trial. If the interim results are less conclusive it is preferred to continue the trial to increase the body of evidence and to collect sufficient (long-term) safety data. Hence, the choice of the error spending function for the primary endpoint is essentially based on the same considerations as in a trial with only one endpoint. In general, a spending function that allocates less significance level at the interim analysis than at the final analysis (O'Brien-Fleming type) is preferred over a more balanced spending function (Pocock type). These latter reasons, however, are not anymore relevant for the secondary endpoint, when stopping of the trial depends on the primary endpoint. In this case power considerations can be the guiding principle for the choice of the error spending function for the secondary endpoint. As shown later, using Pocock boundaries for the secondary endpoint is advantageous in many of the situations considered in this paper.

In Section 2 we introduce different group-sequential test strategies to test hierarchically a primary and a (key) secondary hypothesis. In Section 3 we derive for the different strategies sharp upper bounds for the probability of rejecting the secondary hypothesis, when in fact it is true, and discuss various consequences. In Section 4 we report selected results of an extensive numerical study to investigate the power of rejecting the secondary hypothesis for various combinations of error spending functions. In Section 5 we use a numerical example based on a real case study to illustrate the different hierarchical test strategies. Finally, in Section 6 we give concluding remarks.

2. Group-sequential test strategies for primary and secondary hypotheses

As outlined in the Introduction, we focus on testing hierarchically one primary and one secondary endpoint. For simplicity, we also restrict the description to clinical trials with one interim and one final analysis, resulting in a total of $k=2$ analyses.

Let $H_p:\theta_p\leqslant 0$ denote the one-sided null hypothesis of no effect for the primary endpoint and let $H_s:\theta_s\leqslant 0$ denote the related secondary hypothesis. Let $Z_{p,i}$ and $Z_{s,i}$ denote the test statistics for H_p and H_s, respectively, at the interim analysis $i=1$ and the final analysis $i=2$. Furthermore, let $t_{p,i}$ and $t_{s,i}$ denote respective information fractions at analysis i. For example, $Z_{p,i}$ and $Z_{s,i}$ may denote the two-sample t-test statistics in case of two normally distributed endpoints or the logrank test statistics in case of a time-to-event analysis. The information fractions will typically be given by $t_{p,i}=t_{s,i}=n_i/n$, where n denotes the total number of observed patients per treatment arm and n_i denotes the number of patients per treatment arm whose response is available at the ith analysis. In the time-to-event case, $t_{v,i}$, $v=p, s$, denotes the fraction of events accrued on endpoint v at analysis i relative to the total number of events accrued for the trial. Information fractions may be fixed before the start of the trial or they may be random variables themselves. We will treat them as fixed in the remainder of this paper.

For each hypothesis H_v, $v=p,s$, we define univariate spending functions $a_v(t)$, $t\in[0,1]$. We denote by $\alpha_v=a_v(1)$ the significance level spent up to the final analysis for endpoint $v=p,s$. Although the significance levels α_p and α_s are allowed to be different, in the following we assume $\alpha_p=\alpha_s=\alpha$ for simplicity. We further denote the respective nominal one-sided rejection boundaries for the interim and final analysis by $u_{v,i}$, $v=p,s$, and $i=1,2$. Applying standard error spending approaches [11, 14] implies that the interim test of the null hypothesis H_v is performed at level $\alpha_{v,1}=a_v(t_{v,1})$. At the final analysis, the rejection boundary $u_{v,2}$ of the

final test is the solution of

$$P(Z_{v,1}<u_{v,1}, Z_{v,2}\geqslant u_{v,2})=\alpha-\alpha_{v,1}$$

under the conditions of H_v. Suppose we convert the critical values into a test level by taking $1-F(u_{v,i})$, where $F(\cdot)$ denotes the marginal cumulative distribution function (cdf) of $Z_{v,i}$ under H_v. The quantities $1-F(u_{v,i}), i=1,2$, are then denoted as *nominal* level of the interim and final tests, in contrast to the *actual* or *spent* levels $\alpha_{v,1}$ and $\alpha_{v,2}=\alpha-\alpha_{v,1}$. Note that for the first interim analysis, the nominal and actual levels are the same (i.e. $1-F(u_{v,1})=\alpha_{v,1}$), whereas for the final analysis we have $1-F(u_{v,2})\geqslant\alpha_{v,2}$, $v=p,s$. Pocock [1] and O'Brien and Fleming [2] originally suggested rejection boundary adjustments for the critical values directly, that is $u_{v,1}=u_{v,2}$ (Pocock) and $\sqrt{t_{v,1}}u_{v,1}=u_{v,2}$ (O'Brien-Fleming). As these quantities can be converted to the actual and/or nominal levels of an error spending approach in the case of just two analyses (one interim and the final analysis), we will call them 'Pocock' and 'O'Brien-Fleming' error spending approaches in the remainder of this paper. If, for example, $\alpha=0.025$ and $t_{p,1}=0.5$, then $u_{p,1}=u_{p,2}=2.178$ for Pocock error spending. Equivalently, we can state that the nominal level is $1-F(u_{p,1})=1-F(u_{p,2})=0.0147$ and that the spent levels are $\alpha_{p,1}=0.0147$ and $\alpha_{p,2}=0.0103$, respectively.

In practice, there are different possibilities of implementing a hierarchical test procedure in a group-sequential trial. In the following, we describe various test strategies for the primary and secondary endpoint situation, which accommodate different study objectives. In order to discuss the properties of the four different strategies, let $R_j(H)$ denote the rejection probability of a null hypothesis H by one of the approaches $j=a,b,c,d$ described below (irrespective of whether H is true or not). A hypothesis H_v, $v=p,s$, is rejected if the respective test statistics $Z_{v,i}$ exceeds the critical value $u_{v,i}$ at either the interim or final analysis. We note in passing that this excludes the possibility 're-testing'. That is, if an endpoint is declared significant at the interim analysis, it will not be tested again at the final analysis.

(*a*) *Stagewise hierarchical*: At the interim analysis the primary hypothesis H_p is tested with critical value $u_{p,1}$. If $Z_{p,1}\geqslant u_{p,1}$, we can reject H_p, stop the trial and test the secondary hypothesis H_s with critical value $u_{s,1}$. Otherwise, the trial continues and the primary hypothesis H_p is tested with critical value $u_{p,2}$ at the final analysis. If $Z_{p,2}\geqslant u_{p,2}$, we can reject H_p and test the secondary hypothesis H_s with critical value $u_{s,2}$. Consequently, the marginal probability to reject the secondary hypothesis H_s at either the interim or final analysis using the stagewise hierarchical strategy is

$$R_a(H_s)=P(Z_{p,1}\geqslant u_{p,1}, Z_{s,1}\geqslant u_{s,1})+P(Z_{p,1}<u_{p,1}, Z_{p,2}\geqslant u_{p,2}, Z_{s,2}\geqslant u_{s,2}) \tag{1}$$

(*b*) *Overall hierarchical*: Same principle as the stagewise hierarchical procedure except that the trial does not stop if at the interim analysis H_p can be rejected, but not H_s. In this case the trial continues to the final sample size where H_s is tested again with critical value $u_{s,2}$. Thus,

$$R_b(H_s)=R_a(H_s)+P(Z_{p,1}\geqslant u_{p,1}, Z_{s,1}<u_{s,1}, Z_{s,2}\geqslant u_{s,2}) \tag{2}$$

(*c*) *Partially hierarchical*: Same principle as the overall hierarchical procedure except that H_s can be tested at the final analysis irrespective of whether H_p has been rejected before. In other words, H_s is only tested at the interim analysis if H_p has been rejected at that point. If H_s cannot be rejected at the interim analysis, it can be tested again at the final analysis. Such a strategy could be sensible if the secondary endpoint is actually of prime interest and the primary endpoint is only a surrogate marker to indicate when the secondary endpoint should be tested. Therefore,

$$\begin{aligned} R_c(H_s) &= R_b(H_s)+P(Z_{p,1}<u_{p,1}, Z_{p,2}<u_{p,2}, Z_{s,2}\geqslant u_{s,2}) \\ &= P(Z_{p,1}\geqslant u_{p,1}, Z_{s,1}\geqslant u_{s,1})+P(Z_{p,1}<u_{p,1}, Z_{s,2}\geqslant u_{s,2})+P(Z_{p,1}\geqslant u_{p,1}, Z_{s,1}<u_{s,1}, Z_{s,2}\geqslant u_{s,2}) \end{aligned} \tag{3}$$

(*d*) *Coequal*: Primary and secondary hypotheses are tested separately. The trial stops at interim only if both hypotheses can be rejected. Consequently, the marginal probability to reject the secondary hypothesis H_s using the coequal strategy is

$$\begin{aligned} R_d(H_s) &= R_c(H_s)+P(Z_{p,1}<u_{p,1}, Z_{s,1}\geqslant u_{s,1}, Z_{s,2}<u_{s,2}) \\ &= P(Z_{s,1}\geqslant u_{s,1})+P(Z_{s,1}<u_{s,1}, Z_{s,2}\geqslant u_{s,2}) \end{aligned} \tag{4}$$

Owing to their hierarchical nature, strategies (a) and (b) keep the FWER at level α across both endpoints, if proper error spending approaches, each at level α, are applied separately to both endpoints. Strategies (c) and (d) maintain the Type I error rate separately for each endpoint, but not the FWER across both endpoints. However, strategies (c) and (d) look very similar to strategies (a) and (b) and are sometimes put up for discussion in clinical teams. Furthermore, strict control of the FWER may not always be required in a clinical trial, but still the procedures' properties with respect to this concept can be of interest to assess whether the associated severity of Type I error inflation is deemed acceptable. Interestingly, with the results from Section 3.1 upper bounds for the probability of rejecting the secondary hypothesis H_s can be derived which are attained by all four strategies under the same least favorable parameter configuration.

It follows from the previous formulas that a rejection of H_s by any strategy implies the rejection of the subsequent ones in the order given above. That is,

$$R_a(H_s) \leqslant R_b(H_s) \leqslant R_c(H_s) \leqslant R_d(H_s) \tag{5}$$

holds irrespective of the multivariate distribution of the vector $(Z_{p,1}, Z_{p,2}, Z_{s,1}, Z_{s,2})'$. If H_s is true, inequality (5) describes the relation between the true level of the respective test and otherwise that of power. Note also that for the primary hypothesis we have

$$R_j(H_p) = P(Z_{p,1} \geqslant u_{p,1}) + P(Z_{p,1} < u_{p,1}, Z_{p,2} \geqslant u_{p,2}) \tag{6}$$

for $j = a, b, c, d$. Consequently, the marginal rejection probability of the primary hypothesis H_p is the same for all four strategies.

Further strategies are conceivable. For example, a referee mentioned a strategy that 'looks back' at the interim results of the secondary endpoint when the primary endpoint is significant at the final analysis, but not at the interim. Such a strategy would also keep the FWER. Regarding inequality (5), this strategy would fall between $R_a(H_s)$ and $R_d(H_s)$. Hence, the results from Section 3 also apply to this strategy. We note in passing that such a strategy will less likely be applied in practice, because the power loss from only doing the final test and not the interim test would be very small, while it would be difficult to find a satisfactory explanation if a significant interim result is no longer supported by the final analysis.

For the remainder of this paper, we assume that the vector of test statistics $(Z_{p,1}, Z_{s,1}, Z_{p,2}, Z_{s,2})'$ follows the multivariate normal distribution

$$N\left(\begin{pmatrix} \sqrt{t_{p,1}}\delta_p \\ \sqrt{t_{s,1}}\delta_s \\ \delta_p \\ \delta_s \end{pmatrix}, \begin{pmatrix} 1 & \rho & \sqrt{t_{p,1}} & \rho\sqrt{t_{p,1}} \\ \rho & 1 & \rho\sqrt{t_{s,1}} & \sqrt{t_{s,1}} \\ \sqrt{t_{p,1}} & \rho\sqrt{t_{p,1}} & 1 & \rho \\ \rho\sqrt{t_{s,1}} & \sqrt{t_{s,1}} & \rho & 1 \end{pmatrix}\right) \tag{7}$$

where $\delta_v = E(Z_{v,2})$, $v = p, s$. Here, ρ denotes the correlation between the two endpoints and $\sqrt{t_{v,1}}$ denotes the correlation between the interim tests statistics $Z_{v,1}$ and the final test statistics $Z_{v,2}$ for $v = p, s$. Formula (7) is an obvious generalization of the standard group-sequential 'unified formulation' described in [3, Chapter 3.1] for two correlated endpoints. Asymptotically, the assumption (7) holds for a wide range of tests statistics, see [3] for details.

As an example, consider the case of a two-armed study comparing an active treatment with placebo with respect to a primary endpoint X_p and a secondary endpoint X_s. For the active treatment, let θ_v and σ_v^2 denote the expectations and variances of the endpoints $v = p, s$, respectively. For placebo, assume that the expected value is 0 for both endpoints and that the variances are the same as for the active treatment. Suppose further that the interim analysis is performed after n_1 and the final analysis after n patients per group using a two-sample Z-test. Then we have in the balanced case $\mathrm{corr}(X_p, X_s) = \mathrm{corr}(Z_p, Z_s) = \rho$, $\delta_v = (\theta_v/\sigma_v)\sqrt{n/2}$, $v = p, s$, and $t_{1,p} = t_{1,s} = n_1/n$.

3. Upper bounds for type I error rates

In this section, we derive sharp upper bounds for the probability of rejecting the secondary hypothesis H_s, when in fact it is true, for the different hierarchical test strategies introduced in Section 2. In Section 3.1 we provide the analytical results. In Section 3.2 we discuss various consequences of these results.

3.1. Analytical results

Let $R_{0,j}(H)$ denote the probability to reject H if it is true. If H_s (but not necessarily H_p) is true, upper bounds for the probabilities $R_{0,j}(H_s)$, $j = a, \ldots, d$, can be derived as follows.

We have

$$\begin{aligned} R_{0,d}(H_s) &= P(Z_{s,1} \geqslant u_{s,1}) + P(Z_{s,1} < u_{s,1}, Z_{s,2} \geqslant u_{s,2}) \\ &= P(Z_{s,1} \geqslant u_{s,1} \text{ or } Z_{s,2} \geqslant u_{s,2}) \\ &= 1 - P(Z_{s,1} < u_{s,1}, Z_{s,2} < u_{s,2}) \\ &= 1 - \Phi_{\sqrt{t_{s,1}}}(u_{s,1}, u_{s,2}) \end{aligned} \tag{8}$$

where $\Phi_\rho(.,.)$ denotes the cdf of the bivariate normal distribution with means 0, variances 1 and correlation ρ between the two variables. Hence, due to (5), $1 - \Phi_{\sqrt{t_{s,1}}}(u_{s,1}, u_{s,2})$ is an upper bound for $R_{0,a}(H_s), \ldots, R_{0,d}(H_s)$, which is exact for $R_{0,d}(H_s)$.

We next show that in some cases the upper bound in (8) is attained even for strategy (a), and because of (5) also for the three other strategies introduced in Section 2. A necessary condition for this is that the probability to reject H_s at the interim analysis is equal to the unconditional probability, i.e. $P(Z_{p,1} \geqslant u_{p,1}, Z_{s,1} \geqslant u_{s,1}) = P(Z_{s,1} \geqslant u_{s,1})$. Under the assumption of multivariate normality of the test statistics given by (7), this requires $Z_{p,1} \geqslant Z_{s,1} + u_{p,1} - u_{s,1}$ with probability 1. Under $H_s: \theta_s = 0$, i.e. $\delta_s = 0$, this

is equivalent to $\rho=1$ and $\delta_p \geqslant (u_{p,1}-u_{s,1})/\sqrt{t_{p,1}} \geqslant 0$. The inequality $u_{p,1}-u_{s,1}\geqslant 0$ means that the error spending function for the secondary endpoint is no less than that for the primary endpoint, i.e. the amount spent for the secondary endpoint is at least as much as for the primary endpoint. This includes, for example, the cases of using the same error spending for both endpoints, or using O'Brien-Fleming boundaries for the primary and Pocock boundaries for the secondary endpoint.

Assume now that $t_{p,1}=t_{s,1}$, as will be the case in most applications. It seems intuitively plausible that $R_{0,a}(H_s)$ is maximized at $\delta_p=(u_{p,1}-u_{s,1})/\sqrt{t_{s,1}}$. This is indeed the case: With this choice of δ_p and for $\rho=1$ we have

$$\begin{aligned} R_{0,a}(H_s) &= P(Z_{p,1}\geqslant u_{p,1}, Z_{s,1}\geqslant u_{s,1}) + P(Z_{p,1}<u_{p,1}, Z_{p,2}\geqslant u_{p,2}, Z_{s,2}\geqslant u_{s,2}) \\ &= P(Z_{s,1}\geqslant u_{s,1}) + P(Z_{s,1}<u_{p,1}-\sqrt{t_{s,1}}\delta_p, Z_{s,2}\geqslant \max(u_{p,2}-\delta_p, u_{s,2})) \\ &= 1-\Phi_{\sqrt{t_{s,1}}}(u_{s,1}, u_{s,2}) \end{aligned} \tag{9}$$

because $u_{p,1}\geqslant u_{s,1}$ implies $u_{p,2}-\delta_p \leqslant u_{p,2}\leqslant u_{s,2}$ for $t_{p,1}=t_{s,1}$, and $Z_{s,1}=Z_{p,1}-\sqrt{t_{s,1}}\delta_p$, $Z_{s,2}=Z_{p,2}-\delta_p$ due to $\rho=1$. Hence, the upper bound $1-\Phi_{\sqrt{t_{s,1}}}(u_{s,1}, u_{s,2})$ from (8) is exact for $R_{0,a}(H_s),\ldots,R_{0,d}(H_s)$ under $\sqrt{t_{s,1}}\delta_p=u_{p,1}-u_{s,1}$ and $\rho=1$.

3.2. Implications

The results obtained in the previous section have a number of interesting consequences.

The naive strategy described in Section 1 and introduced as Strategy 1 in [12] is a special case of the stagewise hierarchical test strategy (a) with $u_{s,1}=u_{s,2}=z_{1-\alpha}$, where z_γ denotes the γ-quantile of the standard normal distribution. If the interim analysis is performed after observing 50 per cent of the patients (i.e. $n_1/n=0.5$), equation (9) yields $R_{0,a}(H_s)=R_{0,d}(H_s)=0.080076$ as the maximum rejection probability for $\alpha=0.05$ when H_s is true. Note that the formulas (8) and (9) do not contain the critical values $u_{p,1}$ and $u_{p,2}$ from the primary error spending function. Hence, the value $\delta_p=(u_{p,1}-u_{s,1})/\sqrt{t_{s,1}}$ at which the largest size inflation occurs with the naive strategy depends on the primary error spending approach, but the magnitude of that inflation does not.

Suppose it is desired to control $R_{0,a}(H_s)$ at level α (=0.05, say). If we set $u_{s,1}=u_{s,2}$ and solve equation (9) for $R_{0,a}(H_s)=0.05$, we obtain $u_{s,1}=1.8755$. This is the $(1-0.03036)$-quantile of the standard normal distribution. Thus, the condition $u_{s,1}=u_{s,2}$ leads to the group-sequential boundaries from Pocock [1]. Furthermore, we conclude from Section 3.1 that (i) this approach is uniformly more powerful than Strategy 3 from [12], which essentially performs a Bonferroni test for the secondary endpoint in case of the stagewise hierarchical strategy (a); (ii) there is no uniform improvement of Pocock's error spending approach for the secondary endpoint; (iii) in terms of error spending, there is no gain in testing the secondary endpoint only once: In spite of $R_{0,a}(H_s)\leqslant\cdots\leqslant R_{0,d}(H_s)$, the maximum rejection probability under H_s is the same and no relaxation of nominal levels is possible when performing the stagewise hierarchical strategy instead of the partially hierarchical strategy.

Applying the error spending approach from Pocock [1] to the secondary endpoint is an optimal allocation of the significance level for the stagewise hierarchical test strategy in a minimax sense: the minimal significance level to be used at an unspecified timepoint is maximized.

4. Power comparisons

In Section 3 we showed that using an error spending approach for the primary endpoint and a separate, possibly different error spending approach for the secondary endpoint exhausts the significance level under weak conditions for the stagewise hierarchical, overall hierarchical and partially hierarchical test strategies. It is obvious that the error spending approaches do not dominate each other in terms of power, although special results are available in certain situations. For example, we have seen that using the error spending approach from [1] is optimal in a minimax sense for the stagewise hierarchical strategy. Similarly, in the borderline case of $Z_{p,i}=Z_{s,i}$ for all i, using the same error spending function for both the primary and the secondary endpoint is always better than using different error spending functions.

In this section we report selected results of an extensive numerical study to investigate the power of rejecting H_s for various combinations of error spending functions when applying the stagewise hierarchical strategy (a). Note that the power to reject H_s for the strategies (b) through (d) is very similar to the power in ordinary group-sequential trials, which has been investigated extensively elsewhere (for example, in [3]) and is thus not reported here.

We assume normally distributed test statistics following the distribution (7). The power $R_j(H_s)$ for any of the strategies $j=a,b,c,d$ described in Section 2 depends on the non-centrality parameters δ_p and δ_s, the correlation ρ, the information fractions $t_{p,1}$ and $t_{s,1}$ and the selected error spending approaches. Note that the power calculations can be done analytically and simulations are not necessary. For example, $R_a(H_s)$ is the sum of a two- and a three-dimensional normal integral, which can be calculated using the numerical integration methods from [15, 16].

As mentioned in the Introduction, the choice of the spending function for the primary variable is not primarily guided by statistical power considerations but rather by the wish to stop the trial early only if there is very strong evidence in favor of the new treatment. Hence, in practice spending functions allocating rather small significance levels at the interim analysis (O'Brien-Fleming type) are much more popular than balanced spending functions (Pocock type). Consequently, we use O'Brien-Fleming boundaries [2] for the primary endpoint analysis. For the secondary endpoint, however, the situation is different. When stopping of the trial depends on the primary endpoint, power considerations can be the guiding principle for the choice of the spending function for the secondary endpoint.

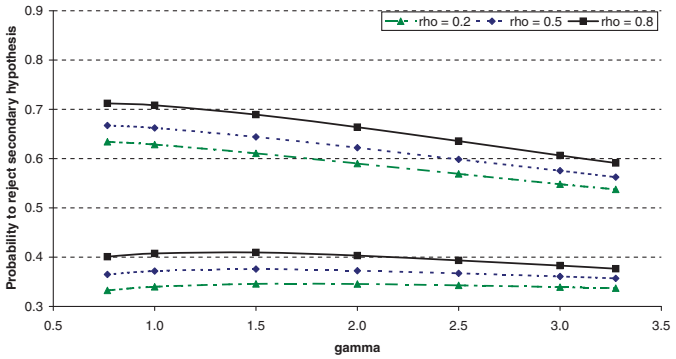


Figure 1. Power to reject the secondary hypothesis H_s with the stagewise hierarchical testing procedure (a) for different values of ρ. Top three curves: $\delta_p = 4, \delta_s = 3$. Bottom three curves: $\delta_p = 3, \delta_s = 2$.

For the power calculations below we set $t_{p,1} = t_{s,1} = \frac{1}{2}$. For the secondary endpoint analysis we used the error spending function

$$a(t) = \min(\alpha \cdot t^{\gamma}, \alpha)$$

from Kim and DeMets [14], where γ is a tuning parameter. In case of just one interim analysis, γ can be calculated such that the resulting critical values $u_{s,1}$ and $u_{s,2}$ match the suggestions from Pocock ($\gamma = 0.7668$) and O'Brien and Fleming ($\gamma = 3.2749$). Values of γ between these two numbers provide error spending functions which spend more of the significance level at the interim analysis than O'Brien-Fleming, but less than Pocock. For the comparisons below, we considered all combinations of $\rho = 0, 0.1, \ldots, 0.9, 1$, $\delta_p = 2, 3, 4, 5$, $\delta_s = 1, 2, 3, 4, 5$, and $\gamma = 0.7668, 1, 1.5, 2, 2.5, 3, 3.2749$.

In Figures 1 and 2 we present selected results of the numerical power study. In each graph we plotted the power to reject the secondary hypothesis H_s against the tuning parameter γ for the stagewise hierarchical strategy (a) and for selected values of δ_p, δ_s, and ρ. The results for the Pocock error spending function are on the left end of the power curves and that of the O'Brien-Fleming error spending function on the right end. The graphs are selected to reflect study sample sizes that are usually used to achieve a reasonably high power for rejecting the primary hypothesis. Note that the conclusions below also hold for the scenarios not plotted here.

The results from the power study seem to indicate that error spending approaches that spend a larger amount at the interim analysis are in general more powerful. For example, choosing Pocock's error spending function has a power advantage over O'Brien-Fleming's error spending function in the scenarios of Figure 1, where the magnitude of the power advantage depends on the magnitude of δ_p and δ_s. Even in cases where the O'Brien-Fleming error spending function is known to be nearly optimal (high correlation, similar effect), the power gain is small, if any. To investigate this effect in more detail, we calculated the power when choosing O'Brien-Fleming's approach under its optimal conditions ($\rho = 1$ and $\delta_p = \delta_s$) and under small deviations thereof. It becomes evident from Figure 2 how quickly the power advantage is lost as either the correlation drops or the mean effects differ. Note that in case of $\delta_s > \delta_p$, the secondary hypothesis is rejected with (conditional) probability 1 if the primary hypothesis is rejected, the correlation between primary and secondary test statistic is 1, and the error spending approach is such that more Type I error is spent for the secondary than for the primary endpoint. This is the reason for two completely flat curves in Figure 2.

The results are in line with [12], who concluded that using the O'Brien-Fleming error spending for both the primary and secondary endpoint is often less powerful than a Bonferroni test for the secondary endpoint. As mentioned in Section 3.2, using Pocock's approach instead of the Bonferroni test is even more powerful. From the results of this power study, we further conclude that the Pocock error spending approach generally performs well when compared with other error spending approaches (i.e. other values of γ).

5. Numerical example

In this section we use a numerical example based on a real case study to illustrate the results of the different hierarchical test strategies introduced in Section 2. In a controlled clinical trial comparing a new respiratory drug with an active control for the chronic obstructive pulmonary disease (COPD), the primary endpoint was the difference between the treatments in standardized area under the curve (AUC) of the forced expiratory volume in one second (FEV_1) after 12 weeks of treatment. The difference in trough FEV_1 was considered as a key secondary endpoint.

An interim analysis was planned after 50 per cent of patients had completed the study with the option to stop early for efficacy. One-sided tests of superiority for the new treatment were considered for both hypotheses. The standardized test statistics $Z_{v,i}$, $i = 1, 2$, for the two hypotheses H_p and H_s can be considered as asymptotically $N(0, 1)$-distributed. It was planned to control

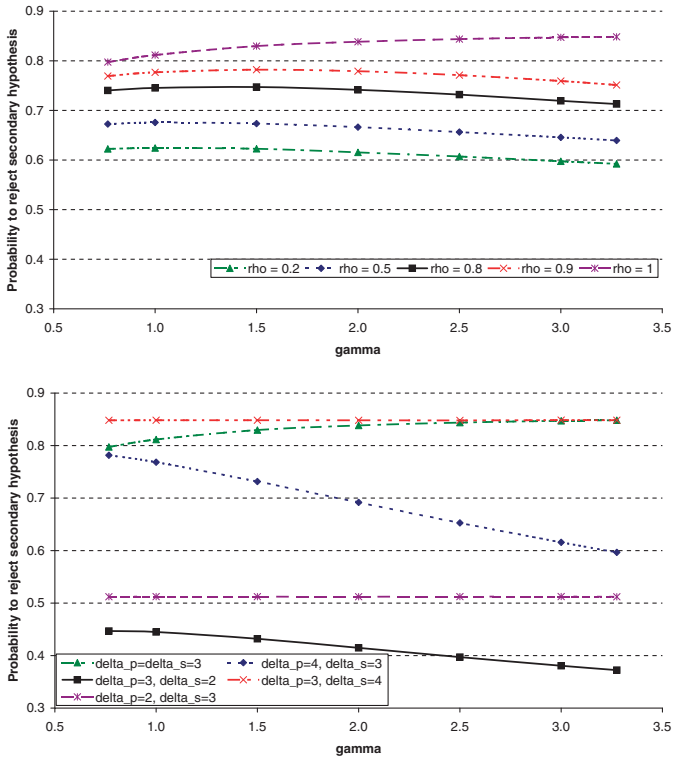


Figure 2. Power to reject the secondary hypothesis H_s with the stagewise hierarchical testing procedure under different scenarios. Top: $\delta_p=\delta_s=3$ and $\rho=0.2, 0.5, 0.8, 0.9, 1$. Bottom: $\rho=1$ and different values for δ_p, δ_s.

the FWER at level $\alpha=0.025$. The study sample size was determined such that a simple Z-test without any adjustments would yield approximately 92 per cent power for the primary and 86 per cent power for the secondary endpoint. This resulted in the recruitment of 542 patients per treatment arm. Note that a power of 92 per cent corresponds to a non-centrality parameter of $\delta_p=\Phi^{-1}(0.975)-\Phi^{-1}(1-0.92)=3.37$ for AUC and $\delta_s=3.04$ for trough FEV_1 at the final analysis. Note also that in the actual trial covariate adjusted analyses were conducted. As these details are not important for the discussion below, we continue with the simplified analyses.

Suppose now that O'Brien-Fleming boundaries were used for the primary analysis and Pocock boundaries for the secondary analysis. We would conclude sufficient evidence of a benefit in AUC, if at the interim analysis $Z_{p,1}\geqslant 2.80$, or if at the final analysis $Z_{p,2}\geqslant 1.98$. As noticed before, these rejection boundaries for the primary hypothesis H_p hold for all four hierarchical test strategies discussed in Section 2. For the secondary endpoint, the significance boundary is 2.18 at both the interim and final analyses. The rejection of the secondary hypothesis H_s of course depends on the selected hierarchical test strategy from Section 2. To illustrate the different strategies, we discuss them below under two scenarios: (i) the situation of stochastic independence between the primary and secondary endpoint and (ii) the case where the correlation between primary and secondary endpoint is 0.8. Furthermore, we make the simplifying assumption that the conditions, under which the power calculations were conducted, are correct and correspond to the true differences between the new treatment and the active control.

(*a*) *Stagewise hierarchical*: With this strategy, the trial is stopped as soon as the primary endpoint is significant. This strategy would be applied if superior efficacy in the primary endpoint alone is important enough to stop the trial (and apply for market approval of the new drug immediately, say). Efficacy of the secondary endpoint would be of substantial additional benefit (for example, for an extended label claim). The probability of rejecting the primary hypothesis H_p at interim is $P(Z_{p,1}\geqslant 2.80)=1-\Phi(2.80-\sqrt{0.5}\cdot 3.37)=33.8\%$. Conditionally on rejecting H_p, the probability of success in the secondary endpoint is $P(Z_{s,1}\geqslant 2.18)=1-\Phi(2.18-\sqrt{0.5}\cdot 3.04)=48.9\%$ in the independent case. In the correlated case, this probability increases to $P(Z_{s,1}\geqslant 2.18|Z_{p,1}\geqslant 2.80)=P(Z_{p,1}\geqslant 2.80, Z_{s,1}\geqslant 2.18)/P(Z_{p,1}\geqslant 2.80)=87.7\%$. If the trial continues to the end and is significant in the primary endpoint, the conditional power of the secondary endpoint is $P(Z_{s,2}\geqslant 2.17)=1-\Phi(2.17-3.04)=80.6\%$ in the independent case. In the correlated case, this value is almost identical: $P(Z_{s,2}\geqslant 2.17|Z_{p,1}<2.80, Z_{p,2}\geqslant 1.98)=80.7\%$. Overall, the power for the secondary endpoint given by formula (1) is 63.3 per cent in the independent case and 76.5 per cent when the correlation between AUC and trough FEV_1 is 0.8.

(*b*) *Overall hierarchical*: Here, the trial would continue even after significance of the primary endpoint at the interim analysis. This would be done if a formal claim of efficacy in trough FEV_1 is important enough to warrant continuation of the trial. In this case, the overall probability of rejection for the secondary endpoint would be 74.6 per cent (see equation (2)), as opposed to 63.3 per cent for strategy (a) in the independent case. For the correlated case, the corresponding power values are 79.8 per cent and 76.5 per cent, respectively. There is a possibility here that a significance for the primary endpoint at interim is 'lost' when AUC is re-analyzed at the final analysis. The probability of this event, however, is remote (0.09 per cent in the scenario considered here). If it occurred, it may be an indication that the two trial stages are not homogeneous.

(*c*) *Partially hierarchical*: This strategy would allow a final test for the secondary hypothesis H_s even if the primary hypothesis H_p is not rejected at all. The power for the secondary endpoint increases to 81.2 per cent in the independent case and 81.3 per cent in the correlated case, see formula (3). By construction, the FWER for both AUC and trough FEV_1 would not be controlled by this approach. This strategy would probably not be appropriate for the case study considered here.

(*d*) *Co-equal*: The difference between this strategy and strategy (c) is that H_s is additionally rejected in the following case: H_p is not rejected at interim, H_s is rejected at interim, but H_s would not have been rejected at the final analysis. Equation (4) gives a power of 82.3 per cent for trough FEV_1. Of course, there is no difference in this power between the correlated and the uncorrelated case here. This strategy may be appropriate if, for example, it is clear from the onset that the trial will be continued to the final analysis for exploratory reasons, and trough FEV_1 is just one of many secondary endpoints that are investigated without strong FWER control.

An interesting side aspect of this trial was that the clinical team also looked into the possibility of conducting an interim analysis after a fixed number of patients had completed 4 weeks of treatment to decide upon stopping or continuing recruitment. The primary (AUC FEV_1 after 12 weeks) and secondary (through FEV_1 after 12 weeks) endpoints would remain the same, but the 4-week readout of AUC would be used to decide whether the study should continue to its second stage or not. In either case, monitoring of all enrolled patients would be continued until completing 12 weeks of treatment.

One way of applying a group-sequential design while controlling the FWER strongly at level α is as follows. Assume that if recruitment continues, the number of patients available at the final analysis (12 weeks of treatment) is n, whereas if the trial stops early, this number is $n_1 < n$. Mathematically, this is equivalent to a design where an interim analysis based on the primary 12-week endpoint is conducted at an information fraction of $t_1 = n_1/n$. Applying this strategy to our case study, we may use, for example, Pocock's boundaries $u_{p,i}$, $i = 1, 2$, for the primary endpoint, based on the information fraction n_1/n. If recruitment is stopped, monitoring continues for the first-stage cohort of patients to collect the complete 12-week data and the final analysis of the n_1 patients is performed with the Pocock boundary $u_{p,1}$. Otherwise, the trial continues and we are allowed to perform the final analysis of the n patients based on the Pocock boundary $u_{p,2}$. (Note that in this case we theoretically can also perform an additional formal interim analysis for the primary endpoint when n_1 patients have completed 12 weeks of treatment. In practice, however, this option would probably not be used, as the benefit of such a late interim analysis will be outweighed by its operational cost, when n patients have already been recruited anyway.) For the secondary endpoint we would apply the stagewise hierarchical test strategy (a). That is, the secondary hypothesis H_s is tested at its own group-sequential boundaries $u_{s,i}$, $i = 1, 2$ which are also calculated based on the information fraction n_1/n. Rejection of the primary hypothesis H_p decides if and when the secondary test is done.

In our case study, one option was to perform the interim analysis after 80 per cent of the patients had completed 4 weeks of treatment. In that case, a recruitment stop would have reduced the sample size only moderately, but the operational advantages might have been worthwhile (being able to tell centers early when to stop recruitment, when to start data cleaning, etc.). If $n_1/n = 0.8$, the critical value for both analyses (with either n_1 or n patients) is 2.11. Owing to the high correlation between the two test statistics, the power loss relative to a trial without the option to adjust for a recruitment stop is rather small. In the scenarios discussed here, even under an early recruitment stop a lower bound of the power is $1 - \Phi(2.11 - \sqrt{0.8} \cdot 3.37) = 82\%$ and $1 - \Phi(2.11 - \sqrt{0.8} \cdot 3.04) = 73\%$ for AUC and trough FEV_1, respectively. This calculation provides a lower bound, because it simply gives the probability of significance for separate interim analyses at 80 per cent of the information, ignoring the decision rule for stopping recruitment and the correlation between 4- and 12-week response.

6. Discussion

Hierarchical testing of a primary and a secondary hypothesis in a group-sequential setting gives rise to interesting questions (and surprising answers) beyond the usual considerations regarding Type I error rate control and power. It follows from Tang and Geller [6] that hierarchical group-sequential testing with separate error spending functions at level α for the primary and secondary hypotheses protects the FWER strongly at level α. For this to be true it is only necessary that the secondary hypothesis H_s is tested whenever the primary hypothesis H_p is rejected. That is, if H_p has been rejected at an interim analysis, H_s can be tested on partial data at that interim analysis as well and again at the final analysis, if it was not significant before. We call this the *overall hierarchical* testing strategy. However, in practice the more restrictive *stagewise hierarchical* testing strategy is often applied, where the trial is stopped early if H_p can be rejected at the interim analysis, irrespective of whether the secondary hypothesis can be rejected or not.

This raises two questions. First, can the rejection boundaries for the stagewise hierarchical testing strategy be relaxed compared with the one based on a separate error spending approach for each hypothesis? Second, can the secondary hypotheses be tested at full level α, either at interim or at the final analysis, since H_s is tested only once in the stagewise hierarchical testing

strategy? The answer to both questions is negative. In this paper, we show that if the nominal rejection levels for H_s are assumed to be the same for the interim and final analysis, the nominal significance levels given by Pocock's error spending function (which are smaller than α) are sharp upper bounds. Nevertheless these levels are larger than $\alpha/2$ obtained from the Bonferroni inequality, which has been proposed previously to protect the Type I error rate in this situation. We also show that in general the boundaries given by any error spending approach for the secondary hypothesis cannot be relaxed simultaneously. We provide cases of parameter values for the joint distribution of primary and secondary test statistics which lead to full exhaustion of the significance level α, even for the most restrictive of the four test strategies considered, the stagewise hierarchical strategy.

The choice of the error spending function for the primary hypothesis as well as the timing of the interim analyses and the maximum sample size are governed by the same criteria as in any clinical trial with a single hypothesis (overall power, expected sample size, minimum number of patients required for safety assessment, etc.). However, the situation is more complex for the secondary hypothesis H_s, as it depends on the error spending function chosen for the primary hypothesis H_p and the selected stopping strategy. For the overall hierarchical strategy, there is no fundamental difference between the criteria governing the choice of the error spending function for the primary and secondary hypothesis. Hence, the same error spending functions may be used for them, or a larger significance level will be used for H_s at the interim analysis in order to increase the chance of stopping the trial early. For the stagewise hierarchical strategy this choice, however, is more important since if H_p is rejected at interim, but H_s is not, there is no second chance to test H_s again.

The power study from Section 4 confirmed and quantified this result for the case of an O'Brien-Fleming spending function for H_p. We show that in this situation the choice of Pocock's spending function for H_s is a good choice, because in most cases it leads to a more powerful procedure compared to other choices of error spending functions for H_s.

In this paper we restrict the discussion to hierarchical tests of one primary and one secondary endpoint in a two-stage design. It is of interest to consider also group-sequential trials with more than one interim analysis. Here, both the stagewise and the overall hierarchical strategy with any choice of error spending function at level α for the primary and the secondary hypothesis will control the FWER at level α. In this situation, however, the probability boundaries for the stagewise hierarchical approach are not sharp anymore. It is an open question whether more powerful procedures can be derived.

Tamhane *et al.* [17] have independently worked on problems similar to the ones discussed in this paper and came to similar conclusions and recommendations with respect to the stagewise hierarchical approach. They as well as the authors of this paper presented their preliminary results—though with different focus and derivations—at the 6th International Conference on Multiple Comparison Procedures in Tokyo (2009). Details of their derivations and results are presented in [17].

The situation of testing one primary and one secondary endpoint repeatedly and hierarchically in a group-sequential design shares some common features with the situation of comparing two treatments with a control on a primary and a secondary endpoint in a fixed-sample trial. Related test procedures have been critically reviewed by Hung and Wang [18] and further discussed with respect to error rates of the secondary hypotheses by Bretz *et al.* [19]. The evaluation of the sharpness of the rejection boundaries and of power is subject of ongoing research.

Acknowledgements

We thank Roger Owen, Karen Thomas and Michelle Henley for providing us the case study and for continuing discussions during the progress of this work. We are also grateful to the anonymous referees for their helpful comments, which improved the presentation of the paper.

References

1. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
2. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **5**:549–556.
3. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall, CRC: Boca Raton, 2000.
4. Proschan MA, Lan KKG, Wittes JT. *Monitoring of Clinical Trials: A Unified Approach*. Springer: New York, 2006.
5. Follmann DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**:325–336.
6. Tang DI, Geller NL. Closed testing procedures for group sequential clinical trials with multiple endpoints. *Biometrics* 1994; **55**:1188–1192.
7. Hellmich M. Monitoring clinical trials with multiple arms. *Biometrics* 2001; **57**:892–898.
8. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
9. Maurer W, Hothorn L, Lehmacher W. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der chemisch-pharmazeutischen Industrie*, Vollmar J (ed.). Fischer Verlag: Stuttgart, 1995; 3–18.
10. Westfall PH, Krishen A. Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.
11. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
12. Hung HMJ, Wang SJ, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics* 2007; **17**:1201–1210.
13. Marcus R, Peritz E, Gabriel KR. On closed testing procedure with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
14. Kim K, DeMets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* 1987; **74**:149–154.
15. Genz A, Bretz F. Methods for the computation of multivariate *t*-probabilities. *Journal of Computational and Graphical Statistics* 2002; **11**:950–971.

16. Genz A, Bretz F. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Computer Science, vol. 195. Springer: Heidelberg, 2009.
17. Tamhane A, Mehta C, Liu L. Testing a primary and a secondary endpoint in a group-sequential design. *Biometrics* 2009; submitted.
18. Hung HMJ, Wang SJ. Some controversial multiple testing problems in regulatory applications. *Journal of Biopharmaceutical Statistics* 2009; **19**:1–11.
19. Bretz F, Maurer W, Gallo P. Discussion of 'Some controversial multiple testing problems in regulatory applications' by H. M. J. Hung and S.-J. Wang. *Journal of Biopharmaceutical Statistics* 2009; **19**:25–34.

Research Article

Statistics in Medicine

Received 15 June 2010, Accepted 17 June 2011 Published online 5 September 2011 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4342

Time-to-event analysis with treatment arm selection at interim

L. Di Scala*† and E. Glimm

This paper discusses the application of an adaptive design for treatment arm selection in an oncology trial, with survival as the primary endpoint and disease progression as a key secondary endpoint. We carried out treatment arm selection at an interim analysis by using Bayesian predictive power combining evidence from the two endpoints. At the final analysis, we carried out a frequentist statistical test of efficacy on the survival endpoint. We investigated several approaches (Bonferroni approach, 'Dunnett-like' approach, a conditional error function approach and a combination p-value approach) with respect to their power and the precise conditions under which type I error control is attained. Copyright © 2011 John Wiley & Sons, Ltd.

Keywords: adaptive design; Bayesian statistics; oncology; treatment selection

1. Introduction

This paper discusses an adaptive oncology phase II/III trial with time-to-event (TTE) analysis for survival. The trial comprises an interim analysis for selecting one of two originally tested treatment regimens. We suggest an approach that combines interim decision making based on Bayesian predictive power with adjustment methods that control the type I error rate of the final efficacy decision (selected treatment versus control) at the prespecified level α. The approach allows for the selection of one treatment at the interim, on the basis of efficacy results or other, external considerations, a stop for futility and/or efficacy at interim or continuation with both treatment arms. Decision making at the interim analysis is improved by incorporating information on progression-free survival (PFS) as an additional TTE endpoint.

We introduce the trial in Section 2. Section 3 reviews the basics of the group-sequential log-rank test. Subsequently, Section 4 discusses the interim decision-making methodology, and Section 5 introduces several adjustment methods used for type I error control. We investigated various combinations of these methods and versions of the decision-making rules in extensive simulation studies. We summarize the results in Section 6, followed by the final discussion in Section 7.

2. The trial

The trial investigates a targeted therapy in the advanced lung cancer setting, administered as an oral agent. Phase I dose-finding studies had established acceptable safety for two different doses/regimens: daily and weekly administration. The daily regimen is assumed to guarantee constant inhibition of markers downstream of the therapeutic target. The weekly schedule, however, might show a better safety profile while still delivering sufficient marker inhibition.

As development approaches phase II planning, the possibility of an adaptive phase II/III study is explored, with the aim of embedding treatment selection into a confirmatory trial for testing the efficacy of the combination of the targeted therapy plus the standard of care (SoC; chemotherapy in this context) versus SoC alone. In oncology, phase II designs have historically been single-arm, multistage studies

Novartis Pharma AG, Basel, Switzerland
**Correspondence to: L. Di Scala, Novartis Pharma AG, Basel, Switzerland.*
†E-mail: lilla.di_scala@novartis.com

with tumor shrinkage as the primary endpoint. These designs are inappropriate in the combination setting, as the presence of the combination partner as active comparator is crucial in properly quantifying treatment benefit. In addition, tumor shrinkage is not an appropriate indicator of treatment benefit of a targeted therapy such as the one investigated here, as the agent would generally be expected to delay tumor growth rather than lead to its disappearance. A TTE endpoint such as PFS, that is, time from start of therapy to tumor progression or death due to any cause (see RECIST guidelines [1]), is often selected as the phase II endpoint for targeted agents. In addition, regulatory agencies recommend that approval should generally be sought by testing efficacy in overall survival (OS) as this is the 'universally accepted direct measure of benefit' [2], whereas PFS is based on tumor measurements and is thus subjective.

Table I. Expected PFS and OS in advanced NSCLC.

Indication	SoC PFS (months)	SoC OS (months)	PFS benefit (%)	OS benefit (%)
First line NSCLC	3.5–5	7.5–10	20–50	20–50

Note: NSCLC, non-small cell lung cancer; OS, overall survival; PFS, progression-free survival; SoC, standard of care.

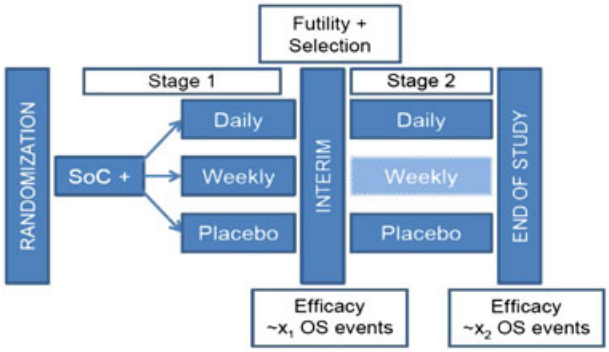


Figure 1. Design schema.

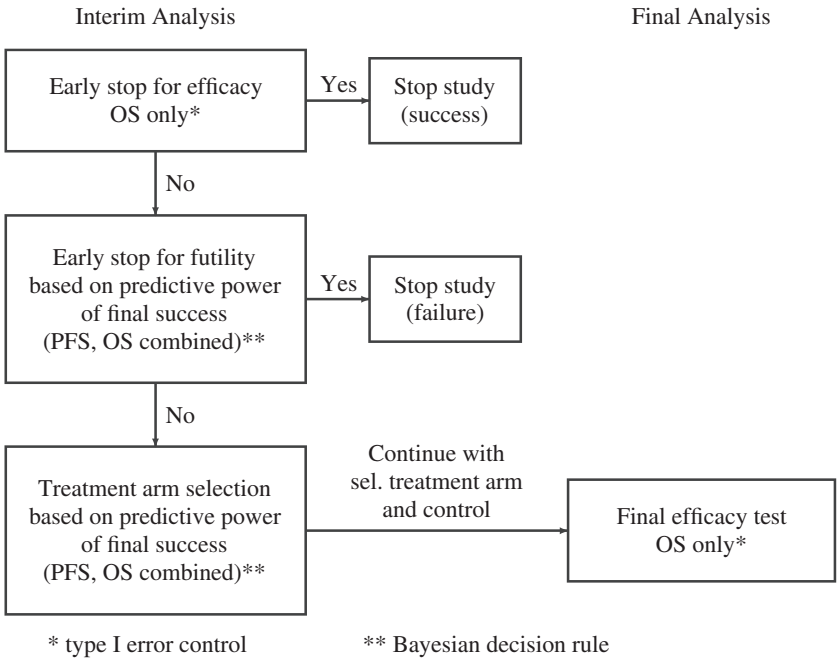


Figure 2. Interim decision flow chart.

Therefore, the proposed trial has OS as the primary endpoint and PFS as a key secondary endpoint. The trial is *event driven*; that is, both interim and final analysis would be carried out after prespecified numbers of observed OS events.

Table I provides an overview of the PFS/OS expectations for non-small cell lung cancer patients with the currently available therapies. The gap between progression and survival expectations would make a phase II trial with an OS endpoint operationally unfeasible and significantly delay development timelines. Therefore, a design that includes

- a phase II part with treatment selection and confirmation of the selected treatment's efficacy and
- a PFS endpoint that would shorten time to treatment selection but an OS endpoint in phase III

was considered by the trial team as a viable option to investigate. A schema is provided in Figure 1, and Figure 2 shows in detail how interim decisions are made.

Regarding the validity of PFS as a surrogate endpoint for survival, it is widely professed in the clinical literature that a delay of progression by treatment corresponds to a simultaneous lengthening of time to death [3]. Furthermore, death is regarded as a progression event by definition. Hence, both endpoints will be influenced by the treatment in a similar way; that is, a prolongation of PFS with a corresponding prolongation of post-progression survival is likely (the simulations performed in Section 6 also explore scenarios where PFS is not a surrogate of survival).

3. Group-sequential log-rank test

This section lays the foundation of subsequent investigations by outlining the joint distribution of a group-sequential log-rank test that is performed on two treatment regimens. Assume that a one-sided test of treatment effect rejects if the value of the log-rank test statistic is 'large'.

Let t_k^* be the time at which an event has occurred in group $j = 1, 2, C$ where 1 and 2 denote the treatments and C the control group. For simplicity, let us assume that all event times are distinct such that

$$\delta_{kj} = \begin{cases} 1 & \text{if an event occurred in group } j \text{ at time } t_k^* \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

Let $t_i, i = 1, 2$ denote the time of interim and final analysis and d_i the number of uncensored events that occurred until t_i. The statistic of the log-rank test for a single treatment $j = 1, 2$ at time t_i is given by

$$l_{ij} = \frac{\sum_{k=1}^{d_i} (\delta_{kj} - p_{kj})}{\sqrt{\sum_{k=1}^{d_i} p_{kj}(1 - p_{kj})}} \tag{2}$$

where $p_{kj} = \frac{r_{kj}}{r_k}$, r_{kj} is the number of patients at risk in group j at time t_k^* and r_k is the total number of patients at risk at t_k^*.

We assume that the two test statistics l_{i1} and l_{i2} are treated separately in the sense that the total number of patients at risk is set to $r_k = r_{kj} + r_{kC}$ at time t_k^* for the comparison of groups C and j and that summation is only over those event times t_k^* where the event occurred either in C or in j. For convenience of notation, we define

$$p_{kj} = \begin{cases} \frac{r_{kj}}{r_{kj} + r_{kC}} & \text{if the event occurred in group } j \text{ or } C \text{ at time } t_k^* \\ 0 & \text{otherwise} \end{cases}$$

such that summation in (2) can be carried out over all event times. If the survival distributions in the three groups are identical and groups are equal in size, the expected value of p_{kj} is 1/2 in this separate analysis.

The asymptotic joint distribution of $\binom{l_{i1}}{l_{i2}}$ is

$$N\left(\begin{pmatrix} i_{i1}\theta_1 \\ i_{i2}\theta_2 \end{pmatrix}, \begin{pmatrix} 1 & v_{i1,i2} \\ v_{i1,i2} & 1 \end{pmatrix}\right) \tag{3}$$

with θ_j being the log hazard ratio between treatment j and control ($\theta_1 = \theta_2 = 0$ under the null hypothesis H_0), $i_{ij} = \sqrt{\sum_{k=1}^{d_i} p_{kj}(1-p_{kj})}$,

$$v_{i1,i2} = \frac{\sum_{k=1}^{d_i} \phi_k}{\sqrt{\sum_{k=1}^{d_i} p_{k1}(1-p_{k1}) \cdot \sum_{k=1}^{d_i} p_{k2}(1-p_{k2})}} \tag{4}$$

where the correlation is derived from the fact that conditional on $\{r_{kj}\}_{j=1,2,C}$, we have

$$\phi_k := \mathrm{Cov}(\delta_{k1}, \delta_{k2}) = E\left((\delta_{k1} - p_{k1})(\delta_{k2} - p_{k2})\right) = p_{k1}p_{k2} \cdot \frac{r_{kC}}{r_{k1} + r_{k2} + r_{kC}}.$$

From this, it is easy to write down the joint asymptotic distribution of $(l_{11}, l_{12}, l_{21}, l_{22})$'. However, in the TTE context, it is more convenient to use the so-called independent increment structure: let

$$\tilde{l}_{2j} = \frac{i_{2j}l_{2j} - i_{1j}l_{1j}}{\sqrt{i_{2j}^2 - i_{1j}^2}}. \tag{5}$$

Then it follows from the standard theory for linear transformations of the multivariate normal distribution that asymptotically,

$$\begin{pmatrix} l_{11} \\ l_{12} \\ \tilde{l}_{21} \\ \tilde{l}_{22} \end{pmatrix} \sim N\left(\begin{pmatrix} i_{11}\theta_1 \\ i_{12}\theta_2 \\ \sqrt{i_{21}^2 - i_{11}^2}\theta_1 \\ \sqrt{i_{22}^2 - i_{12}^2}\theta_2 \end{pmatrix}, \begin{pmatrix} 1 & v_{11,12} & 0 & 0 \\ v_{11,12} & 1 & 0 & 0 \\ 0 & 0 & 1 & v^*_{21,22} \\ 0 & 0 & v^*_{21,22} & 1 \end{pmatrix}\right) \tag{6}$$

with $v^*_{21,22} = \frac{\sum_{k=1}^{d_2}\phi_k - \sum_{k=1}^{d_1}\phi_k}{\sqrt{i_{21}^2 - i_{11}^2}\cdot\sqrt{i_{22}^2 - i_{12}^2}}$.

In contrast to the analysis of non-time-dependent data such as normal endpoints or rates, the statistics l_{1j} and $\tilde{l}_{2j}$ are not simply the 'stage-1' and 'stage-2' test statistics, respectively, calculated from observations before and after the interim. It can be shown that a naive calculation of test statistics just splitting into events before and after the interim leads to an invalid analysis. We will discuss this aspect in Section 5.

Without treatment arm selection, a group-sequential approach with the possibility to stop early for efficacy if at least one treatment is statistically significantly superior to control could be implemented as follows. Critical values c_1 and c_2 have to be determined in such a way that

$$Pr_{H_0}(\max(l_{11}, l_{12}) \geqslant c_1) = \alpha_1 \text{ and} \tag{7}$$

$$Pr_{H_0}(\max(l_{11}, l_{12}) < c_1, \max(l_{21}, l_{22}) \geqslant c_2) = \alpha - \alpha_1 \tag{8}$$

where α is the overall test level and $\alpha_1 < \alpha$ is the 'level spent' at the interim. It is easy to obtain c_1 and c_2 from (3) and (6) by numerical integration, for example, using the methods described in [4].

The concrete value of α_1 is determined as some function of the *information fraction* $\frac{I_{1j}}{I_{2j}}$ where $I_{ij} = i_{ij}^2$. In later sections, we use the α-spending approach by Lan and DeMets [5]. In *event-driven* trials, I_{2j} is usually approximated by the total number of events to be collected divided by 4 (which is the expected value of I_{2j} under the null hypothesis of equal event rates in all treatments). The interim α_1 is then calculated from the ratio of actually observed events to this estimate of I_{2j}. The final critical value c_2 is re-calculated after all data were obtained such that (8) is met (see, e.g. [6, Chapter 3.7]). Note that these calculations are carried out based on the overall events observed, assuming H_0, not on the actually observed information fractions, which depend also on how all events split onto the treatment groups. Here, we also follow this general idea. However, treatment selection requires some modifications of the general approach. These are detailed in Section 5.

Only the so-called *non-binding* futility bounds are considered in this paper. The 'α reclaim', which is a consequence of binding futility bounds, is mathematically easy to handle but generally discouraged by health authorities.

The generalization to a stratified log-rank test is straightforward but not discussed here.

4. Interim decision making

The aim of this paper was to modify the group-sequential approach outlined in Section 3 to accommodate the three possible interim decisions as shown in Figure 2, that is, stop for efficacy, stop for futility and selection of a treatment arm to continue into stage 2 of the trial. The control arm is continued to the end of the trial and not subject to any interim decision.

The stop for efficacy will be based on a conventional group-sequential α-spending rule for the primary OS endpoint. For the other decisions, it is desired to exploit information available from a surrogate endpoint, that is, PFS, as an indicator of subsequently increased risk of death.

4.1. *Early stop for efficacy*

Because efficacy claims are ultimately based on benefit in OS, the decision to stop the trial early for efficacy can only be based on OS information. Following common practice in event-driven TTE trials, an O'Brien–Fleming type Lan–DeMets α-spending approach is used (see, e.g. [6, p. 148, formula (7.3)]): the trial is stopped early if the log-rank test at interim is significant at level α_1, which will be calculated from the function

$$\alpha(t) = 2 - 2\Phi\left(\frac{u_{1-\alpha/2}}{\sqrt{t}}\right)$$

at t equal to the expected value of the accrued information ratio I_{1S}/I_{2S} where S denotes the selected treatment, $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution and u_α its α-quantile. Early efficacy is claimed if $\max(l_{11}, l_{12}) > c_1$ where c_1 is the critical value defined in (7). Alternatively, it would also be possible to simply fix an $\alpha_1 < \alpha$ as deemed appropriate.

The expected value of I_{1S}/I_{2S} is approximated by

$$\frac{\frac{2}{3}d_1^*}{d_2^* - \frac{d_1^*}{3}} \tag{9}$$

where d_i^* is the total number of events after stage i. The trial will be planned for fixed d_2^* and d_1^*. Hence, (9) is the ratio of expected number of events under H_0 in the selected treatment arm and the control arm. Other ways of approximating the information fraction are conceivable. The approximation of the information fraction is only used to calculate α_1 according to the selected α-spending approach. It is not used to calculate $\tilde{l}_{2S}$, which is carried out via formula (5).

Simulations indicate that the impact of different ways of approximating the information fraction under H_0 on both α-level control and power is negligible. This is unsurprising, as the various estimates of information fraction are very similar, vary in a rather narrow range and only marginally influence critical values in this range.

4.2. *Hybrid Bayesian framework for interim decision making*

An early stop for efficacy can only be justified with clear-cut evidence from the primary endpoint. Substituting lacking evidence on OS with evidence from PFS is risky in the context of a submission-relevant trial and requires assumptions on the link between the two. In contrast, stopping for futility or treatment arm selection does not directly affect efficacy claims. Hence, it is desired to improve decision making at interim on these two aspects by 'borrowing strength' from observed PFS events.

To achieve this, Bayesian tools are suggested. In particular, Bayesian predictive probabilities will be calculated for both the futility decision and the treatment arm selection.

4.2.1. Predictive power calculation and its use. In a Bayesian framework, the log-hazard ratios θ_{ij} are considered as random variables with a prior distribution. The location of their prior distribution reflects the knowledge about the expected magnitude of the treatment effect, whereas the covariance reflects the (un)certainty expressed regarding such prior knowledge. Bayesian inference of the treatment effects is based on the posterior distribution. This is defined as the conditional distribution of the θ_{ij}'s given the observations coming (in this setting) from stage 1 of the trial.

Because of the confirmatory nature of the trial, a traditional statistical test of treatment effect is carried out. Thus, the suggested hybrid Bayesian/frequentist approach is not based on direct inference using the

posterior distribution of θ_{ij}'s but rather on the predictive distribution of their estimates $\hat{\theta}_{ij}$. The predictive distribution relates the estimates $\hat{\theta}_{ij}$ to the posterior and is obtained by integrating the conditional distributions $\hat{\theta}_{ij}$ given θ_{ij} by the posterior distribution of θ_{ij}. Thus, one obtains the distribution of new estimates $\hat{\theta}_{ij}$ when both prior information on θ_{ij} and old estimates (in this case, through stage 1) are available. Brannath *et al.* [7] used a similar approach in the context of confirmatory clinical trials. For more information on this methodology see, e.g., [8] or [9].

In the context of the trial discussed here, the approach works as follows: $\hat{\theta}_{1j} = \frac{l_{1j}}{i_{1j}}$ and $\hat{\theta}_{2j} = \frac{\tilde{l}_{2j}}{\sqrt{I_{2j}-I_{1j}}}$ are estimates of the log-hazard ratios for one of the considered endpoints (either OS or PFS). According to formula (3), their asymptotic joint covariance is $\mathbf{I}_1^{-1} = \begin{pmatrix} I_{11}^{-1} & v_{11,12} i_{11}^{-1} i_{12}^{-1} \\ v_{11,12} i_{11}^{-1} i_{12}^{-1} & I_{12}^{-1} \end{pmatrix}$. Assuming that $\binom{\theta_1}{\theta_2}$ has the prior distribution $N(\theta_0, \mathbf{I}_0^{-1})$, the posterior distribution of $\binom{\theta_1}{\theta_2}$ given $\binom{\hat{\theta}_{11}}{\hat{\theta}_{12}}$ is

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \left| \begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix} \sim N\left((\mathbf{I}_0 + \mathbf{I}_1)^{-1} \left(\mathbf{I}_0\theta_0 + \mathbf{I}_1 \begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix} \right), (\mathbf{I}_0 + \mathbf{I}_1)^{-1} \right).$$

Because the interim decision should be driven by the first stage data and not much influenced by any prior assumptions about the putative treatment effects, a vague prior is used by letting $\mathbf{I}_0 \to \mathbf{0}$ [10]. Thus, the posterior simplifies to

$$\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \left| \begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix} \sim N\left(\begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix}, \mathbf{I}_1^{-1} \right). \right. \tag{10}$$

In line with the general asymptotic approach to the analysis of survival data, this assumes $\mathbf{I}_1$ to be known.

The predictive distribution is defined as the conditional distribution of $\binom{\hat{\theta}_{21}}{\hat{\theta}_{22}}$ given $\binom{\hat{\theta}_{11}}{\hat{\theta}_{12}}$. A simple application of the convolution theorem for normal distributions shows that

$$\begin{pmatrix}\hat{\theta}_{21}\\ \hat{\theta}_{22}\end{pmatrix} \left| \begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix} \sim N\left(\begin{pmatrix}\hat{\theta}_{11}\\ \hat{\theta}_{12}\end{pmatrix}, \mathbf{I}_1^{-1} + \begin{pmatrix} (I_{21}-I_{11})^{-1} & v \\ v & (I_{22}-I_{12})^{-1} \end{pmatrix} \right), \right. \tag{11}$$

with $v = \frac{v^*_{21,22}}{\sqrt{(I_{21}-I_{11})(I_{22}-I_{12})}} = \frac{\sum_{k=1}^{d_2}\phi_k - \sum_{k=1}^{d_1}\phi_k}{(I_{21}-I_{11})(I_{22}-I_{12})}$ (see formula (6)) in case of the vague prior.

Considering the two treatments $j = 1, 2$ separately, the predictive probability that the final log-rank statistic lies above some fixed threshold c is obtained from (11) as

$$Pr\left(\tilde{l}_{2j} > c \,\middle|\, \hat{\theta}_{1j}\right) = \Phi\left(\frac{i_{1j}}{i_{2j}} \left(c - \sqrt{I_{2j} - I_{1j}}\hat{\theta}_{1j} \right) \right). \tag{12}$$

The resulting predictive power (12) is calculated at interim for both treatments and for both PFS and OS. It is the probability of a successful trial (i.e. a significant trial, irrespective of 'correct' treatment arm selection), given the interim results and a vague Normal prior for the true log-hazard ratios θ_j, $j = 1, 2$. Only one of the four predictive probabilities corresponds to 'reality': the one for OS and the selected treatment. The predictive probability for OS in the deselected treatment arm corresponds to a 'what if' scenario; the predictive probabilities for PFS correspond to 'virtual' tests of PFS, which play no role in the final decision on the primary endpoint but are used for decision making. The calculation of the PFS predictive probability requires additional approximations of expected events (whereas the total and interim number of cases is fixed for OS). Details are found in Appendix A.2. Appendix A.1 shows how the threshold c in (12) is selected in the concrete application.

The suggested approach makes a number of simplifying assumptions. In particular, it considers separate marginal prior (and hence posterior) distributions of PFS and OS. More sophisticated approaches might attempt to consider some joint prior distribution of the hazard ratios of PFS and OS. This was not carried out here mainly because of two reasons: (i) setting up a reasonable multivariate prior for PFS and OS jointly in the absence of strong prior beliefs is a very difficult task and (ii) a relatively crude rule serves our purpose well enough, as this is only used for selecting a treatment to be continued into stage 2

and not for the ultimate decision about treatment efficacy and/or effect size. In addition, it should be considered that (maybe counter-intuitively) the correlation between PFS and OS in the bivariate exponential distribution assumed in simulations does not have a large impact on the power of the final statistical test (in contrast to the ratio of their respective hazard ratios; see Section 6.2).

Denote the four marginal predictive probabilities by $prob_{j,OS}, prob_{j,PFS}, j = 1,2$. In the following, we suggest ways of combining them for the futility and the selection decision.

4.2.2. Stop for futility. The general idea of the futility check is to stop the trial if the chance of success measured via predictive probabilities is too slim. Two thresholds are thus prespecified at study start. The trial is stopped for futility if $\max_j(prob_{j,OS}) < t_{OS}$ and $\max_j(prob_{j,PFS}) < t_{PFS}$.

4.2.3. Treatment arm selection. To select a treatment arm, evidence from PFS and OS is combined via the utility function

$$util_j = w_j \cdot prob_{j,PFS} + (1 - w_j) \cdot prob_{j,OS}. \tag{13}$$

The weight w_j governs the relative importance that we assign to the corresponding endpoints and should reflect not only the hierarchy between the two endpoints but also the actual number of events accrued up to the interim analysis. In Section 6, different suggestions for the weights w_j are investigated via simulation, one of which is

$$w_j = \frac{d_{1j,PFS}}{d_{1j,PFS} + 2 \cdot d_{1j,OS}} \tag{14}$$

where $d_{1j,PFS}$ and $d_{1j,OS}$ denote events (progressions and deaths, respectively) in treatment arm j up to interim (stage 1). Deaths are thus weighted twice.

It must be noted here that type I error rate violations can arise as a 'side effect' of treatment arm selection, if the selection of a treatment arm implies a non-prespecified change in the recruitment rate [11]. This can happen if, for example, the possibility of keeping only one or both treatment arms after the interim analysis is left open, but the recruitment rate is fixed irrespective of the number of continued treatment arms, such that in a retained arm, it is higher if the other arm is dropped. For such a design, asymptotic type I error control cannot be guaranteed. The problem is avoided if the recruitment rate of any retained arm follows a prespecified rule. An easy way to enforce this is to require that only one treatment arm is carried forward into stage 2.

It should also be noticed that this is a potential problem for ordinary non-adaptive group-sequential TTE trials as well. From a theoretical perspective, asymptotic type I error control via the independent increments argument is valid, if any change in the recruitment rate after the interim analysis is stochastically independent of the observed value of the test statistic. In practice, a change of the recruitment rate after the interim analysis is usually accepted if there are convincing operational reasons explaining it (e.g. slower or faster recruitment rates than anticipated or delays in feedback from centres, etc.).

5. Final decision making: adjustment for interim treatment selection

If stage 2 of the design is performed, a final decision must be reached regarding efficacy of the selected treatment. In this context, the group-sequential log-rank test discussed in Section 3 will be used with an adjustment for the selection. There is an extensive literature on adjustments for treatment arm selection (e.g. [12–15]), but relatively little has been written about the intricacies associated with applying these general approaches in the context of TTE analysis. In addition, these methods are often slightly different in their aims: some focus on establishing efficacy for the *best* treatment, which does not necessarily have to coincide with the *selected* treatment (e.g. [14]).

Also, some approaches require that only the most efficacious treatment is selected at interim and/or that exactly one treatment is selected. In contrast, in what follows, we investigate a number of approaches that do not require to select the most efficacious treatment at interim (e.g. if safety concerns suggest to go with a marginally less efficacious but clearly better tolerable treatment). The approaches are dealt with in Sections 5.2 to 5.5. Section 5.1 sets the scene.

5.1. *Approximations of missing information: expected events, correlations and weights between stages*

In the following, let us assume that treatment 1 is selected at interim. Obviously, we do not have a value for $\tilde{l}_{22}$ or for $v^*_{21,22}$ from the joint distribution (6). However, for an adjustment of the final test, an approximation of the correlation $v^*_{21,22}$ under the null hypothesis of no treatment effect is needed. Under this null hypothesis and with equal sample sizes in all three groups, we expect that $p_{kj} \approx 1/2$, and thus, $p_{kj}(1-p_{kj}) \approx 1/4$. Hence, the approximation of the correlation becomes

$$v^*_{21,22} = \frac{(d_2-d_1)\frac{1}{2}\frac{1}{2}\frac{1}{3}}{\sqrt{\frac{2}{3}\frac{1}{2}\frac{1}{2}(d_2-d_1)}\sqrt{\frac{2}{3}\frac{1}{2}\frac{1}{2}(d_2-d_1)}} = \frac{1}{2}.$$

Here, we need to set $d_i = d_{i1} + d_{iC}$, excluding the non-selected treatment.

Simulations indicate that it makes almost no difference if we use these simplified approximations or more complicated ones based on the actually observed data. As briefly touched upon in Section 4.1, similar to the approximations of the information fraction, this is not surprising. In addition, the approximation is needed under H_0 anyway, such that it is unclear in what sense a 'data-driven approximation' should be 'better'.

5.2. *Conservative Dunnett approach*

Using the results from the previous subsection for the calculation of critical values, this approach works as follows:

(1) At interim, calculate l_{11} and l_{12} and c_1 from (7). Reject H_0 if $\max(l_{11}, l_{12}) \geqslant c_1$.
(2) If $\max(l_{11}, l_{12}) < c_1$, drop the worse treatment where 'worse' may mean 'less efficient' or 'less safe' or something similar. At the final analysis, calculate l_{2S} where S is the selected treatment. Reject H_0, if $l_{2S} \geqslant c_2$ where c_2 is calculated to fulfil (8).

This approach is asymptotically conservative, as it compares l_{2S} with a critical value that is 'intended' for the maximum of two test statistics.

5.3. *Bonferroni approach*

This approach is almost the same as the conservative Dunnett approach in Section 5.2. The only difference is that rather than using the critical value c_2 for the final test, the normal quantile $\Phi^{-1}(1-(\alpha-\alpha_1)/2)$ is used. Thus, only the 'stage-2-α' is split by the Bonferroni method. $\Phi^{-1}(1-(\alpha-\alpha_1)/2)$ is always larger than c_2; hence, this approach is more conservative than the Dunnett approach in Section 5.2. Its primary purpose is to provide a benchmark for assessing the magnitude of a potential power loss from using a very simple adjustment method for the final comparison.

5.4. *Conditional error function approach*

König *et al.* [12] have developed a conditional error function (CEF) approach for selecting treatments at an interim analysis. We will adapt this approach to the situation at hand.

In case of only two treatments, one of which is selected at interim, the approach can be summarized as follows:

(1) After stage 1, calculate the two test statistics l_{11} and l_{12}. If $\max(l_{11}, l_{12}) > c_1$, we stop early for efficacy. Otherwise, obtain the conditional probabilities

$$q_{12} = Pr_{H_0}(\max(l_{21}, l_{22}) \geqslant c_2 | l_{11}, l_{12}) \tag{15}$$

and

$$q_S = Pr_{H_0}(l_{2S} \geqslant c_S | l_{1S}) \tag{16}$$

where $S \in \{1, 2\}$ denotes the selected treatment. Here, c_2 and c_S are chosen such that they fulfil (8) and

$$Pr_{H_0}(\max(l_{11}, l_{12}) < c_1, l_{21} \geqslant c_S) = \alpha - \alpha_1, \text{ respectively.}$$

As mentioned previously, it is not difficult to obtain c_2 and c_S by a numerical search.

(2) At the final analysis, $\tilde{l}_{2S}$ is calculated as in (5), and the corresponding p-value p_{2S} is obtained. If $p_{2S} < \min(q_{12}, q_S)$, superiority of the treatment over control is concluded; otherwise, H_0 cannot be rejected.

To calculate the quantities q_{12} and q_S, the approximation of $v^*_{21,22}$ in Section 5.1 is plugged into the distribution (6). The null distribution is then derived using standard results for conditional multivariate normal distributions (see, e.g. [16, Theorem 3.2.4]):

$$\sqrt{\frac{d_2}{d_2-d_1}}\begin{pmatrix} l_{21} \\ l_{22}\end{pmatrix} - \sqrt{\frac{d_1}{d_2-d_1}}\begin{pmatrix} l_{11} \\ l_{12}\end{pmatrix} \Bigg| \begin{pmatrix} l_{11} \\ l_{12}\end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0\end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1\end{pmatrix}\right).$$

Hence,

$$q_{12} = 1 - \Phi_{0.5}\left(\sqrt{\frac{d_2}{d_2-d_1}}c_{12} - \sqrt{\frac{d_1}{d_2-d_1}}l_{11}, \sqrt{\frac{d_2}{d_2-d_1}}c_{12} - \sqrt{\frac{d_1}{d_2-d_1}}l_{12}\right)$$

and

$$q_S = 1 - \Phi\left(\sqrt{\frac{d_2}{d_2-d_1}}c_S - \sqrt{\frac{d_1}{d_2-d_1}}l_{1S}\right)$$

where $\Phi_\rho(\cdot,\cdot)$ denotes the cumulative distribution function of the bivariate standard normal distribution with correlation ρ between the two variables.

König *et al.* [12] show that this CEF approach is uniformly more powerful than the conservative Dunnett method in Section 5.2. However, it must be kept in mind that with the CEF approach (asymptotic) α-level control is guaranteed only if $\tilde{l}_{2S}$ is stochastically independent of l_{11} and l_{12}. This assumption may be violated if some feature of patients recruited before the first interim analysis impacts l_{2S} (or, equivalently, $\tilde{l}_{2S}$) without being accounted for by the adjustment of the final analysis. As an example, this can happen, if the selection of a treatment arm at stage 1 is not only based on the observed values of l_{11} and l_{12} but also on progression events.

The conservative Dunnett approach (Section 5.2) and the Bonferroni approach (Section 5.3) are not affected by this problem. The key difference is that these two approaches do not explicitly use the (asymptotic) distribution of l_{2S} but rather replace it (conservatively) by that of $\max(l_{21}, l_{22})$.

5.5. *Combination p-value approach*

This approach [17] requires that we fix weights $w_1 > 0$ and $w_2 > 0$, $w_1^2 + w_2^2 = 1$ before the final analysis. Furthermore, we define the p-values $p_{1,12} = 1 - \Phi_{0.5}(\max(l_{11}, l_{12}), \max(l_{11}, l_{12}))$, $p_{1S} = 1 - \Phi(l_{1S})$ and $p_{2S} = 1 - \Phi\left(\tilde{l}_{2S}\right)$ where S denotes the selected treatment.

The null hypothesis of no differences between treatment and placebo is rejected if either $p_{1,12} \leqslant \alpha_1$ after stage 1 or $w_1 \cdot \Phi^{-1}(1-\max(p_{1,12}, p_{1S})) + w_2 \cdot \Phi^{-1}(1-p_{2S}) \geqslant c_2$ where, in analogy with formulae (7) and (8), c_2 is chosen such that under H_0,

$$1 - \Phi_{w_1}\left(\Phi^{-1}(1-\alpha_1), c_2\right) = \alpha \tag{17}$$

is kept. If $\alpha_1 = 0$ (i.e. if there is only a selection at interim, no potential stop for efficacy), then $c_2 = \Phi^{-1}(1-\alpha)$. A natural choice for w_1^2 and w_2^2 would be $w_1^2 = \frac{d_1}{d_2}$ and $w_2^2 = 1 - \frac{d_1}{d_2}$ or some other approximation of expected information fractions in the two stages.

This approach is very similar to, but not identical with, the CEF approach in Section 5.4. Although the combination p-value approach can be converted into a CEF approach [18], this does not yield the CEF approach used here. Like the CEF approach, the combination p-value approach requires stochastic independence of $\tilde{l}_{2S}$ from (l_{11}, l_{12}), such that the same caveats as in Section 5.4 apply regarding potential type I error violations.

6. Simulations

This section investigates the operating characteristics of the planned trial by simulations. The scenarios range from optimistic via realistic to 'approximately worst case'. Both power and type I error control

are investigated. First, the set-up and the assumptions of the simulations are described, then results are reported and discussed. Comparisons are made across the decision-making criteria of Section 4 and the methodologies described in Section 5 for treatment arm selection. As a benchmark, a standard sequential design without treatment arm selection is considered.

6.1. *Design assumptions and simulation set-up*

The design assumptions are based on the concrete clinical context of the trial. Hence, a total number of 1000 patients is recruited with the final analysis occurring after 600 deaths. The randomization ratio is 1:1(:1) in Stage 2 (1). The null hypothesis for efficacy testing is $H_0 : \log(\text{HazardRatio}) = 0$, which is tested as a one-sided hypothesis with $\alpha = 5\%$. The level of proof t_{PFS}, t_{OS} for the futility stop (Section 4.2.2) is set to 35%. Furthermore, it is assumed that a single site recruits an average of 0.5 patients every month for a total accrual period of 25 months. Regarding entry of centres into the trial, a staggered recruitment scheme based on projections for the trial is implemented. Minimum follow-up time is assumed to be 6 months.

The interim analysis time point is set at an information rate (IR) of either 20% or 30% of the final OS events. The final test for efficacy is conducted on OS only. All four approaches of Sections 5.2 to 5.5 (conservative Dunnett (CD) test, Bonferroni (Bonf) approach, CEF approach and combination (Combi) p-value approach) are investigated.

Several ways of combining PFS and OS events into a decision criterion according to the general strategy outlined in Section 4.2.3 are considered. These are described in the succeeding section together with the corresponding simulation results. Regarding the simulation of interim decision making, the information ratio for PFS is estimated from the simulated number of PFS events in every simulation run.

With respect to data generation within the simulation, observed progression and survival times are assumed to arise from a Moran–Downton bivariate exponential distribution [19]. This distribution depends on the hazard ratios of each endpoint and the correlation between them. Many variations of these parameters were investigated. Selected results are presented in Section 6.2.

6.2. *Simulation results*

All subsequently reported results are based on 10,000 simulation runs per scenario.

Figure 3 (top) shows the power of the suggested approaches for various scenarios. In addition to the fixed assumptions given in Section 6.1, these use a correlation of 0.4 between PFS and OS test statistic, nominal type I error $\alpha = 5\%$, $IR = 30\%$ and the combined selection rule (13) with fixed weights $1/3$ and $2/3$ for PFS and OS, respectively. The hazard ratios for PFS and OS were prespecified according to the targeted treatment benefit shown in Table I: median OS for control was assumed to be 7.5 months and the median OS for the better treatment 10 months in the *strong* effect scenario and 8.5 months in the *weak* effect scenario. The effect of the remaining (worse) treatment arm is set 'halfway' between 'better' and control (e.g. to median OS 8.75 months if median OS is 10 months for the better treatment corresponding to a hazard ratio of 0.857 for the worse treatment). Both of these OS scenarios were repeated with different assumptions about PFS: median PFS for control was set to 3.5 months and the targeted median for PFS for the better treatment was set to (3.5, 4, 4.5, 5) months with the same halfway interpolation for the other treatment as for OS.

As an informal comparison with a *benchmark* design, a three-arm non-adaptive group-sequential phase III trial without treatment arm selection at the interim analysis is considered. With respect to the number of events observed at final, the benchmark design considers 400 events for each arm thus reflecting the 600 events of the simulated three-way designs with treatment arm selection.

As one possibility, a simple Bonferroni adjustment (i.e. testing both treatment arms at the split level $\alpha/2$) was considered. This option is sometimes used in practice because of its simplicity (e.g. [20]). For this purpose, EAST v.5.0 ([21]; Cytel, Inc., Cambridge, MA, USA) was used, which allows to specify the same staggered recruitment scheme as in the simulations carried out, using a one-sided α-level of 2.5% and two-arms only (active versus control). The resulting design would have a power of around 82% and a maximum accrual of around 1000 patients. This design did not foresee a stop for futility.

Alternatively, we also simulated the use of a Dunnett-type test in this non-adaptive benchmark design. For this, we applied the same futility stopping rules that were used for the adaptive designs. The power of this second benchmark option depends on the assumed PFS and OS hazard ratios, the correlation between PFS and OS and the futility rule used. For the situation shown in Figure 3 (top panel, left), it ranges from 80% to 84%. Similarly, in other scenarios, the power is somewhat lower than that of the

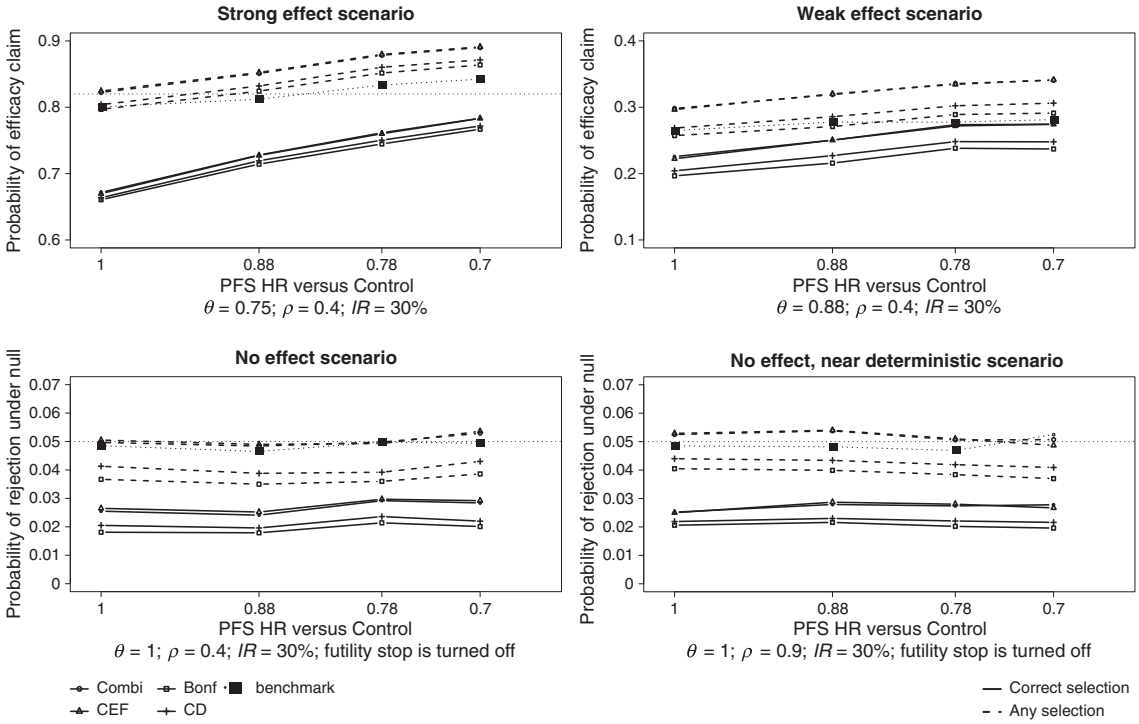


Figure 3. Progression-free survival (PFS) versus chance of success at final analysis across methods. Correct selection = selection of better arm and rejection of null; any selection = rejection of null; dotted horizontal line at α-level or indicating benchmark design power.

treatment-selection designs. The differences in power primarily depend on the true hazard ratios. It is smaller when there is no treatment effect on PFS but a strong one on OS. When hazard ratios are similar in PFS and OS, the power gain of the treatment-arm-selecting designs over the Dunnett-type benchmark is typically 2% to 3% in the scenarios investigated.

Figure 3 shows that under the *strong effect* scenario (top panel, left), the methods achieve good power (above 80%) with a substantial gain over the benchmark approaches. As expected, the power increases monotonously with PFS benefit. Under both scenarios (left and right top panel), the CEF and the combination p-value approaches provide very similar results and are consistently better than the conservative Dunnett and the Bonferroni approaches. Obviously, there is some positive probability that the less effective treatment is picked at interim and still the demonstration of a significant treatment effect at the final analysis is achieved. The solid and dashed lines in Figure 3 illustrate the difference between the corresponding two power concepts. The difference between these powers is only a minor concern. If treatment effects are substantially different, the probability of selecting the inferior treatment and achieving significance is low. If, however, the difference in effect size is small, selection of the 'right' treatment is not of particular concern.

Figure 3 (bottom, left) is based on the same assumptions as Figure 3 (top) except that the median OS under the treatments is set to 7.5 months—the same as for the control. Hence, the graph shows the actual type I error rates in these scenarios when H_0 holds true. The error seems to be generally controlled (irrespective of the correlation between PFS and OS). The CEF and the combination p-value approaches are the most liberal, whereas the Dunnett and the Bonferroni approaches reveal the expected slight conservatism under the null hypothesis.

To see the extent of the potential type I error violation discussed for the CEF and the p-value combination approaches in Section 5, Figure 3 (bottom, right) shows simulation results under a null scenario that attempts to approximate a 'worst-case scenario' where deaths follow progression with a fixed, non-stochastic time lag and the interim time point is placed such that progressions have already occurred in every patient at interim. This implies that the exact order of stage 2 deaths from stage 1 recruits who survive the interim time point is already determined at interim. As another feature of this worst-case scenario, the treatment arm corresponding to the better log-rank test on PFS is the one always chosen at interim. For this scenario, theoretical calculations suggest that the true rejection probability of the design

could go up to a maximum of approximately 7%. As expected, this extremely pessimistic boundary is never reached in any of the simulations carried out: even in the case of a very strong correlation between PFS and OS ($\rho = 0.9$), the simulated rejection rate never exceeded 5.6%. This statement also holds for simulations of the other weighting rules for combining OS and PFS events and for the IR of 30 %. In all these null simulations, the futility stop was turned off.

Figure 4 provides a comparison of interim criteria for treatment arm selection. The criteria are as follows: PP(PFS) = predictive power of PFS only, PP(OS) = predictive power of OS only, PP(PFS,OS)FW = predictive power of weighted average of OS and PFS with fixed weights (2/3 and 1/3, respectively), PP(PFS,OS)OW = predictive power of weighted average of OS and PFS with weights based on observed number of events (see formula (14)) and LOG-RANK = selection based on the log-rank test for OS. The combination p-value approach is used here. Results for the other approaches are very similar.

Figure 4 shows that the Bayesian predictive power tool allows to borrow strength at interim across the endpoints PFS and OS by combining them explicitly into a single utility index thus obtaining clear

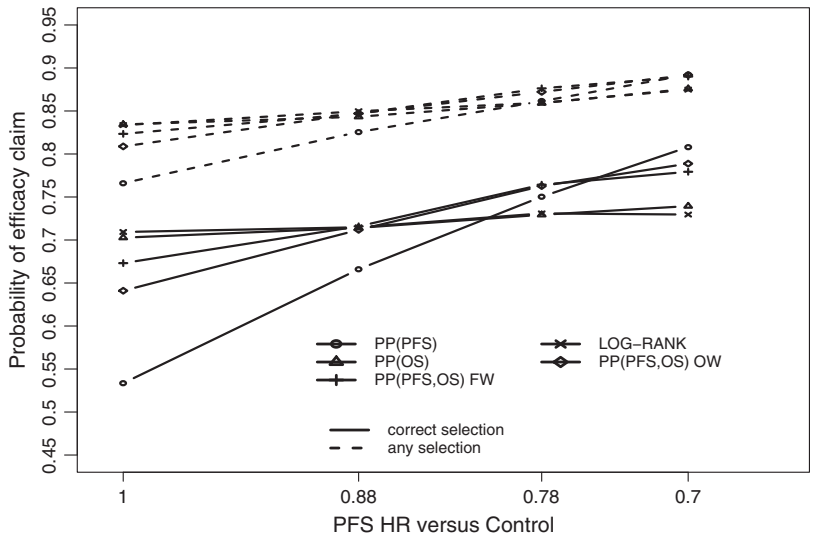


Figure 4. Progression-free survival (PFS) versus chance of success at final analysis across selection criteria: *strong effect scenario* ($\theta_{OS} = 0.75$, $\rho = 0.4$, $IR = 30\%$).

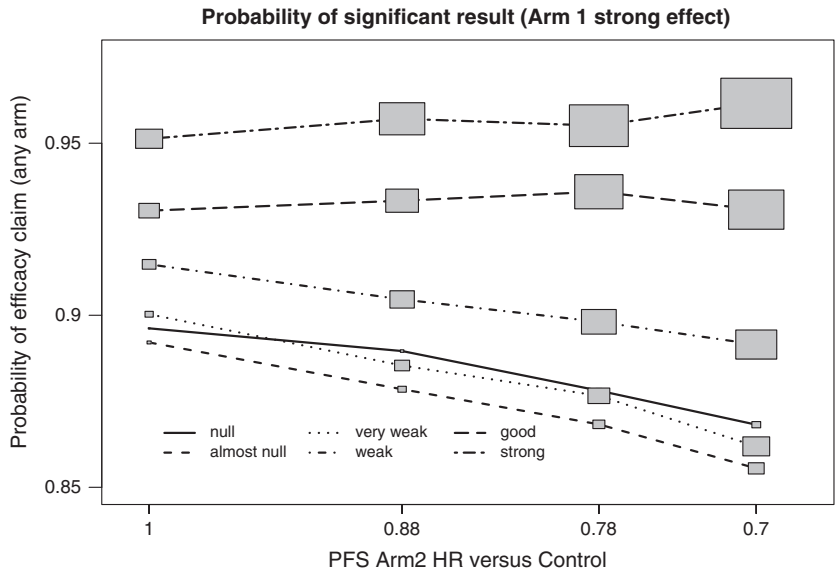


Figure 5. Probability of significant result with the combination p-value approach (arm 1 *strong effect scenario*: $\theta_{OS}^{Arm1} = 0.75$, $\theta_{PFS}^{Arm1} = 0.75$, $\rho = 0.4$, $IR = 30\%$).

advantages over the single endpoint. The combined (PFS,OS) interim decision rule fares better than a decision based on survival alone and also fares better than a decision based on progressions alone.

Finally, Figure 5 provides additional insight into the case of a *strong arm 1 effect* (in both OS and PFS) and the role of the inferior arm 2 as it varies across a range of OS and PFS values. It quantifies the trade-off between the chance of declaring efficacy at the final analysis and the probability of wrong selection at interim (i.e. the probability of arm 2 selection). The size of the boxes in the figure is proportional to the probability of arm 2 selection. The categories (*null*, *almost null*,. . . ,*strong*) of arm 2 treatment effects represent hazard ratios of 1.00, 0.94, 0.88, 0.83, 0.79 and 0.75, respectively, corresponding to median OS times of 7.5 to 10 months in steps of 0.5 month.

Obviously, if the arm 2 OS effect is *strong*, then the probability of selecting the wrong treatment is high, but there is no consequence on power, as both arms contribute to it. Slightly more surprising, power is worst if the arm 2 OS effect is *almost null* or *very weak*. In these situations, the trade-off between false selection and low power is at its worse. In general, independent of the arm 2 OS value, a mismatch between the PFS and the OS effect is always detrimental.

Regarding the correlation between PFS and OS, additional simulations show little impact of the correlation on the power in the range of realistic correlations between 0.2 and 0.8. These results are thus not presented in detail.

7. Discussion

In this paper, we investigate adaptive design options for an oncology phase II/III trial. It is seen that the use of Bayesian decision-making tools, combined with methods that control the type I error of the ultimate efficacy test, leads to designs that are substantially more efficient than conventional, non-adaptive approaches.

Care is needed when implementing adaptive designs in the TTE context. The 'independent increments structure' needed in 'Cox-type' survival analyses is not a trivially obvious property here. It can be violated in rather subtle ways (see, e.g. [22]). In essence, the problem arises from patients who are recruited before the interim analysis but are under investigation beyond the interim analysis time point. If design modifications at interim are based on information from such patients, then this information must either be stochastically independent of the primary endpoint or the influence of the information on the final analysis of the primary endpoint must explicitly be accounted for. Both of these solutions are difficult to attain. For example, in the oncology set-up explored in this paper, safety information, progressions and the like will usually be correlated with expected remaining survival time and thus not be stochastically independent of survival. In several publications on adaptive designs for TTE trials (e.g. [23], in the context of sample size re-estimation), this is not mentioned. Hence, these publications work with the implicit assumption that any interim modification of the trial is based exclusively on the observed value of the primary test statistic.

Here, the study foresees treatment arm selection of one out of two treatment arms. Thus, the potential for violation of the true type I error rate due to interim decisions is limited. Regarding the methods described in Section 5, the two conservative approaches (Bonferroni and conservative Dunnett) do not entail any (asymptotic) α-level violation (except the potential of an 'indirect' violation via the recruitment rate as discussed in Section 4.2.3). In contrast, for the CEF and the p-value combination approaches, the possibility cannot be ruled out. To address this concern, thorough investigations of the true type I error rates were carried out via simulation, including very pessimistic scenarios that were deliberately designed to expose any α-level violation. The results were reassuring: α-level violation occurs in extreme scenarios but is very minor (at most 5.6% and at nominal 5%). Thus, the use of the CEF and the p-value combination approaches is regarded as acceptable here. They have power advantage over the two other methods. In simulations, the CEF and the p-value combination approaches are virtually indistinguishable with respect to power. A preference for either of them is difficult to express. The p-value combination may be preferred for simplicity and the CEF approach for slightly higher flexibility and elegance. If one is concerned about potential challenges regarding type I error violations, the conservative Dunnett approach provides an attractive alternative, as it suffers only from a moderate power loss relative to the two 'non-conservative' approaches.

Regarding the approximations of information fractions, asymptotic distributions of the log-rank test statistics, the approach suggested here is not fundamentally different from the conventional TTE analysis. The very good convergence of the log-rank test to its asymptotic normal distribution is well known, and these results extend to the bivariate normal distribution of the related log-rank tests for two

treatments (e.g. [24]). Likewise, the log-rank test is reasonably robust against non-proportional hazards, as long as one treatment is consistently better than the other. Still, all derivations in Sections 3 and 5 are asymptotic, so some additional simulations are advisable when using similar methodology for a different trial.

Regarding the interim decision rules, it seems worthwhile to incorporate additional information like PFS by way of a Bayesian predictive power calculation as outlined in Section 4. The results are easy to communicate. There is a worthwhile power gain in case of a treatment benefit on both PFS and OS. According to the performed simulations, this result is not much affected by the correlation between the PFS and the OS endpoint. Only if PFS and OS are affected differently by the treatment (e.g. if time to death is prolonged by the treatment but time to progression is not), there is no gain (and maybe even a minor power loss) from including PFS in decision making. However, this latter situation will not occur in practice, as it is hardly conceivable that a cancer treatment prolongs time to progression but reduces time to death, or vice versa. If, however, treatment effects on PFS and OS are in the same direction, but of different magnitude, there is still some power gain from a decision making which includes PFS.

Another approach for the treatment arm selection at interim, which tries to overcome the issue of patients recruited before interim contributing to post-interim decisions, is to split patients into a stage 1 and a stage 2 recruits statistic. Following this approach, one would know both test statistics only at the end of the trial. This of course would preclude a possible stop for efficacy at interim and was therefore not investigated here.

The proper implementation of flexible designs in phase II/III trials is somewhat more intricate in the TTE situation than in conventional trials. Of course, closely monitoring study conduct so as not to introduce any operational bias (e.g. at the time when interim decisions are communicated) requires careful planning and implementation [25]. The difficulties are, however, not insurmountable. In important trials, the gains from making the effort can be substantial.

APPENDIX A. Computational and methodological details

A.1. Technical details concerning Section 4.2.1

In Section 4.2.1, c has been introduced as a given constant. To avoid arbitrarily setting this to some number, the following approach is adopted. In Section 5.4, we have derived

$$q_{12} = 1 - \Phi_{0.5}\left(\sqrt{\frac{d_2}{d_2 - d_1}}c_{12} - \sqrt{\frac{d_1}{d_2 - d_1}}l_{11}, \sqrt{\frac{d_2}{d_2 - d_1}}c_{12} - \sqrt{\frac{d_1}{d_2 - d_1}}l_{12}\right)$$

and

$$q_S = 1 - \Phi\left(\sqrt{\frac{d_2}{d_2 - d_1}}c_S - \sqrt{\frac{d_1}{d_2 - d_1}}l_{1S}\right).$$

To claim efficacy after stage 2, we must have $-\tilde{l}_{2S} < \Phi^{-1}(\min(q_{12}, q_S))$ (S denotes selected treatment and $\Phi^{-1}(\alpha)$ the standard normal α-quantile). Thus, it seems intuitively appealing to use $c = \Phi^{-1}(\min(q_{12}, q_S))$.

Mainly for convenience, the expected information ratio (9) is used instead of $\frac{d_1}{d_2}$ in the calculation of q_{12} and q_S. As d_1 and d_2 are known at interim ($d_1 = d_{1S} + d_{1C}$ is observed and $d_2 = d_{2S} + d_{2C} + d_1$ is the planned number of total events minus the observed number of events in the deselected arm), these values could also be used.

A.2. Approximating expected events and information fraction for progression-free survival

As briefly mentioned in Section 4.2.1, the calculation of the predictive probability for PFS requires some additional approximations of expected numbers of progression events. This is necessary, because the trial is planned such that interim and final analyses happen after certain, fixed numbers of deaths have occurred. This means that the number of progression events at these time points is random and thus needs to be approximated for the incorporation of a 'virtual test' on PFS that is used in the treatment arm selection as outlined in Section 4.2.3. Formula (12) requires an approximation of the PFS information fraction $\frac{i_{1j}}{i_{2j}}$, and w_j in Section 4.2.3 requires a closely related approximation of progression events $d_{1j,PFS}$.

In Section 4.1, we described how the OS information fraction is approximated. Here, we only outline the basic idea that was implemented to obtain a similar approximation for PFS: in formula (9), we are replacing the d_i^*'s by the expected number of progression events at time i, assuming that progression events from all three treatment groups follow the same exponential distribution with common intensity rate λ_{PFS}. This is carried out for the expected time point of the interim and the final analysis, which in turn is calculated from a corresponding assumption of exponentially distributed death times with a different intensity rate λ_{OS}.

Again, these approximations are only used to determine the information fraction for the virtual PFS test and the weights in the utility function $util_j$ in Section 4.2.3. The corresponding test statistics are calculated exactly the same way as for OS by using the approaches described in Section 5. This means that our approximations of expected progression events can be quite rough: they only impact treatment selection and not the confirmatory efficacy decision on OS.

Acknowledgements

We would like to thank Norbert Holländer and three anonymous reviewers for their constructive comments that helped improve the paper.

References

1. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *European Journal of Cancer* 2009; **45**:228–247.
2. FDA. Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. *FDA Doc. Ref. UCM071590*, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071590.pdf[May2007].
3. Lara PN, Redman MW, Kelly K, Edelman MJ, Williamson SK, Crowley JJ, Gandara DR. Disease control rate at 8 weeks predicts clinical benefit in advanced non-small-cell lung cancer: results from Southwest Oncology Group randomized trials. *Journal of Clinical Oncology* 2008; **26**:463–467.
4. Genz A, Bretz F. *Computation of multivariate normal and t probabilities*. Springer: Heidelberg, 2009.
5. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
6. Jennison C, Turnbull BW. *Group sequential methods with applications to clinical trials*. Chapman & Hall: London, 2000.
7. Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine* 2009; **28**:1445–1463.
8. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis,* 2nd edition. Chapman & Hall: Boca Raton, 2004.
9. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. Wiley: New York, 2004.
10. Bernardo JM, Smith AFM. *Bayesian Theory*. Wiley: Chichester, 1994.
11. Bauer P, Posch M. Letter to be editor: modification of the sample size and the schedule of interim analyses in survival trials based on data inspections. *Statistics in Medicine* 2004; **23**:1333–1335.
12. König F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
13. Sampson AR, Sill M. Drop-the-losers design: normal case. *Biometrical Journal* 2005; **47**:257–268.
14. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703.
15. Stallard N, Friede T. Flexible group-sequential designs for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
16. Mardia KV, Kent JT, Bibby MJ. *Multivariate Analysis*. Academic Press: London, 1979.
17. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.
18. Posch P, Bauer P. Adaptive two-stage design and the conditional error function. *Biometrical Journal* 1999; **41**:689–696.
19. Iliopoulos I. Estimation of parametric functions in Downton's bivariate exponential distribution. *Journal of Statistical Planning and Inference* 2003; **117**:169–184.
20. Barnes P, Pocock SJ, Magnussen H, Iqbal A, Kramer B, Higgins M, Lawrence D. Integrating indacaterol dose selection in a clinical study in COPD using an adaptive seamless design. *Pulmonary Pharmacology & Therapeutics* 2010; **23**:165–171.
21. Cytel Inc. *East 5.0 Manual*, 2007.
22. Jahn-Eimermacher K, Ingel A. Adaptive trial design: a general methodology for censored time to event data. *Contemporary Clinical Trials* 2009; **30**:171–177.
23. Li G, Shih WJ, Wang Y. Two-stage adaptive design for clinical trials with survival data. *Journal of Biopharmaceutical Statistics* 2005; **15**:701–718.
24. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer: New York, 2000.
25. FDA. Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics (Draft). *FDA Doc. Ref. UCM201790*, available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM201790.pdf [February 2010].

MAIN PAPER

Pharmaceutical Statistics

(wileyonlinelibrary.com) DOI: 10.1002/pst.504

Published online 15 February 2012 in Wiley Online Library

An evaluation of methods for testing hypotheses relating to two endpoints in a single clinical trial

Ting-Li Su,[a] Ekkehard Glimm,[b] John Whitehead,[a]* and Mike Branson[b]

The issues and dangers involved in testing multiple hypotheses are well recognised within the pharmaceutical industry. In reporting clinical trials, strenuous efforts are taken to avoid the inflation of type I error, with procedures such as the Bonferroni adjustment and its many elaborations and refinements being widely employed. Typically, such methods are conservative. They tend to be accurate if the multiple test statistics involved are mutually independent and achieve less than the type I error rate specified if these statistics are positively correlated. An alternative approach is to estimate the correlations between the test statistics and to perform a test that is conditional on those estimates being the true correlations.

In this paper, we begin by assuming that test statistics are normally distributed and that their correlations are known. Under these circumstances, we explore several approaches to multiple testing, adapt them so that type I error is preserved exactly and then compare their powers over a range of true parameter values. For simplicity, the explorations are confined to the bivariate case. Having described the relative strengths and weaknesses of the approaches under study, we use simulation to assess the accuracy of the approximate theory developed when the correlations are estimated from the study data rather than being known in advance and when data are binary so that test statistics are only approximately normally distributed.

Keywords: Bonferroni; clinical trials; multiple hypotheses; O'Brien test; Simes test

1. INTRODUCTION

Whereas classical theory of design and analysis of clinical trials addresses comparisons of one experimental treatment with one control treatment in respect of one patient response, the reality of many studies is more complicated. The evaluation of several patient responses leads to simultaneous testing of several hypotheses. Many approaches have been suggested to reconcile preservation of type I error with a multiplicity of inferences to be drawn.

The problem of testing the general null hypothesis H_0: $\boldsymbol{\theta} = 0$ against restricted alternatives, where the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$, has been tackled from different perspectives both in the multiple testing literature and in the literature on multivariate statistics. Common multiple testing approaches are the Bonferroni adjustment, or the variation due to Simes [1]. Discussions of these methods are available [2, 3], and recently, Bretz *et al.* [4] introduced a general graphical approach for representing many multiple testing procedures in an intuitive way.

A simple multivariate approach is to test whether a specific linear combination of the θ_i is equal to zero [5]. More sophisticated are likelihood-ratio tests against the alternative that all $\theta_i \geq 0$ and at least one $\theta_i > 0$, which have been developed assuming that the covariance matrix of the estimates of the θ_i is known [6] or unknown [7]. Silvapulle and Sen [8] give an overview of methods suggested for tests against restricted alternatives such as $0 < \theta_1 < \theta_2 < \ldots < \theta_p$. Another approach uses the unrestricted alternative $\boldsymbol{\theta} \neq 0$ but constructs tests to have high power against specific alternatives such as $\theta_1 = \cdots = \theta_p > 0$ [9–11]. When the alternative is unrestricted, rejection of H_0 might occur due to data supporting an alternative other than that for which high power has been set [12], but modifications to avoid this problem can be made [13].

In this paper, we consider the following situation that gives rise to the simultaneous testing of two null hypotheses. Patients are randomised between two treatments, E and C. They are assessed according to two different endpoints. The parameter θ_i denotes the advantage of E over C in terms of the i-*th* of these endpoints. We investigate the properties of the Bonferroni, Simes and O'Brien approaches and modifications to them, but instead of relying on the conservative properties of these procedures, use estimated correlations to sharpen their efficiency. In doing so, we follow Senn and Bretz [14] who explore the relationship between the power of a Bonferroni procedure and the value of the correlation coefficient. We restrict attention to the case in which two hypotheses are to be tested. In large samples, statistical hypothesis tests are often based on asymptotically normally distributed test statistics. Given such normality, we investigate 'exact' results for corresponding multivariate normal distributions.

[a] *Medical and Pharmaceutical Statistics Research Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, UK*

[b] *Novartis Pharma, Basel, Switzerland*

**Correspondence to: John Whitehead, Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster LA1 4YF, UK.*
E-mail: j.whitehead@lancaster.ac.uk

Intrinsically, conservative procedures are adjusted to have a precise prespecified type I error level so that fair comparisons of power can be made. Our objective is to identify which procedures have the greatest power to detect specific types of departure from the null hypothesis. Having made comparisons in the ideal situation of test statistics with exact normal distributions, we investigate the validity of the conclusions when the data are binary using simulation.

2. SPECIFICATION OF SIX POTENTIAL TESTING PROCEDURES

We will follow the closed-testing procedure, which begins with a test of the null hypothesis $H_{0,12}$: $\theta_1 = \theta_2 = 0$. This means that E has no effect on either response. If that null hypothesis is rejected, then we proceed to test the individual hypotheses $H_{0,1}$: $\theta_1 = 0$ and $H_{0,2}$: $\theta_2 = 0$. If each rejection is made at the one-sided level α, then the procedure maintains the multiple type I error rate, that is

(i) P(reject $H_{0,12} \mid (\theta_1, \theta_2) = (0,0)) \leq \alpha$;
(ii) P(reject $H_{0,1} \mid (\theta_1, \theta_2) = (0, \theta_2)) \leq \alpha$, for any value of θ_2;
(iii) P(reject $H_{0,2} \mid (\theta_1, \theta_2) = (\theta_1, 0)) \leq \alpha$, for any value of θ_1

hold simultaneously. Hence, the procedure maintains control of the family-wise error rate in the strong sense [2, 3, 15].

The component tests could be based on score statistics. Denote the efficient score statistic for θ_i by Z_i and its null variance by V_i, $i = 1, 2$. In large samples, Z_i is normally distributed with mean $\theta_i V_i$ and variance V_i. Alternatively, Wald tests of $H_{0,i}$ are based on the maximum likelihood estimate $\hat{\theta}_i$ and its standard error $se(\hat{\theta}_i)$. As $\hat{\theta}_i$ $se(\hat{\theta}_i)^2$ is normally distributed with mean $\theta_i se(\hat{\theta}_i)^2$ and variance $se(\hat{\theta}_i)^2$, the statistics Z_i and V_i could be replaced by $\hat{\theta}_i$ $se(\hat{\theta}_i)^2$ and $se(\hat{\theta}_i)^2$ in what follows. Let $corr(Z_1, Z_2) = \rho$. Treating V_1 and V_2 as fixed, it follows that $corr(\hat{\theta}_1, \hat{\theta}_2) = \rho$ as well. The value of ρ can be estimated from the study data for various response types [16–18].

The test of $H_{0,i}$ will reject if $Z_i/\sqrt{V_i} \geq \Phi^{-1}(1-\alpha)$ or equivalently if the one-sided p-value $p_i = 1 - \Phi(Z_i/\sqrt{V_i}) \leq \alpha$, $i = 1, 2$. Possible procedures differ in terms of how $H_{0,12}$ is tested. Six possible procedures are shown in Figure 1. Figure 1(a) shows a conservative Bonferroni test constructed for the case $\alpha = 0.025$. The joint null hypothesis $H_{0,12}$ will be rejected if $p_{bon} \leq 0.025$, where $p_{bon} = 2\min(p_1, p_2)$. As $\Phi^{-1}(1 - 0.0125) = 2.2414$, it follows that $H_{0,12}$ will be rejected if either $Z_1 \geq 2.2414\sqrt{V_1}$ and $Z_2 \leq 2.2414\sqrt{V_2}$ or $Z_2 \geq 2.2414\sqrt{V_2}$ and $Z_1 \leq 2.2414\sqrt{V_1}$ or $Z_1 \geq 2.2414\sqrt{V_1}$ and $Z_2 \geq 2.2414\sqrt{V_2}$. This is the shaded rejection region in Figure 1(a). Following the closed test procedure, if *in addition* $Z_i \geq 1.9600\sqrt{V_i}$, then $H_{0,i}$ will be rejected. This is equivalent to the Bonferroni–Holm procedure [19].

Figure 1(b) shows a Simes test that will reject $H_{0,12}$ if $p_{sim} \leq 0.025$, where $p_{sim} = \min\{2\min(p_1, p_2), \max(p_1, p_2)\}$. Rejection occurs if $Z_1 \geq 2.2414\sqrt{V_1}$ and $Z_2 \leq 1.9600\sqrt{V_2}$ or if $Z_2 \geq 2.2414\sqrt{V_2}$ and $Z_1 \leq 1.9600\sqrt{V_1}$ or if $Z_1 \geq 1.9600\sqrt{V_1}$ and $Z_2 \geq$

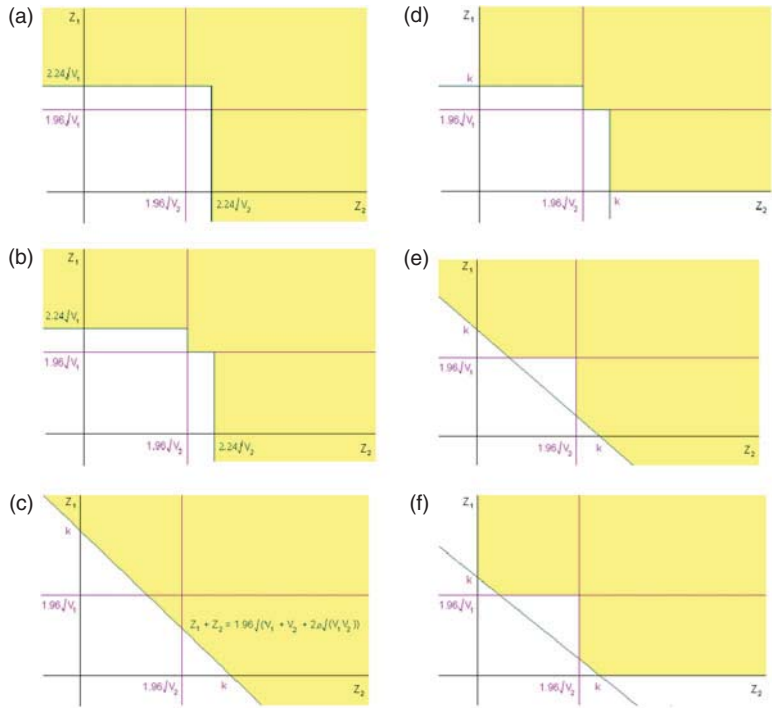


Figure 1. The rejection regions (shaded) of six possible procedures for testing two hypotheses (with $\alpha = 0.025$). (a) Bonferroni; (b) Simes; (c) combined; (d) restricted Simes; (e) consonant combined; (f) restricted consonant combined.

$1.9600\sqrt{V_2}$. The rejection region is indicated in Figure 1(b), and it strictly contains that of the Bonferroni test, so that the Simes test shown must be more powerful than the Bonferroni test.

Figure 1(c) shows a *combined test* that is based on the test statistic $Z_1 + Z_2$ and is an asymptotic version of the approach of O'Brien [5]. Under $H_{0,12}$, $Z_1 + Z_2$ is normally distributed with mean 0 and variance $V_1 + V_2 + 2\rho\sqrt{(V_1V_2)}$, where $\rho = \text{corr}(Z_1, Z_2)$. Thus, $H_{0,12}$ is rejected if $Z_1 + Z_2 \geq 1.9600\sqrt{\{V_1 + V_2 + 2\rho\sqrt{(V_1V_2)}\}}$, as shown in Figure 1(c). If *in addition* $Z_i \geq 1.9600\sqrt{V_i}$, then $H_{0,i}$ will be rejected. The test of $H_{0,12}$ will not be strictly conservative, as ρ has to be estimated from the study data. By the Neyman–Pearson Lemma, the combined test (considered alone) is the most powerful test of $H_{0,12}$: $\theta_1\sqrt{V_1} = \theta_2\sqrt{V_2} = 0$ vs $H_{1,12}$: $\theta_1\sqrt{V_1} = \theta_2\sqrt{V_2} > 0$.

A further modification of the Simes test is shown in Figure 1(d). Recall that E is being assessed according to two responses and rejection of $H_{0,12}$ will lead to a decision to investigate it further. If the trial in question is a phase II study, this will mean proceeding to phase III. If the trial in question is phase III, this will mean seeking a licence or promoting adoption of the treatment in practice. It may be that such a positive decision would be taken only if $H_{0,12}$ were rejected and $Z_1 \geq 0$ and $Z_2 \geq 0$. Thus, any indication that the experimental treatment is inferior to the control in either assessment of efficacy will be fatal to its progress. Brannath *et al.* [20] have shown that this modification of the Simes test (without the additional modifications of k discussed in Section 3) maintains the type I error in the presence of unknown correlation between the two test statistics, whereas the original Simes test keeps it only under the assumption that the correlation is larger than 0. Often such a restriction would apply, although Z_1 and Z_2 might be compared with small positive or small negative values rather than 0 (here the value 0 is retained for illustration). This procedure will be referred to as the restricted Simes test.

Returning to the combined test shown in Figure 1(c), it can be seen that it is possible to reject $H_{0,12}$ while not rejecting either of $H_{0,1}$ or $H_{0,2}$. This property is called *dissonance* (and its opposite *consonance*) [21, 22]. By contrast, the Bonferroni and Simes tests are consonant. Figure 1(e) shows a modified form of the combined test in which rejection of at least one of $H_{0,1}$ and $H_{0,2}$ is required to reject $H_{0,12}$ and the diagonal boundary is moved downwards and to the left to compensate and preserve type I error. This is referred to as the *combined consonant test*. Figure 1(f) shows a version of this procedure that is restricted to rejection only if in addition Z_1 and $Z_2 \geq 0$.

3. PROPERTIES OF THE TESTING PROCEDURES

In this section, we will investigate the six testing procedures introduced in Section 2, except that the Bonferroni procedure will be restricted to rejection of $H_{0,12}$ only if both Z_1 and $Z_2 \geq 0$. In each case, the value of ρ is treated as known, and the values of V_1 and V_2 are each set at 1. The value of α is set at 0.025, and the critical values for rejection of $H_{0,1}$ and $H_{0,2}$ taken to be 1.96. The critical values for rejection of $H_{0,12}$ are denoted by k for each procedure, with k replacing $2.24\sqrt{V_1}$ and $2.24\sqrt{V_2}$ in Figure 1(a) and (b) and being shown directly in Figure 1(c)–(f).

Table I shows the results of searching for suitable values of k on basis of numerical integration of the bivariate normal distribution over each rejection region, for 11 given values of ρ. When $\rho = 0$, and indeed for all values of $\rho \in (0, 0.0145)$, the restricted Bonferroni and restricted Simes procedures coincide and will reject if either Z_1 or $Z_2 \geq k$, and both Z_1 and $Z_2 \geq 0$, where $k \leq 1.9600$. For $\rho = 0$, $k = 1.9488$. The values of k differ between these two procedures for other tabulated values of ρ. The reductions to k shown for the Simes procedures itself, relative to the conservative value of $k = 2.2414$ that is traditionally applied, are modest unless ρ is close to 1. Dunnett and Tamhane [23] have suggested this approach and have generalised it to more than two comparisons.

The last three columns of Table I show the intercepts of the diagonal boundary (also denoted by k) for the three versions of the combined test, computed for the selected values of ρ. When ρ is small, the *restricted consonant combined test* becomes conservative, and when $\rho = 0$, the use of the critical value $k = 1.96$ results in a rejection rate of $\alpha - (1/2)\alpha^2$. The restriction makes very little difference for larger values of ρ.

In cases where $V_1 = V_2 = V$, with $V \neq 1$, the aforementioned results apply for the critical value $k\sqrt{V}$, with k as tabulated in Table I. Often, with the two null hypotheses relating to two endpoints for each study subject, the information measures V_1 and V_2 will indeed take similar values. When they do not, or when we have reason to believe that θ_1 is substantially different from θ_2,

Table I. Critical values (k) for six joint testing procedures with exact type 1 error and various values of known correlation (ρ).

ρ	Restricted Bonferroni	Simes	Restricted Simes	Combined	Consonant combined	Restricted consonant combined
0.0	1.9488	2.2414	1.9488	2.7718	2.2962	1.9600*
0.1	2.0202	2.2405	2.0204	2.9071	2.5053	2.4149
0.2	2.0788	2.2387	2.0802	3.0364	2.7058	2.6740
0.3	2.1247	2.2356	2.1281	3.1603	2.8966	2.8862
0.4	2.1579	2.2308	2.1639	3.2796	3.0770	3.0743
0.5	2.1778	2.2234	2.1871	3.3948	3.2465	3.2460
0.6	2.1841	2.2125	2.1968	3.5061	3.4048	3.4048
0.7	2.1758	2.1963	2.1919	3.6140	3.5517	3.5517
0.8	2.1520	2.1711	2.1706	3.7188	3.6870	3.6870
0.9	2.1081	2.1275	2.1275	3.8207	3.8103	3.8103
0.9999	1.9605	1.9607	1.9607	3.9200	3.9199	3.9199

*When $\rho = 0$, the critical value $k = 1.9600$ leads to a conservative test in this case.

the simple test statistic $Z_1 + Z_2$ used in the combined procedures might be replaced by $c_1Z_1 + c_2Z_2$ for suitable constants c_1 and c_2.

The six testing procedures can also be compared in terms of power. We again assume that $V_1 = V_2 = 1$, and allow θ_1 and θ_2 to range between 0 and 6. This setting provides a reasonable range of powers: the results can be scaled to give the properties of the tests for other values of V_1 and V_2. Figures 2 and 3 shows plots of the powers of these tests when $\rho = 0.2$ and $\rho = 0.8$, respectively. It can be seen that the Simes test is less powerful than the combined test when $\theta_1 = \theta_2$, this being more marked for the higher correlation. In fact, as stated in Section 2, the combined test is most powerful in this setting. This comparative feature can still be seen in the restricted versions of the tests when $\rho = 0.8$. Numerical results are presented in Table II, for $\rho = 0.5$ as well as for $\rho = 0.2$ and 0.8, and this allows more precise comparisons to be made. The results in Table II follow from numerical integration and not from simulation. The restricted tests have little power when θ_1 or $\theta_2 = 0$, which is intended, as power in this case is actually the risk of type I error. In principle, they should gain power elsewhere as a result. In practice, this occurs in regions where power is already high, and so it is of limited value. For example, in Table II see the row for $\rho = 0.2$, $\theta_1 = 3$ and $\theta_2 = 4$. The powers of the Simes and restricted Simes tests are 0.987 and 0.990, respectively. Imposing the restriction indeed enhances power but not to any meaningful extent. Even for these values, the restricted consonant combined test is not as powerful as the unrestricted version. An advantage of imposing the restriction can also be seen for the row of Table II in which $\rho = 0.5$, $\theta_1 = 3$ and $\theta_2 = 3$, but once more the gain is slight.

4. EVALUATION IN THE CONTEXT OF BINARY DATA

In this section, we study the characteristics of multiple testing using the statistics Z and V when the two correlated endpoints are

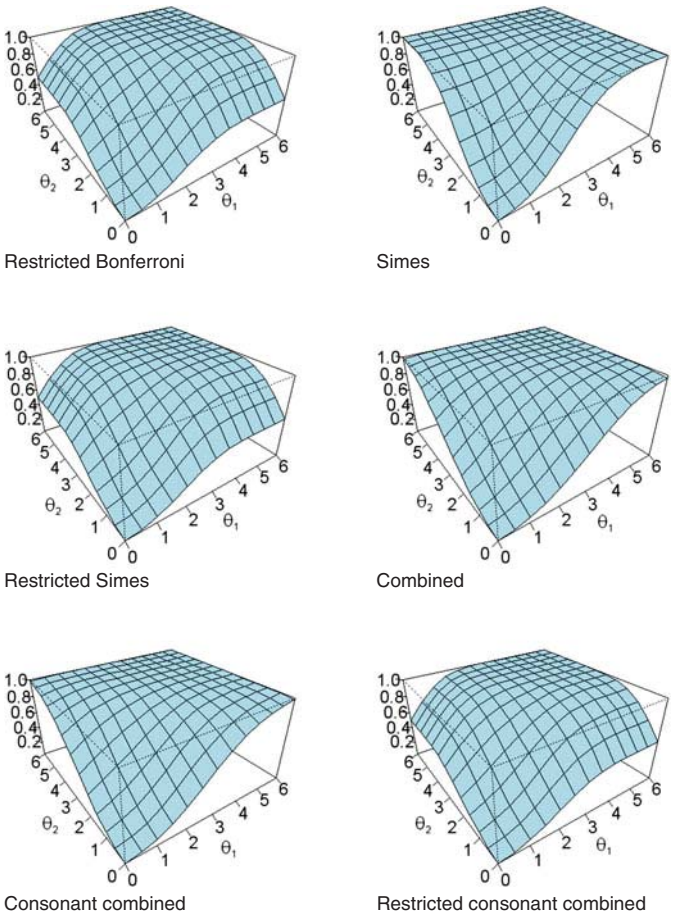


Figure 2. Power plots for the restricted Bonferroni, Simes, restricted Simes, combined, consonant combined and restricted consonant combined tests when $V_1 = V_2 = 1$ for θ_1 and θ_2 between 0 and 6 when the correlation $\rho = 0.2$.

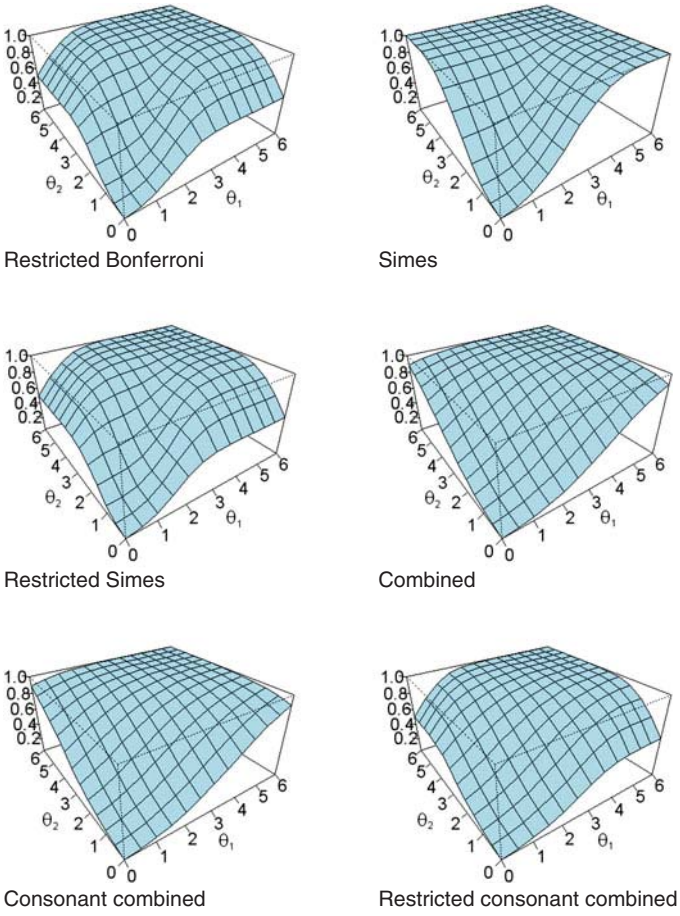


Figure 3. Power plots for the restricted Bonferroni, Simes, restricted Simes, combined, consonant combined and restricted consonant combined tests when $V_1 = V_2 = 1$ for θ_1 and θ_2 between 0 and 6 when the correlation $\rho = 0.8$.

binary. Assume for each endpoint, the results of a two-treatment-arm comparison study can be summarised as shown in Table III. The advantage of the experimental treatment over the control can be expressed as the log-odds ratio:

$$\theta = \log\left\{\frac{p_E(1-p_C)}{p_C(1-p_E)}\right\}, \tag{1}$$

where p_E and p_C denote the success probabilities on the experimental and control treatments, respectively. Taking an approach based on the conditional likelihood given the total number of successes, S, the efficient score statistic for θ, Z, and Fisher's information, V, are given respectively [24] by

$$Z = \frac{n_C S_E - n_E S_C}{n} \quad \text{and} \quad V = \frac{n_E n_C SF}{n^2(n-1)}. \tag{2}$$

Asymptotically, for large samples and small θ, conditional on the value of S, $Z \sim N(\theta V, V)$. When there are two binary endpoints B_1 and B_2 leading to two score statistics Z_1 and Z_2, the covariance C between them is

$$C = \frac{n_E n_C (nS_{12} - S_1 S_2)}{n^2(n-1)}, \tag{3}$$

where S_1 and S_2 now denote the total number of successes according to the two endpoints B_1 and B_2, respectively, and S_{12} denotes the number of patients who have succeeded according to both endpoints [16–18]. The correlation between Z_1 and Z_2 is thus $\rho = C/\sqrt{(V_1 V_2)}$.

The results presented in Section 3 are based on test statistics with an exact bivariate normal distribution with known correlation. Asymptotically, these results will be valid in the case of bivariate binary data, and to evaluate their accuracy for moderate and small sample sizes, a simulation investigation was conducted. The starting point was a model based on data reported in Bolland *et al.* [25]. That paper concerned results from 1372 patients who suffered from stroke and were randomised between an experimental

Table II. The power of the testing procedures for various values of θ_1, θ_2 and ρ.

ρ	θ_1	θ_2	Restricted Bonferroni	Simes	Restricted Simes	Combined	Consonant combined	Restricted consonant combined
0.2	0	3	0.433	0.779	0.433	0.491	0.566	0.418
		4	0.491	0.961	0.491	0.733	0.797	0.487
		5	0.500	0.997	0.500	0.898	0.931	0.499
	1	3	0.720	0.794	0.720	0.733	0.767	0.718
		4	0.824	0.963	0.824	0.898	0.926	0.821
		5	0.840	0.997	0.840	0.972	0.983	0.840
	2	3	0.869	0.853	0.870	0.898	0.890	0.884
		4	0.959	0.972	0.960	0.972	0.976	0.962
		5	0.976	0.998	0.976	0.995	0.997	0.976
	3	3	0.952	0.938	0.953	0.972	0.963	0.962
		4	0.990	0.987	0.990	0.995	0.993	0.992
		5	0.998	0.999	0.998	0.999	0.999	0.998
	4	4	0.998	0.997	0.998	0.999	0.999	0.999
		5	1	1	1	1	1	1
	5	5	1	1	1	1	1	1
0.5	0	3	0.455	0.782	0.454	0.410	0.443	0.385
		4	0.496	0.962	0.496	0.637	0.668	0.477
		5	0.500	0.997	0.500	0.823	0.844	0.498
	1	3	0.714	0.787	0.714	0.637	0.662	0.651
		4	0.826	0.962	0.826	0.823	0.844	0.793
		5	0.841	0.997	0.841	0.934	0.944	0.836
	2	3	0.821	0.823	0.824	0.823	0.827	0.827
		4	0.951	0.966	0.951	0.934	0.942	0.937
		5	0.976	0.997	0.976	0.981	0.985	0.971
	3	3	0.909	0.908	0.913	0.934	0.930	0.930
		4	0.978	0.978	0.979	0.981	0.982	0.982
		5	0.997	0.998	0.997	0.996	0.997	0.996
	4	4	0.993	0.993	0.993	0.996	0.996	0.996
		5	0.999	0.999	0.999	0.999	0.999	0.999
	5	5	1	1	1	1	1	1
0.8	0	3	0.490	0.796	0.489	0.352	0.359	0.350
		4	0.500	0.966	0.500	0.559	0.566	0.475
		5	0.500	0.998	0.500	0.750	0.756	0.499
	1	3	0.754	0.797	0.750	0.559	0.565	0.565
		4	0.838	0.966	0.837	0.750	0.756	0.748
		5	0.841	0.998	0.841	0.885	0.889	0.831
	2	3	0.807	0.806	0.805	0.750	0.754	0.754
		4	0.957	0.966	0.956	0.885	0.889	0.888
		5	0.977	0.998	0.977	0.958	0.960	0.957
	3	3	0.872	0.875	0.875	0.885	0.885	0.885
		4	0.970	0.970	0.970	0.958	0.959	0.959
		5	0.997	0.998	0.997	0.988	0.988	0.988
	4	4	0.985	0.986	0.986	0.988	0.988	0.988
		5	0.998	0.998	0.998	0.997	0.997	0.997
	5	5	0.999	0.999	0.999	1	1	1

treatment (the drug citcoline) and a control treatment (placebo). Assessments made 12 weeks after randomisation included their scores according to the modified Rankin scale (mRS) and the National Institutes of Health Stroke Scale (NIHSS). These were dichotomised as success (mRS ≤ 1, NIHSS ≤ 1) or failure. Table IV presents the proportions of subjects in Table III of Bolland *et al.*

Table III. Results from a comparison of two treatments based on a single binary endpoint.

Endpoint	Experimental	Control	Total
Success	S_E	S_C	S
Failure	F_E	F_C	F
Total	n_E	n_C	n

Table IV. Probabilities of each joint outcome from trials of citicoline.

		NIHSS			
		Experimental		Control	
		Success	Failure	Success	Failure
mRS	Success	0.1610	0.0634	0.1286	0.060
	Failure	0.0532	0.7224	0.0464	0.765

with each joint outcome on mRS and NIHSS. In this paper, only the results from the control group will be used: those for the experimental group are shown for completeness. The simulations that follow are made under a global null hypothesis in which all subjects followed the joint distribution given for the control subjects in Table IV.

In each simulated trial, patients were randomised in a 1:1 ratio, and seven total sample sizes $n = 24$, 50, 100, 200, 500, 1000 and 2000 were investigated. Each scenario was replicated 100 000 times with the exception of $n = 100$ where 1 000 000 replications were run. For each simulated dataset, an estimate of the correlation, ρ, between the score statistics for mRS and NIHSS was found, and using this value as if it were the truth, a critical region was derived using a search procedure as described in the previous section. The null hypotheses $H_{0,12}$, $H_{0,1}$ and $H_{0,2}$ were then tested using each of the procedures under consideration. The proportions of rejections of $H_{0,12}$, of ($H_{0,12}$ and $H_{0,1}$) and of ($H_{0,12}$ and $H_{0,2}$) are given in Table V, together with the average of the estimates of ρ over all runs.

Despite the high correlation between the two measures of patient outcome, there remains a small chance of observing a positive outcome on one measure and a negative outcome on the

Table VI. Probabilities of each joint outcome for simulation under the alternative.

		NIHSS			
		Experimental		Control	
		Success	Failure	Success	Failure
mRS	Success	0.1268	0.0732	0.0172	0.1828
	Failure	0.2555	0.5445	0.0005	0.7995

Table V. Proportions of rejections in 100 000 (*1 000 000 for $n = 100$) replicate simulations under the global null scenario.

	n	24	50	100*	200	500	1000	2000
	Average estimated	0.6321	0.6396	0.6417	0.6427	0.6430	0.6432	0.6432
Method	Proportion $\hat{\rho} < 0$	0.0213	0.0010	0.0000	0	0	0	0
Restricted Bonferroni	Reject $H_{0,12}$	0.0166	0.0227	0.0246	0.0253	0.0255	0.0249	0.0249
	Reject $H_{0,1}$ and $H_{0,12}$	0.0093	0.0130	0.0140	0.0144	0.0144	0.0141	0.0144
	Reject $H_{0,2}$ and $H_{0,12}$	0.0101	0.0133	0.0141	0.0146	0.0145	0.0144	0.0139
Simes	Reject $H_{0,12}$	0.0164	0.0230	0.0242	0.0252	0.0256	0.0246	0.0252
	Reject $H_{0,1}$ and $H_{0,12}$	0.0107	0.0147	0.0155	0.0161	0.0164	0.0157	0.0165
	Reject $H_{0,2}$ and $H_{0,12}$	0.0112	0.0151	0.0156	0.0162	0.0164	0.0159	0.0160
Restricted Simes	Reject $H_{0,12}$	0.0170	0.0230	0.0245	0.0252	0.0255	0.0247	0.0252
	Reject $H_{0,1}$ and $H_{0,12}$	0.0109	0.0146	0.0157	0.0162	0.0163	0.0157	0.0164
	Reject $H_{0,2}$ and $H_{0,12}$	0.0116	0.0151	0.0158	0.0162	0.0164	0.0160	0.0161
Combined	Reject $H_{0,12}$	0.0197	0.0235	0.0245	0.0254	0.0250	0.0244	0.0251
	Reject $H_{0,1}$ and $H_{0,12}$	0.0111	0.0137	0.0147	0.0155	0.0153	0.0147	0.0153
	Reject $H_{0,2}$ and $H_{0,12}$	0.0116	0.0145	0.0150	0.0151	0.0153	0.0151	0.0154
Consonant combined	Reject $H_{0,12}$	0.0189	0.0234	0.0246	0.0252	0.0251	0.0244	0.0252
	Reject $H_{0,1}$ and $H_{0,12}$	0.0119	0.0147	0.0156	0.0164	0.0162	0.0155	0.0162
	Reject $H_{0,2}$ and $H_{0,12}$	0.0124	0.0154	0.0159	0.0160	0.0161	0.0160	0.0163
Restricted consonant combined	Reject $H_{0,12}$	0.0189	0.0234	0.0246	0.0252	0.0251	0.0244	0.0252
	Reject $H_{0,1}$ and $H_{0,12}$	0.0119	0.0147	0.0156	0.0164	0.0162	0.0155	0.0162
	Reject $H_{0,2}$ and $H_{0,12}$	0.0125	0.0154	0.0159	0.0160	0.0161	0.0160	0.0163
Original Bonferroni	Reject $H_{0,12}$	0.0140	0.0201	0.0209	0.0225	0.0226	0.0219	0.0219
	Reject $H_{0,1}$ and $H_{,12}$	0.0077	0.0112	0.0117	0.0128	0.0127	0.0125	0.0127
	Reject $H_{0,2}$ and $H_{0,12}$	0.0084	0.0117	0.0118	0.0128	0.0127	0.0124	0.0120
Original Simes	Reject $H_{0,12}$	0.0156	0.0214	0.0222	0.0236	0.0239	0.0229	0.0234
	Reject $H_{0,1}$ and $H_{0,12}$	0.0103	0.0139	0.0145	0.0153	0.0155	0.0149	0.0155
	Reject $H_{0,2}$ and $H_{0,12}$	0.0108	0.0143	0.0146	0.0154	0.0156	0.0150	0.0151

other, leading to a negative estimate of correlation. This happens infrequently (the proportions of runs involved indicated in the 'Proportion $\hat{\rho} < 0$' row) except when $n = 24$ or when the overall success rate is small. When the estimated correlation is negative, the critical region for this simulation was found as if two outcomes are uncorrelated. The restricted consonant combined test is conservative in this situation, as it cannot be adjusted by lowering the critical region to achieve the desired type I error when ρ is very small or 0. It is possible for there to be zero information ($V = 0$) about the treatment effect from one of the two outcomes in a small sample size study; for example, all the enrolled patients succeed or else all fail. In this case, a treatment effect cannot be claimed for this measure, neither can a combined treatment effect be claimed, and so we do not reject $H_{0,i}$ or $H_{0,12}$. The consonant combined and restricted consonant combined tests give very similar results, showing differences only when the sample

Table VII. Probabilities of each joint outcome used in the simulation study.

ρ	θ_1	θ_2	n	$\bar{p}$	mRS	NIHSS			
						Experimental		Control	
						Success	Failure	Success	Failure
0.2		5	50	0.2	Success	0.1268	0.0732	0.0172	0.1828
					Failure	0.2555	0.5445	0.0001	0.7995
			50	0.6	Success	0.5726	0.0274	0.2434	0.3566
					Failure	0.3171	0.0829	0.0669	0.3331
			100	0.2	Success	0.1162	0.0838	0.0278	0.1722
					Failure	0.2403	0.5597	0.0157	0.7843
			100	0.6	Success	0.5362	0.0638	0.2798	0.3202
					Failure	0.2870	0.1130	0.0970	0.3030
			500	0.2	Success	0.0944	0.1056	0.0496	0.1504
					Failure	0.1904	0.6096	0.0656	0.7344
			500	0.6	Success	0.4702	0.1298	0.3458	0.2542
					Failure	0.2372	0.1628	0.1468	0.2532
			2000	0.2	Success	0.0835	0.1165	0.0605	0.1395
					Failure	0.1606	0.6394	0.0954	0.7046
			2000	0.6	Success	0.4396	0.1604	0.3764	0.2236
					Failure	0.2149	0.1851	0.1691	0.2309
			500	0.03	Success	0.0107	0.0193	0.0027	0.0273
					Failure	0.0450	0.9250	0.0016	0.9684
0.5	2	4	50	0.2	Success	0.2128	0.0913	0.0272	0.0687
					Failure	0.1541	0.5419	0.0059	0.8981
			50	0.6	Success	0.6911	0.0432	0.2689	0.1968
					Failure	0.1552	0.1104	0.0848	0.4496
			100	0.2	Success	0.1937	0.0830	0.0463	0.0770
					Failure	0.1422	0.5811	0.0178	0.8589
			100	0.6	Success	0.6351	0.0614	0.3249	0.1786
					Failure	0.1494	0.1542	0.0906	0.4058
			500	0.2	Success	0.1566	0.0789	0.0834	0.0811
					Failure	0.1126	0.6519	0.0474	0.7881
			500	0.6	Success	0.5517	0.0920	0.4083	0.1480
					Failure	0.1349	0.2214	0.1051	0.3386
			2000	0.2	Success	0.1387	0.0792	0.1013	0.0808
					Failure	0.0968	0.6853	0.0632	0.7547
			2000	0.6	Success	0.5161	0.1058	0.4439	0.1342
					Failure	0.1276	0.2505	0.1124	0.3095
			500	0.03	Success	0.0256	0.0185	0.0053	0.0106
					Failure	0.0276	0.9283	0.0015	0.9826

size is small ($n \leq 100$) as then estimated correlations could sometimes be zero or negative. Also shown in the table are results from the original, unmodified versions of the Bonferroni and Simes procedures, in which k is set at the value 2.2414, for comparison. As intended, these procedures are conservative in their preservation of the type I error rate.

The average correlation between Z_1 and Z_2 for the simulation results shown in Table V is 0.643. Both original Bonferroni test and original Simes are conservative, with null rejection rates for $H_{0,12}$ strictly less than 0.025. For the remaining tests, type I error rates are close to 0.025, except when $n = 24$ when they drop below 0.02. It would appear that in realistic situations, it could be safe to plug in an estimated correlation and then make use of that value as if it were the true correlation. The proportion of runs in which the global null hypothesis and one of the marginal null hypotheses is rejected reaches around 0.015 for the large sample sizes.

By simulating under the global null hypothesis, it has been demonstrated that the probability of rejecting $H_{0,12}$ is close to the target type I error rate of 0.025. Simulations have also been conducted under alternative hypotheses in order to check predictions of power. Returning to Table II, we find asymptotic values for power when $V_1 = V_2 = 1$, $\rho = 0.2$, 0.5 or 0.8 and θ_1 and θ_2 range between 0 and 6, and we now construct binary scenarios that lead to test statistics with approximately normal properties that correspond to some of the rows in Table II. This allows us to check simulated powers against theoretical values.

To illustrate how the binary scenarios are constructed, suppose that the outcome probabilities are as shown in Table VI. Marginal success rates according to mRS are 0.2 on both experimental and control so that $\theta_1 = 0$. Marginal success rates according to NIHSS are 0.3823 on experimental and 0.0177 on control so that $\theta_2 = 3.5365$. For both scales, the average success rate over the two treatment groups is 0.2. Suppose that responses on $n = 50$ patients, equally divided between experimental and control, are observed. Under this model, the score statistic Z_1 will approximately follow the $N(0, V_1)$ distribution and Z_2 the $N(\theta_2 V_2, V_2)$ distribution, where $V_1 = V_2 \approx (1/4) \times 50 \times 0.2 \times 0.8 = 2$. Now consider the transformation $Z_i^* = Z_i/\sqrt{2}$ so that $Z_i^* \sim N(\theta_i^* V_i^*, V_i^*)$, where $V_i^* = V_i/2$ and $\theta_i^* = \theta_i\sqrt{2}$, $i = 1, 2$. Then $V_1^* = V_2^* = 1$, $\theta_1^* = 0$ and $\theta_2^* = 3.5365 \times \sqrt{2} = 5$. The correlation between Z_1 and Z_2 is given by $\rho \approx (\bar{p}_{12} - \bar{p}_1\bar{p}_2)/\sqrt{\{\bar{p}_1\bar{p}_2(1-\bar{p}_1)(1-\bar{p}_2)\}}$, where $\bar{p}_1$ and $\bar{p}_2$ are the probabilities of success according to mRS and NIHSS, respectively, and $\bar{p}_{12}$ is the probability of success according to both mRS and NIHSS simultaneously, averaged over the two treatments. Hence, $\rho = (0.072 - 0.2 \times 0.2)/\{0.2(1-0.2)\} = 0.2$. Notice that the transformation to Z_i^* is considered only to demonstrate that the properties of the testing procedures should achieve the values shown in the third row of Table II. It is not used in the testing procedures themselves, which are carried out using only the available data in the manner described in Section 3 with ρ replaced by its estimate given by the variance and covariance estimates of equations (2) and (3) of this section.

Table VIII(a). Proportions of rejections in 100 000 replicate simulations under the scenarios specified in Table VII, $\theta_1 = 0$; $\theta_2 = 5$; $\rho = 0.2$.

	n	50	50	100	100	500	500	2000	2000
	$\bar{p}$	0.2	0.6	0.2	0.6	0.2	0.6	0.2	0.6
	Average estimated ρ	0.2064	0.2054	0.2007	0.2008	0.1998	0.1998	0.1998	0.2000
	Proportion $\hat{\rho} < 0$	0.1079	0.0798	0.0388	0.0228	0.0000	0	0	0
Restricted Bonferroni	Reject $H_{0,12}$	0.4727	0.4864	0.4915	0.4980	0.4971	0.4971	0.5023	0.4964
0.500	Reject $H_{0,1}$ and $H_{0,12}$	0.4724	0.4864	0.4915	0.4980	0.4971	0.4970	0.5023	0.4964
	Reject $H_{0,2}$ and $H_{0,12}$	0.0145	0.0161	0.0168	0.0169	0.0188	0.0181	0.0188	0.0184
Simes	Reject $H_{0,12}$	0.9193	0.9890	0.9781	0.9941	0.9956	0.9965	0.9968	0.9973
0.997	Reject $H_{0,1}$ and $H_{0,12}$	0.9192	0.9890	0.9781	0.9941	0.9956	0.9965	0.9968	0.9973
	Reject $H_{0,2}$ and $H_{0,12}$	0.0240	0.0252	0.0252	0.0252	0.0255	0.0248	0.0251	0.0244
Restricted Simes	Reject $H_{0,12}$	0.4725	0.4864	0.4915	0.4980	0.4971	0.4971	0.5023	0.4964
0.500	Reject $H_{0,1}$ and $H_{0,12}$	0.4724	0.4864	0.4915	0.4980	0.4971	0.4970	0.5023	0.4964
	Reject $H_{0,2}$ and $H_{0,12}$	0.0241	0.0252	0.0252	0.0252	0.0255	0.0248	0.0251	0.0244
Combined	Reject $H_{0,12}$	0.5514	0.7906	0.7362	0.8493	0.8697	0.8876	0.8889	0.8932
0.898	Reject $H_{0,1}$ and $H_{0,12}$	0.5507	0.7905	0.7360	0.8493	0.8696	0.8876	0.8888	0.8932
	Reject $H_{0,2}$ and $H_{0,12}$	0.0241	0.0252	0.0252	0.0252	0.0255	0.0248	0.0252	0.0244
Consonant combined	Reject $H_{0,12}$	0.6501	0.8537	0.8100	0.8975	0.9104	0.9232	0.9245	0.9285
0.931	Reject $H_{0,1}$ and $H_{0,12}$	0.6501	0.8537	0.8100	0.8975	0.9104	0.9232	0.9245	0.9285
	Reject $H_{0,2}$ and $H_{0,12}$	0.0242	0.0252	0.0252	0.0252	0.0255	0.0248	0.0252	0.0244
Rest consonant combined	Reject $H_{0,12}$	0.4500	0.4821	0.4859	0.4961	0.4963	0.4964	0.5017	0.4959
0.499	Reject $H_{0,1}$ and $H_{0,12}$	0.4498	0.4821	0.4859	0.4961	0.4963	0.4963	0.5016	0.4959
	Reject $H_{0,2}$ and $H_{0,12}$	0.0242	0.0252	0.0252	0.0252	0.0255	0.0248	0.0252	0.0244
Original Bonferroni	Reject $H_{0,12}$	0.9184	0.9887	0.9780	0.9940	0.9955	0.9964	0.9967	0.9973
	Reject $H_{0,1}$ and $H_{0,12}$	0.9182	0.9886	0.9780	0.9940	0.9955	0.9964	0.9967	0.9972
	Reject $H_{0,2}$ and $H_{0,12}$	0.0109	0.0124	0.0120	0.0119	0.0129	0.0123	0.0126	0.0126
Original Simes	Reject $H_{0,12}$	0.9186	0.9887	0.9781	0.9940	0.9956	0.9964	0.9967	0.9973
	Reject $H_{0,1}$ and $H_{0,12}$	0.9185	0.9887	0.9781	0.9940	0.9956	0.9964	0.9967	0.9973
	Reject $H_{0,2}$ and $H_{0,12}$	0.0240	0.0252	0.0252	0.0252	0.0255	0.0248	0.0251	0.0244

The values in Table VI have been chosen to convert the score statistics into scaled versions of the case depicted in the third row of Table II. Consequently, the powers of the restricted Bonferroni, Simes, restricted Simes, combined, consonant combined and restricted consonant combined tests should take the values predicted in that table: namely 0.500, 0.997, 0.500, 0.898, 0.931 and 0.499, respectively. Simulations based on 100 000 replicates led to rejection of $H_{0,12}$ using these tests in the following proportions of runs: 0.472, 0.894, 0.471, 0.550, 0.650 and 0.449. In those simulations, the average estimated correlation between score statistics was 0.2062.

Extensive simulations along these lines have been carried out on the basis of the rows of Table II in which (a) $\rho = 0.2$, $\theta_1 = 0$, $\theta_2 = 5$ and (b) $\rho = 0.5$, $\theta_1 = 2$, $\theta_2 = 4$. Scenarios leading to scaled versions of these situations, for various samples sizes, are shown in Table VII: these have been constructed with $\bar{p} = 0.2$ or 0.6. Rejection rates for these scenarios are given in Table VIII(a). As for the simulations under the alternative hypothesis, there is a small chance of observing a negative estimate of correlation, and the proportion of runs in which this happened is indicated in the table: it was more frequent for small sample sizes.

In Table VIII(a), below the name of each test, the theoretical power for rejecting the global null hypothesis $H_{0,12}$ is written in bold. For Table VIII(a), these values are taken from the third row of Table II ($\rho = 0.2$, $\theta_1 = 0$, $\theta_2 = 5$), and for Table VIII(b) from the 23rd row of Table II ($\rho = 0.5$, $\theta_1 = 2$, $\theta_2 = 4$). It can be seen that in general, these target powers are achieved for sample sizes of 500 and 2000, except when the overall success rate is 0.03. Powers for sample sizes of 50 are generally poor, whereas those for 100 are better, but still substantially below the higher target values. The alternative values chosen in Table VIII(a) are such that restricted procedures are intended to lose power, as one marginal null is true. For the nonrestricted procedures, it is unusual for the closed-test procedure to fail to reject a null hypothesis that would be rejected by a naive marginal test, but this is very common for the restricted procedures. The greater theoretical power for rejecting the global null possessed by the Simes and the consonant combined tests are reflected in the good performance of these methods across the sample sizes, with the Simes test performing particularly well even in the smaller sample sizes. The alternative chosen for Table VIII(b) is one in which restricted and nonrestricted tests should achieve similar powers. This is borne out in the simulation results. The restricted methods always show lower power than their nonrestricted counterparts, but this reflects their greater conservatism under the global null hypothesis. The alternative does not involve equal values for θ_1 and θ_2, and so it is not the ideal situation for a combined test. Table II shows that the combined test has better theoretical properties than the Simes test or the other approaches when $\theta_1 = \theta_2 > 0$, and Table VIII(a) shows that in moderate to large samples of binary data those properties are attained. Thus, it can be anticipated that simulations under alternatives with equal treatment effects would favour the combined approach. Also shown in Tables VIII(a) and VIII(b) are results from the original, unmodified versions of the

Table VIII(b). Proportions of rejections in 100 000 replicate simulations under the scenarios specified in Table VII, $\theta_1 = 2$; $\theta_2 = 4$; $\rho = 0.6$.

	n	50	50	100	100	500	500	2000	2000
	$\bar{p}$	0.2	0.6	0.2	0.6	0.2	0.6	0.2	0.6
	Average estimated ρ	0.4965	0.4996	0.4987	0.5000	0.4996	0.5002	0.5000	0.4999
	Proportion with $\hat{\rho} < 0$	0.0031	0.0001	0.0000	0	0	0	0	0
Restricted Bonferroni	Reject $H_{0,12}$	0.8489	0.9285	0.9116	0.9405	0.9444	0.9490	0.9495	0.9522
0.951	Reject $H_{0,1}$ and $H_{0,12}$	0.8319	0.9222	0.9045	0.9365	0.9413	0.9464	0.9472	0.9497
	Reject $H_{0,2}$ and $H_{0,12}$	0.3578	0.4072	0.3979	0.4187	0.4241	0.4248	0.4272	0.4314
Simes	Reject $H_{0,12}$	0.8574	0.9427	0.9214	0.9542	0.9589	0.9627	0.9642	0.9665
0.966	Reject $H_{0,1}$ and $H_{0,12}$	0.8489	0.9399	0.9181	0.9524	0.9575	0.9614	0.9631	0.9653
	Reject $H_{0,2}$ and $H_{0,12}$	0.4444	0.4897	0.4829	0.5030	0.5101	0.5111	0.5154	0.5160
Restricted Simes	Reject $H_{0,12}$	0.8516	0.9287	0.9117	0.9406	0.9447	0.9491	0.9497	0.9524
0.951	Reject $H_{0,1}$ and $H_{0,12}$	0.8423	0.9256	0.9081	0.9385	0.9431	0.9478	0.9485	0.9511
	Reject $H_{0,2}$ and $H_{0,12}$	0.4452	0.4900	0.4832	0.5032	0.5102	0.5112	0.5155	0.5161
Combined	Reject $H_{0,12}$	0.8383	0.9122	0.8903	0.9237	0.9253	0.9309	0.9333	0.9341
0.934	Reject $H_{0,1}$ and $H_{0,12}$	0.8163	0.9050	0.8814	0.9185	0.9215	0.9272	0.9300	0.9308
	Reject $H_{0,2}$ and $H_{0,12}$	0.4487	0.4913	0.4851	0.5043	0.5109	0.5119	0.5162	0.5166
Consonant combined	Reject $H_{0,12}$	0.8498	0.9234	0.9026	0.9336	0.9345	0.9398	0.9413	0.9424
0.942	Reject $H_{0,1}$ and $H_{0,12}$	0.8366	0.9187	0.8968	0.9304	0.9322	0.9377	0.9394	0.9405
	Reject $H_{0,2}$ and $H_{0,12}$	0.4491	0.4915	0.4853	0.5043	0.5110	0.5120	0.5162	0.5167
Rest consonant combined	Reject $H_{0,12}$	0.8464	0.9169	0.8989	0.9278	0.9300	0.9356	0.9365	0.9378
0.937	Reject $H_{0,1}$ and $H_{0,12}$	0.8332	0.9122	0.8931	0.9246	0.9277	0.9334	0.9346	0.9360
	Reject $H_{0,2}$ and $H_{0,12}$	0.4491	0.4915	0.4853	0.5043	0.5110	0.5120	0.5162	0.5167
Original Bonferroni	Reject $H_{0,12}$	0.8450	0.9374	0.9157	0.9509	0.9557	0.9597	0.9615	0.9639
	Reject $H_{0,1}$ and $H_{0,12}$	0.8279	0.9309	0.9083	0.9468	0.9524	0.9568	0.9590	0.9613
	Reject $H_{0,2}$ and $H_{0,12}$	0.3300	0.3760	0.3668	0.3899	0.3986	0.4000	0.4029	0.4062
Original Simes	Reject $H_{0,12}$	0.8520	0.9401	0.9192	0.9528	0.9573	0.9610	0.9629	0.9651
	Reject $H_{0,1}$ and $H_{0,12}$	0.8441	0.9375	0.9161	0.9510	0.9560	0.9598	0.9619	0.9640
	Reject $H_{0,2}$ and $H_{0,12}$	0.4438	0.4895	0.4827	0.5030	0.5100	0.5111	0.5154	0.5160

Bonferroni and Simes procedures. It was seen in Table V that these approaches gave more away in terms of conservatism of type I error rates than their modified counterparts. Despite this finding, the gain in power due to the modification is slight for the range of values explored here, throughout the range of sample sizes considered.

5. CONCLUSIONS

In this paper, we explore the following strategy for testing multiple hypotheses. Devise an exact procedure for the case in which test statistics are exactly normally distributed with known correlations. Base that actual tests on approximately normal test statistics, and substitute the estimated correlations between them as if they were known values, making no allowance for their estimation. Six testing procedures were evaluated in the case of bivariate tests. The principal differences lay between the restricted tests, in which rejection of the global test statistic is allowed only if there are at least positive trends towards each of the alternative hypothesis of a positive treatment effect. Imposition of such a restriction has a marked effect on the power surface, when plotted against the two actual treatment effects. Power is reduced when one treatment effect is small or zero and enhanced when the treatment effects are both positive and roughly equal. However, in well powered studies, the power in this region will already be large, and so the gain in applying a restricted test will be insubstantial. In practice, restricted testing procedures would not be chosen to increase power but because of a scientific belief that a negative effect on one measure of treatment advantage cannot be outweighed by a large advantage in the other. When such considerations apply, the messages of this paper are that an accurate test procedure can be carried out and that for moderate to large sample sizes, little power will be lost in the alternative regions of interest.

The strategy of conducting tests that depend on the correlation between test statistics, and using the estimated correlation without allowance for that estimation, has been found to be valid. Type I error rates are closely achieved in large sample sizes and conservatively bounded in smaller sample sizes or when success rates are small. The choice between a Simes approach and a combined approach depends on the alternative that is anticipated: if the effect sizes θ_1 and θ_2 are likely to be equal to one another, then the combined tests have the greater powers. However, the advantages seen relative to conventional procedures based on Bonferroni or Simes tests are modest. It should be added that these conclusions are based on limited explorations of bivariate binary tests. Although it would appear to be more satisfying to base tests on correlation values seen in the data, it must be admitted that although no invalidity appears to occur, the gains in doing so are only small.

REFERENCES

[1] Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986; **73**:751–754.
[2] Hochberg Y, Tamhane AC. *Multiple Comparison Procedures*. Wiley: New York, 1987.
[3] Hsu JC. *Multiple Comparisons: Theory and Methods*. Chapman & Hall: London, 1996.
[4] Bretz F, Maurer W, Brannath W, Posch M. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:586–604.
[5] O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics* 1984; **40**:1079–1087.
[6] Kudô A. A multivariate analog of the one-sided test. *Biometrika* 1963; **50**:403–418.
[7] Perlman MD. One-sided problems in multivariate analysis. *Annals of Mathematical Statistics* 1969; **40**:549–567.
[8] Silvapulle MJ, Sen PK. *Constrained Statistical Inference: Inequality, Order and Shape Restrictions*. Wiley: New York, 2005.
[9] Läuter J. Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* 1996; **52**:964–970.
[10] Follmann DA. Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine* 1995; **14**:1163–1175.
[11] Follmann DA. A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* 1996; **91**:854–861.
[12] Silvapulle MJ. A curious example involving the likelihood ratio test against one-sided hypotheses. *American Statistician* 1997; **51**: 178–180.
[13] Glimm E, Läuter J. Directional multivariate tests rejecting null and negative effects in all variables. *Biometrical Journal* 2010; **52**: 757–770.
[14] Senn SJ, Bretz F. Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics* 2007; **6**:161–170.
[15] Bretz F, Hothorn T, Westfall P. *Multiple Comparison Procedures Using R*. Chapman & Hall/CRC: Boca Raton, 2010.
[16] Pocock SJ, Geller NL, Tsiatis AA. The analysis of multiple endpoints in clinical trials. *Biometrics* 1987; **43**:487–498.
[17] Todd S. An adaptive approach to implementing bivariate group sequential clinical trial designs. *Journal of Biopharmaceutical Statistics* 2003; **13**:605–619.
[18] Whitehead J, Branson M, Todd S. A combined score test for binary and ordinal endpoints from clinical trials. *Statistics in Medicine* 2010; **29**:521–532.
[19] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.
[20] Brannath W, Bretz F, Maurer W, Sarkar S. Trimmed weighted Simes' test for two one-sided hypotheses with arbitrarily correlated test statistics. *Biometrical Journal* 2009; **51**:885–898.
[21] Gabriel KR. Simultaneous test procedures – some theory of multiple comparisons. *Annals of Mathematical Statistics* 1969; **40**:224–250.
[22] Bittman RM, Romano JP, Vallarino C, Wolf M. Optimal testing of multiple hypotheses with common effect direction. *Biometrika* 2009; **96**:399–410.
[23] Dunnett CW, Tamhane AC. A step-up multiple test procedure. *Journal of the American Statistical Association* 1992; **87**:162–170.
[24] Whitehead J. *The Design and Analysis of Sequential Clinical Trials* (Revised 2nd edn). Wiley: Chichester, 1997.
[25] Bolland K, Whitehead J, Cobo E, Secades JJ. Evaluation of a sequential global test of improved recovery following stroke as applied to the ICTUS trial of citicoline. *Pharmaceutical Statistics* 2009; **8**: 136–149.

Biometrical Journal **42** (2000) 8, 937–950

Spherical Tests in Balanced Multivariate Mixed Models

EKKEHARD GLIMM

Department of Statistics
University of Toronto
Toronto, Ontario
Canada

Summary

This paper deals with the application of spherical tests in balanced multivariate mixed models. A general representation of the univariate mixed model, given by HOCKING (1985), is adapted to the multivariate case and it is demonstrated how spherical tests, introduced by LÄUTER (1996), can be applied to test hypotheses about the covariance structure and the means. The methods are illustrated by an example. A simulation experiment shows the superiority of spherical tests over traditional methods, if the multivariate data arise from a factor structure.

Key words: Spherical tests; Balanced mixed models; Variance components.

1. Introduction

Some years ago, Läuter has introduced the concept of spherical tests for multivariate linear hypotheses (LÄUTER, 1996). The idea was first introduced for tests of the mean vectors from multivariate normal distributions. Subsequently, Läuter and co-workers have extended the concept with respect to several aspects (LÄUTER, GLIMM, and KROPF, 1996, 1998; KROPF, LÄUTER, and GLIMM, 1997; KROPF and GLIMM, 1996; KROPF and LÄUTER, 2000). The basic concept has also been picked up by other researchers. FANG et al. (1998), FANG and LIANG (1999), LIANG and FANG (2000) and LIANG et al. (2000) have used it to derive tests of multivariate normality. Their techniques can be regarded as nonparametic spherical methods. The theoretical work by LÄUTER, GLIMM, and KROPF (1998) provides a general framework for the construction of tests for a multitude of statistical models. Not all of these opportunities have been exploited. This paper is intended to demonstrate how spherical tests can be used to test hypotheses in multivariate mixed models. The emphasis is on tests about the covariance structure.

2. Theoretical Foundations

The spherical tests treated in this paper are all based on the following theorem 2.1 which is a slightly simplified special case of theorem 2 by LÄUTER et al. (1998).

Theorem 2.1: *Let*

$$\boldsymbol{X} \sim N_{n\times p}(\boldsymbol{\Theta}, \boldsymbol{I}_n \otimes \boldsymbol{\Phi})$$

be a matrix of n independent observations from p-dimensional normal distributions. Let $\boldsymbol{E}$ *be a fixed* $n \times f$*-matrix with* $\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}_f$ *and* $\boldsymbol{E}'\boldsymbol{\Theta} = \boldsymbol{0}$*. Let the* $p \times q$*-matrix* $\boldsymbol{D}$ *be a Borel function of* $\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}$*, where* $\boldsymbol{P}$ *is an idempotent, symmetric* $n \times n$*-matrix with* $\boldsymbol{P}\boldsymbol{E} = \boldsymbol{E}$*,* $q \leq \min(p, f)$ *and* $\boldsymbol{D}$ *such that* $\operatorname{rank}(\boldsymbol{E}'\boldsymbol{X}\boldsymbol{D}) = q$ *with probability 1. Let* $F_0(\boldsymbol{Z})$ *be a Borel function defined for all full-rank* $f \times q$*-matrices* $\boldsymbol{Z}$*. Finally, let* $\boldsymbol{Y} = \boldsymbol{E}'\boldsymbol{X}\boldsymbol{D}$ *and* $F(\boldsymbol{X}) = F_0(\boldsymbol{Y} \cdot (\sqrt{\boldsymbol{Y}'\boldsymbol{Y}})^{-1})$*, where* $\sqrt{\boldsymbol{A}}$ *is any root of the non-singular* $q \times q$*-matrix* $\boldsymbol{A}$*, i.e.* $\sqrt{\boldsymbol{A}}$ *is a* $q \times q$*-matrix with* $\sqrt{\boldsymbol{A}}'\sqrt{\boldsymbol{A}} = \boldsymbol{A}$*.*

Then the distribution of $F(\boldsymbol{X})$ *does not depend on* p, $\boldsymbol{D}$, $\boldsymbol{E}$, $\boldsymbol{P}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$.

Theorem 2.1 provides the basis for exact multivariate level-α tests in a variety of situations. The proof of the theorem is a straightforward modification of the proof of theorem 2 given by LÄUTER et al. (1998). Using this theorem, LÄUTER et al. (1996, 1998, 1999) have investigated hypotheses on the mean of normal observations and tests of correlation between normal samples. KROPF and LÄUTER (2000) have derived further tests for the correlation problem, while Fang and Liang (FANG and LIANG, 1999; LIANG et al., 2000) have concentrated on tests for multivariate normality.

Here, we are concerned with observations from balanced multivariate mixed models. These models are straightforward multivariate extensions of well-known univariate mixed models, such as the split-plot design. HOCKING (1985), chapter 9, discusses those univariate models in detail.

The situation investigated here may stem from an experiment with k factors, some of which may be fixed and some of which are random. It is assumed that p-dimensional observations are obtained for each combination of those factors. Let a_i denote the number of levels of the ith factor, $i = 1, \ldots, k$. Complete balance implies that there are n observations on each of the $\prod_{i=1}^{k} a_i$ combinations, so there is a total of $N = n \cdot \prod_{i=1}^{k} a_i$ p-dimensional observations.

Let $\boldsymbol{X} \sim N_{N\times p}(\boldsymbol{\Theta}, \boldsymbol{\Sigma})$ be the $N \times p$-matrix comprising those observations in its rows. The random factors in the experiment cause correlations between some observations and hence imply a certain covariance structure, whereas the fixed effects have a bearing on the mean of different observations. To characterize the statistical models arising from this situation, some notation has to be introduced. First of all, we define the set $T = \{1, \ldots, k, 12, \ldots, (k-1)\,k, 123, \ldots, 123\cdots k\}$ of possible combinations of indices of the k factors in the experiment. Just to avoid confusion, it is assumed that no factor occupies more than a single digit in those combinations, such that 123 stands for factors 1, 2 and 3, not for factors 1 and 23, say. Let $T_0 = T \cup \{0\}$. T_0 represents all the main effects and interaction effects in the balanced mixed model. We further assume that T_1 is the subset of fixed effects, T_2 the subset of random effects, $T_1 \cup T_2 \subset T_0$, $T_1 \cap T_2 = \emptyset$. Let us

denote the overall mean by $\boldsymbol{\mu}$, the effects of the fixed factors by $\mathbf{B}_j$ and the multivariate variance components from the random effects by $\boldsymbol{\Sigma}_i$. Then the models investigated in this paper are characterized by the mean structure

$$\boldsymbol{\Theta} = E(\boldsymbol{X}) = \mathbf{1}_N \boldsymbol{\mu}' + \sum_{j \in T_I} M_j \mathbf{B}_j', \tag{1}$$

and the covariance structure

$$\boldsymbol{\Sigma} = \boldsymbol{I}_N \otimes \boldsymbol{\Sigma}_0 + \sum_{i \in T_2} \boldsymbol{V}_i \otimes \boldsymbol{\Sigma}_i. \tag{2}$$

$\boldsymbol{V}_i$ and M_j are the design matrices of the model. $\boldsymbol{V}_i$, $\boldsymbol{i} \in T_2$ is defined by

$$\boldsymbol{V}_i = \boldsymbol{T}_1^{(i)} \otimes \cdots \boldsymbol{T}_k^{(i)} \otimes \mathbf{1}_n \mathbf{1}_n' \tag{3}$$

with

$$\boldsymbol{T}_l^{(i)} = \begin{cases} \boldsymbol{I}_{a_l} & \textit{if} \quad l \in \boldsymbol{i} \\ \mathbf{1}_{a_l} \mathbf{1}_{a_l}' & \textit{if} \quad l \notin \boldsymbol{i} \end{cases}, \tag{4}$$

where "$l \in \boldsymbol{i}$" is used to indicate that the single digit l appears in the sequence of digits $\boldsymbol{i}$. As an example, if $\boldsymbol{i} = 13 \in T_2$, then $\boldsymbol{V}_{13} = \boldsymbol{I}_{a_1} \otimes \mathbf{1}_{a_2} \mathbf{1}_{a_2}' \otimes \boldsymbol{I}_{a_3} \otimes \mathbf{1}_{a_4} \mathbf{1}_{a_4}' \otimes \cdots \otimes \mathbf{1}_{a_k} \mathbf{1}_{a_k}' \otimes \mathbf{1}_n \mathbf{1}_n'$.

In the same vein, $M_{\boldsymbol{j}} = \boldsymbol{Z}_1^{(j)} \otimes \cdots \otimes \boldsymbol{Z}_k^{(j)} \otimes \mathbf{1}_n$ with

$$\boldsymbol{Z}_l^{(j)} = \begin{cases} \begin{pmatrix} \boldsymbol{I}_{a_l - 1} \\ -\mathbf{1}_{a_l - 1}' \end{pmatrix} & \text{if} \quad l \in \boldsymbol{j} \\ \mathbf{1}_{a_l} & \text{if} \quad l \notin \boldsymbol{j} \end{cases}$$

for $\boldsymbol{j} \in T_1$.

In this derivation, the variance components $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ are generated by random effects in the model and thus are bound to be positive definite. We note in passing that this is not necessarily the only situation from which such a covariance structure may arise and hence the positive definiteness is not an imperative condition. Of course, $\boldsymbol{\Sigma}$ has to be positive definite. An aspect that does not occur in univariate mixed models is the fact that all of the variance components have to be symmetric. This implies complete interchangeability of the p variables. As a characteristic feature of the outlined model, covariance and mean parameters do not interfere. This property is important for the existence of exact tests, as will be illustrated in the following.

Example 2.1: Suppose, p observations are made on each of s randomly selected patients at t time points. It is assumed that in the course of a cure, say, the responses change with time and that each patient has an individual response level, such that observations from the same patient at different times are correlated. However, it is also assumed that this correlation remains the same for any two times. This is called the compound-symmetry structure. In terms of the more gen-

eral model outlined above, we have $\boldsymbol{X} = \left(\boldsymbol{X}_1', \cdots, \boldsymbol{X}_t'\right)'$ with

$$\boldsymbol{X} \sim N_{st\times p}\left(\mathbf{1}_{st}\boldsymbol{\mu}' + \left[\begin{pmatrix} \boldsymbol{I}_{t-1} \\ -\mathbf{1}_{t-1}' \end{pmatrix} \otimes \mathbf{1}_s\right] \mathbf{B}_1, \boldsymbol{I}_{st} \otimes \boldsymbol{\Sigma}_0 + \mathbf{1}_t\mathbf{1}_t' \otimes \boldsymbol{I}_s \otimes \boldsymbol{\Sigma}_2\right).$$

Note that $n = 1$, since there is just one observation per patient and time. Furthermore, $T_1 = \{1\}$ and $T_2 = \{2\}$. □

Tests and estimates of the parameters in the mixed models introduced in this section are based on quadratic forms calculated from the data. To derive the distributions of these quadratic forms, a canonical representation of the covariance structure is needed:

$$\boldsymbol{\Sigma} = \boldsymbol{A}_\mu \otimes \boldsymbol{\Lambda}_\mu + \boldsymbol{A}_0 \otimes \boldsymbol{\Lambda}_0 + \sum_{\boldsymbol{i}\in T} \boldsymbol{A}_{\boldsymbol{i}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{i}},$$

where

$$\boldsymbol{A}_{\boldsymbol{i}} = \boldsymbol{G}_1^{(\boldsymbol{i})} \otimes \cdots \otimes \boldsymbol{G}_k^{(\boldsymbol{i})} \otimes \frac{1}{n}\mathbf{1}_n\mathbf{1}_n',$$

$$\boldsymbol{G}_l^{(\boldsymbol{i})} = \begin{cases} \boldsymbol{I}_{a_l} - \dfrac{1}{a_l}\mathbf{1}_{a_l}\mathbf{1}_{a_l}' & \text{if} \quad l \in \boldsymbol{i} \\ \dfrac{1}{a_l}\mathbf{1}_{a_l}\mathbf{1}_{a_l}' & \text{if} \quad l \notin \boldsymbol{i} \end{cases}, \quad \boldsymbol{i} \in T,$$

$$\boldsymbol{A}_\mu = \frac{1}{N}\cdot\mathbf{1}_N\mathbf{1}_N', \qquad \boldsymbol{A}_0 = \boldsymbol{I}_N - \sum_{\boldsymbol{i}\in T}\boldsymbol{A}_{\boldsymbol{i}} - \boldsymbol{A}_\mu,$$

and

$$\boldsymbol{\Lambda}_0 = \boldsymbol{\Sigma}_0, \qquad \boldsymbol{\Lambda}_{\boldsymbol{i}} = \boldsymbol{\Sigma}_0 + n\cdot\sum_{\boldsymbol{j}\in T, \boldsymbol{i}\subseteq\boldsymbol{j}} a_{\boldsymbol{j}}^*\boldsymbol{\Sigma}_{\boldsymbol{j}}, \qquad \boldsymbol{\Lambda}_\mu = \boldsymbol{\Sigma}_0 + n\cdot\sum_{\boldsymbol{i}\in T} a_{\boldsymbol{i}}^*\boldsymbol{\Sigma}_{\boldsymbol{i}},$$

with $a_{\boldsymbol{i}}^* = \prod_{l\notin\boldsymbol{i}} a_l$, $a_{\boldsymbol{i}}^* = 1$ if $\boldsymbol{i} = 12\cdots k$. For example, with $k = 3$ factors, $\boldsymbol{\Lambda}_{23} = \boldsymbol{\Sigma}_0 + n\cdot a_1\cdot\boldsymbol{\Sigma}_{23} + n\boldsymbol{\Sigma}_{123}$. These expressions are exactly analogue to the ones given by HOCKING (1985), definition 9.3, for the univariate case.

The advantage of this canonical form over the initial representation lies in the projection matrices $\boldsymbol{A}_{\boldsymbol{i}}$, $\boldsymbol{i} \in T$, $\boldsymbol{A}_0$ and $\boldsymbol{A}_\mu$. Since these fulfill $\boldsymbol{A}_{\boldsymbol{i}_1}\boldsymbol{A}_{\boldsymbol{i}_2} = \mathbf{0}$, $\boldsymbol{i}_1 \neq \boldsymbol{i}_2$, a well-known extension of Cochran's theorem (see e.g. AHRENS and LÄUTER, 1981, p. 24, or RAO, 1973, ch. 8b.2, (ii) and (iii), p. 535ff.) can be used to derive independent, unbiased estimates of the canonical variance components $\boldsymbol{\Lambda}_{\boldsymbol{i}}$:

Theorem 2.2:
i) $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}}\boldsymbol{X} \sim W_p(\boldsymbol{\Lambda}_{\boldsymbol{i}}, \operatorname{rank}(\boldsymbol{A}_{\boldsymbol{i}}))$ *for all* $\boldsymbol{i} \in T_0 - T_1$.
ii) $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}_1}\boldsymbol{X}$ *and* $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}_2}\boldsymbol{X}$ *are independent,* $\boldsymbol{i}_1 \neq \boldsymbol{i}_2$; $\boldsymbol{i}_1, \boldsymbol{i}_2 \in T_0$.

This theorem is a straightforward extension of theorems 9.8 and 9.9 by HOCKING (1985).

Example 2.2: [example 2.1 continued] Since $\boldsymbol{\Sigma}_1 = \mathbf{0}$ and $\boldsymbol{\Sigma}_{12} = \mathbf{0}$ by the model definition, we have

$$\boldsymbol{\Lambda}_0 = \boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_{12} = \boldsymbol{\Sigma}_0, \qquad \boldsymbol{\Lambda}_2 = \boldsymbol{\Lambda}_\mu = \boldsymbol{\Sigma}_0 + t \cdot \boldsymbol{\Sigma}_2 ,$$

$$\boldsymbol{A}_0 = \mathbf{0}, \qquad \boldsymbol{A}_1 = \left(\boldsymbol{I}_t - \frac{1}{t}\, \mathbf{1}_t \mathbf{1}_t' \right) \otimes \frac{1}{s}\, \mathbf{1}_s \mathbf{1}_s' ,$$

$$\boldsymbol{A}_2 = \frac{1}{t}\, \mathbf{1}_t \mathbf{1}_t' \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\, \mathbf{1}_s \mathbf{1}_s' \right) ,$$

$$\boldsymbol{A}_{12} = \left(\boldsymbol{I}_t - \frac{1}{t}\, \mathbf{1}_t \mathbf{1}_t' \right) \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\, \mathbf{1}_s \mathbf{1}_s' \right) .$$

Consequently,

$$\boldsymbol{X}' \boldsymbol{A}_2 \boldsymbol{X} \sim W_p(\boldsymbol{\Lambda}_2, s-1), \qquad \boldsymbol{X}' \boldsymbol{A}_{12} \boldsymbol{X} \sim W_p(\boldsymbol{\Lambda}_{12}, (t-1)\,(s-1)) .$$

and these two quadratic forms are stochastically independent. □

3. Spherical Tests

3.1 *Tests of the variance components*

To test the hypothesis that a variance component $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ is zero, one has to select two quadratic forms $\boldsymbol{X}' \boldsymbol{A}_{i_1} \boldsymbol{X}$ and $\boldsymbol{X}' \boldsymbol{A}_{i_2} \boldsymbol{X}$, $\boldsymbol{i}_1, \boldsymbol{i}_2 \in T_0 - T_1$, in such a way that the same covariance parameters occur in their respective Wishart distributions under the null hypothesis, but ones that differ by a multiple of $\boldsymbol{\Sigma}_{\boldsymbol{i}}$ in case of the alternative. Usually, but not always, this is possible. One can then apply one of the usual multivariate tests, e.g. Wilks' Λ with $\Lambda = \dfrac{|\boldsymbol{X}' \boldsymbol{A}_{i_2} \boldsymbol{X}|}{|\boldsymbol{X}' (\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2})\, \boldsymbol{X}|}$ to test the hypothesis. In example 2.2, $\boldsymbol{\Lambda}_{12} = \boldsymbol{\Sigma}_0$ is the covariance parameter corresponding to $\boldsymbol{X}' \boldsymbol{A}_{12} \boldsymbol{X}$ and $\boldsymbol{\Lambda}_2 = \boldsymbol{\Sigma}_0 + t \cdot \boldsymbol{\Sigma}_2$ pertains to $\boldsymbol{X}' \boldsymbol{A}_2 \boldsymbol{X}$, so these are the two quadratic forms suitable for the test of no intra-patient correlation, $H_0 : \boldsymbol{\Sigma}_2 = \mathbf{0}$.

Theorem 2.1 can be applied to the function $F_0(\boldsymbol{Z}) = \left| \boldsymbol{Z}' \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{r_2} \end{pmatrix} \boldsymbol{Z} \right|$ defined for $(r_1 + r_2) \times q$-matrices $\boldsymbol{Z}$ with any pre-selected $q \leq \min(p, r_2)$, where $r_1 = \operatorname{rank}(\boldsymbol{A}_{i_1})$, $r_2 = \operatorname{rank}(\boldsymbol{A}_{i_2})$ and $\mathbf{0}$ denotes appropriately sized matrices of zeros. Let $\boldsymbol{E}$ with $\boldsymbol{E}' = (\boldsymbol{E}_1', \boldsymbol{E}_2')$, where $\boldsymbol{E}_l$ is an $r_l \times N$-matrix, be any root of the (usually singular) $N \times N$-matrix $\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2}$, i.e. $\boldsymbol{E}_1' \boldsymbol{E}_1 = \boldsymbol{A}_{i_1}$, $\boldsymbol{E}_2' \boldsymbol{E}_2 = \boldsymbol{A}_{i_2}$ and $\boldsymbol{E}\boldsymbol{E}' = \boldsymbol{I}_{r_1 + r_2}$, since $\boldsymbol{A}_{i_1}' \boldsymbol{A}_{i_2} = \mathbf{0}$, $\boldsymbol{i}_1 \neq \boldsymbol{i}_2$. The theorem states that $F(\boldsymbol{X}) = \dfrac{|\boldsymbol{D}' \boldsymbol{X}' \boldsymbol{A}_{i_2} \boldsymbol{X} \boldsymbol{D}|}{|\boldsymbol{D}' \boldsymbol{X}' (\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2})\, \boldsymbol{X} \boldsymbol{D}|}$ has the same null distribution as Wilks' Λ, $\Lambda = \dfrac{|\boldsymbol{G}|}{|\boldsymbol{H} + \boldsymbol{G}|}$, for independent quadratic forms $\boldsymbol{H} \sim W_q(\boldsymbol{I}_q, r_1)$ and $\boldsymbol{G} \sim W_q(\boldsymbol{I}_q, r_2)$, if $\boldsymbol{D}$ is any Borel function of $\boldsymbol{X}' (\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2})\, \boldsymbol{X}$.

Example 3.1: [example 2.2 continued] Suppose $q = 1$, such that $\boldsymbol{D}$ is a vector, $\boldsymbol{D} = \boldsymbol{d}$. LÄUTER et al. (1996, 1998) have proposed several rules for obtaining a $\boldsymbol{d}$. One of these is the so-called Principal-Component (PC-)rule: $\boldsymbol{d}$ is the eigenvector corresponding to the largest eigenvalue λ of $\boldsymbol{X}'(\boldsymbol{A}_2 + \boldsymbol{A}_{12})\,\boldsymbol{X}\boldsymbol{d} = \text{Diag}\,(\boldsymbol{X}'(\boldsymbol{A}_2 + \boldsymbol{A}_{12})\,\boldsymbol{X})\,\boldsymbol{d}\lambda$. Then $\dfrac{|\boldsymbol{d}'\boldsymbol{X}'\boldsymbol{A}_2\boldsymbol{X}\boldsymbol{d}|}{|\boldsymbol{d}'\boldsymbol{X}'(\boldsymbol{A}_2 + \boldsymbol{A}_{12})\,\boldsymbol{X}\boldsymbol{d}|}$ has a Wilks' Λ-distribution under H_0 which is equivalent to

$$(t-1)\cdot\frac{\boldsymbol{d}'\boldsymbol{X}'\boldsymbol{A}_2\boldsymbol{X}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{X}'\boldsymbol{A}_{12}\boldsymbol{X}\boldsymbol{d}} \overset{H_0}{\sim} F(s-1,(t-1)\,(s-1))\,. \tag{5}$$

The PC-test of $H_0 : \boldsymbol{\Sigma}_2 = 0$ is based on this statistic. □

Due to theorem 2.1, spherical tests like this one are exact level-α tests. The problem here is to find an appropriate spherical test with satisfactory power properties. The following provides a heuristic argument for the PC test which has proved to be powerful in the context of mean and correlation hypotheses (GLIMM and LÄUTER, 2000).

The argument involves the assumption of a single latent variable, y, say, underlying the p variables in $\boldsymbol{X}$. Suppose that for each of the N observations there is such an underlying, unobserved variable y_i. Summarizing the y_is into the vector $\boldsymbol{y}$, we obtain

$$\boldsymbol{y} \sim N\left(\boldsymbol{1}_N\mu_{(y)} + \sum_{\boldsymbol{j}\in T_1} \boldsymbol{M}_{\boldsymbol{j}}\beta_{(y)\boldsymbol{j}}\,,\ \sigma_0\boldsymbol{I}_N + \sum_{\boldsymbol{i}\in T_2}\sigma_{\boldsymbol{i}}\boldsymbol{V}_{\boldsymbol{i}}\right),$$

$$\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\vartheta}' + \boldsymbol{U}$$

with $\boldsymbol{U} \sim N_{N\times p}(\boldsymbol{0}, \boldsymbol{I}_n \otimes \mathbf{K})$ independent of $\boldsymbol{y}$, where $\boldsymbol{\vartheta}$ is a p-dimensional vector of factor loadings and $\mathbf{K}$ is a positive definite diagonal matrix of individual errors for each of the p variables. These conditions define a one-factor model. The vector $\boldsymbol{y}$ is often called the vector of factor scores. LÄUTER (1992) has examined this model in detail. He has also used it to derive the PC test for hypotheses about the means of multivariate normal observations (LÄUTER, 1996; LÄUTER et al., 1996). The one-factor model implies

$$\boldsymbol{X} \sim N_{N\times p}\left(\boldsymbol{1}_N\mu_{(y)}\boldsymbol{\vartheta}' + \sum_{\boldsymbol{j}\in\boldsymbol{T_I}}\boldsymbol{M}_{\boldsymbol{j}}\beta_{(y)\boldsymbol{j}}\boldsymbol{\vartheta}'\,,\right.$$

$$\left.\boldsymbol{I}_N \otimes (\mathbf{K} + \sigma_0\boldsymbol{\vartheta}\boldsymbol{\vartheta}') + \sum_{\boldsymbol{i}\in T_2}(\boldsymbol{V}_{\boldsymbol{i}} \otimes \sigma_{\boldsymbol{i}}\boldsymbol{\vartheta}\boldsymbol{\vartheta}')\right).$$

We thus have $\boldsymbol{\Sigma}_0 = \mathbf{K} + \sigma_{\boldsymbol{0}}\boldsymbol{\vartheta}\boldsymbol{\vartheta}'$, $\boldsymbol{\Sigma}_{\boldsymbol{i}} = \sigma_{\boldsymbol{i}}\boldsymbol{\vartheta}\boldsymbol{\vartheta}'$ and any test of $\boldsymbol{\Sigma}_{\boldsymbol{i}} = \boldsymbol{0}$ is reduced to a test of $\sigma_{\boldsymbol{i}} = 0$. Note that the same error structure $\mathbf{K}$ and the same vector of factor loadings $\boldsymbol{\vartheta}$ apply to each of the N observations so that every observation reflects the latent variable in the same way. The latent variable itself follows a univariate mixed model with the same structure as that of $\boldsymbol{X}$. Hence, we are looking for an approximation of $\boldsymbol{y}$ as the basis for a test of the variance component $\sigma_{\boldsymbol{i}}$.

The best linear approximation of $\boldsymbol{y}$ in case of known $\mathbf{K}$, $\boldsymbol{\vartheta}$ is proportional to $\mathbf{K}^{-1}\boldsymbol{\vartheta}$ (see, for example, ANDERSON, 1984, ch. 14.7). $\mathbf{K}^{-1}\boldsymbol{\vartheta}$ is also a solution of the generalized eigenvalue problem

$$\boldsymbol{\Lambda}_{\boldsymbol{i}}\boldsymbol{d} = \mathbf{K}\boldsymbol{d}\lambda \tag{6}$$

for each $\boldsymbol{i} \in T_0 - T_1$: Since

$$\boldsymbol{\Lambda}_{\boldsymbol{i}} = \boldsymbol{\Sigma}_0 + n \sum_{\boldsymbol{j} \in T, \boldsymbol{i} \subseteq \boldsymbol{j}} a_{\boldsymbol{j}}^{*} \boldsymbol{\Sigma}_{\boldsymbol{j}} = \mathbf{K} + \left(\sigma_0 + n \sum_{\boldsymbol{j} \in T, \boldsymbol{i} \subseteq \boldsymbol{j}} a_{\boldsymbol{j}}^{*} \cdot \sigma_{\boldsymbol{j}} \right) \boldsymbol{\vartheta}\boldsymbol{\vartheta}', \qquad \boldsymbol{i} \in T_0 - T_1 ,$$

one eigenvalue of (6) is $1 + \left(\sigma_0 + n \sum_{\boldsymbol{j} \in T, \boldsymbol{i} \subseteq \boldsymbol{j}} a_{\boldsymbol{j}}{}^{*} \cdot \sigma_{\boldsymbol{j}} \right) \cdot \boldsymbol{\vartheta}'\mathbf{K}^{-1}\boldsymbol{\vartheta}$, while the $p-1$ others are all 1. If $\sigma_{\boldsymbol{i}} \geq 0$ for all $\boldsymbol{i} \in T_2$, $\mathbf{K}^{-1}\boldsymbol{\vartheta}$ corresponds to the largest eigenvalue of (6). Concerning the test of $\boldsymbol{\Sigma}_{\boldsymbol{i}} = \mathbf{0}$ with $\boldsymbol{\Lambda}_{\boldsymbol{i}_1}$ and $\boldsymbol{\Lambda}_{\boldsymbol{i}_2}$, $\boldsymbol{i}_1, \boldsymbol{i}_2 \in T_0 - T_1$, such that $\boldsymbol{\Lambda}_{\boldsymbol{i}_1} - \boldsymbol{\Lambda}_{\boldsymbol{i}_2} \propto \boldsymbol{\Sigma}_{\boldsymbol{i}}$, the corresponding quadratic forms $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}_1}\boldsymbol{X}$ and $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}_2}\boldsymbol{X}$ both have central Wishart distributions, so $E(\boldsymbol{X}'(\boldsymbol{A}_{\boldsymbol{i}_1} + \boldsymbol{A}_{\boldsymbol{i}_2})\,\boldsymbol{X})$ is proportional to $\mathbf{K} + c \cdot \boldsymbol{\vartheta}\boldsymbol{\vartheta}'$, where c is a positive real number. Thus, $\mathbf{K}^{-1}\boldsymbol{\vartheta}$ can be approximated by the eigenvector $\boldsymbol{d}$ pertaining to the largest eigenvalue λ of

$$\boldsymbol{X}'(\boldsymbol{A}_{\boldsymbol{i}_1} + \boldsymbol{A}_{\boldsymbol{i}_2})\,\boldsymbol{X}\boldsymbol{d} = \hat{\mathbf{K}}\boldsymbol{d}\lambda , \tag{7}$$

if we can replace $\mathbf{K}$ by a satisfactory estimate $\hat{\mathbf{K}}$.

Several factor analytic methods are available for the estimation of $\mathbf{K}$ (see, for example, BARTHOLOMEW and KNOTT, 1999, chapter 3). However, it has to be kept in mind that these must be applied to $\boldsymbol{Q} = \boldsymbol{X}'(\boldsymbol{A}_{\boldsymbol{i}_1} + \boldsymbol{A}_{\boldsymbol{i}_2})\,\boldsymbol{X}$ or a "larger" matrix in order to keep the conditions of theorem 2.1. LÄUTER et al. (1999) present several proposals for the estimation of $\mathbf{K}$. As a further possibility, an iterated principal factor algorithm may be applied. This starts with the eigenvalue decomposition of $\boldsymbol{R} = \text{Diag}\,(\boldsymbol{Q})^{-\frac{1}{2}}\,\boldsymbol{Q}\,\text{Diag}\,(\boldsymbol{Q})^{-\frac{1}{2}}$. The first step in the iteration gives $\hat{\mathbf{K}} = \text{Diag}\,(\boldsymbol{R} - \phi_1\boldsymbol{e}_1\boldsymbol{e}_1')$, where ϕ_1 is the largest eigenvalue of $\boldsymbol{R}$ and $\boldsymbol{e}_1$ is the corresponding eigenvector. In the second step, the same calculation is done with $\boldsymbol{R} - \hat{\mathbf{K}}$ instead of $\boldsymbol{R}$, giving a new estimate $\hat{\mathbf{K}}$ and so on, until convergence is achieved. There are several ways to justify this algorithm. BARTHOLOMEW and KNOTT (1999), p. 53ff., give some of them. In rare cases, the algorithm will lead to negative estimates in the diagonal of $\hat{\mathbf{K}}$. The estimate obtained after the first step will be positive definite with probability 1.

Example 3.2: [example 3.1 continued] For every patient i and every time j, there is an underlying, unobserved variable y_{ij}. In vector notation, we have

$$\boldsymbol{y} \sim N\left(\mathbf{1}_{st} \cdot \mu_{(y)} + \left[\begin{pmatrix} \boldsymbol{I}_{t-1} \\ -\mathbf{1}_{t-1}' \end{pmatrix} \otimes \mathbf{1}_s \right] \beta_{(y)1},\ \sigma_0 \boldsymbol{I}_t \otimes \boldsymbol{I}_s + \sigma_2 \mathbf{1}_t\mathbf{1}_t' \otimes \boldsymbol{I}_s \right),$$

$$\boldsymbol{X} = \boldsymbol{y}\boldsymbol{\vartheta}' + \boldsymbol{U}, \qquad \boldsymbol{U} \sim N_{st \times p}(\mathbf{0}, \boldsymbol{I}_{st} \otimes \mathbf{K}) .$$

For each time and each patient the matrix of individual error variances $\mathbf{K}$ of the p variables is always the same. Furthermore, the same vector $\boldsymbol{\vartheta}$ of factor load-

ings is in effect at each time. From example 2.2, $\boldsymbol{A}_2 + \boldsymbol{A}_{12} = \boldsymbol{I}_t \otimes (\boldsymbol{I}_s - \frac{1}{s}\mathbf{1}_s\mathbf{1}_s')$, such that

$$E(\boldsymbol{X}'(\boldsymbol{A}_2 + \boldsymbol{A}_{12})\,\boldsymbol{X}) = \sum_{i=1}^{t} E\left(\boldsymbol{X}_i'(\boldsymbol{I}_s - \frac{1}{s}\mathbf{1}_s\mathbf{1}_s')\,\boldsymbol{X}_i\right)$$
$$= t(s-1)\cdot\left(\mathbf{K} + (\sigma_0 + \sigma_2)\,\boldsymbol{\vartheta}\boldsymbol{\vartheta}'\right),$$

showing that this is indeed proportional to $\mathbf{K} + \mathbf{c}\cdot\boldsymbol{\vartheta}\boldsymbol{\vartheta}'$ with some $c > 0$. In this case, $\sigma_2 = 0$ would be tested by

$$\frac{(t-1)\cdot\boldsymbol{d}'\boldsymbol{X}'\left(\frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t' \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\right)\boldsymbol{X}\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{X}'\left(\left(\boldsymbol{I}_t - \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t'\right) \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\right)\boldsymbol{X}\boldsymbol{d}} \overset{H_0}{\sim} F(s-1, (t-1)\,(s-1))$$

with $\boldsymbol{d}$ as the eigenvector corresponding to the largest eigenvalue from

$$\boldsymbol{Q}\boldsymbol{d} = \hat{\mathbf{K}}\boldsymbol{d}\lambda\,,$$

where $\boldsymbol{Q} = \sum_{i=1}^{t} \boldsymbol{X}_i'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_i$ and $\hat{\mathbf{K}}$ has been derived from the iterative principal factor estimation method mentioned above, applied to $\text{Diag}\,(\boldsymbol{Q})^{-1/2}\boldsymbol{Q}\,\text{Diag}\,(\boldsymbol{Q})^{-1/2}$. It is noteworthy that $\boldsymbol{Q}$ is proportional to the sum of the estimated covariance matrices for each of the t times. The same coefficient vector $\boldsymbol{d}$ is used at every time point to calculate the corresponding score vector $\boldsymbol{X}_i\boldsymbol{d}$. □

If a one-factor model seems inappropriate, the reasoning leading to (6) and (7) can easily be extended to situations, where the observations stem from $q > 1$ latent factors leading to corresponding q-dimensional spherical tests as outlined at the beginning of this section.

$E\left(\boldsymbol{X}'(\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2} + \sum_j \boldsymbol{A}_{i_j})\,\boldsymbol{X}\right) \propto \mathbf{K} + c\cdot\boldsymbol{\vartheta}\boldsymbol{\vartheta}'$ with some $c > 0$ still holds for any $\boldsymbol{i}_j \in T_0$ (in case of $\sigma_{\boldsymbol{i}} \geq 0$ for all $\boldsymbol{i}$), even if $\boldsymbol{X}'\boldsymbol{A}_{\boldsymbol{i}_j}\boldsymbol{X}$ does not have a central Wishart-distribution. In that case, c is a weighted sum of $\sigma_{\boldsymbol{i}}$'s plus a weighted sum of squared components from the mean structure. According to theorem 2.1, an exact level-α test is still obtained, if $\boldsymbol{X}'(\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2})\,\boldsymbol{X}$ is replaced by some $\boldsymbol{X}'\left(\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2} + \sum_j \boldsymbol{A}_{i_j}\right)\boldsymbol{X}$ on either side of (7).

3.2 *Tests of the mean structure*

Theorem 9.8 by HOCKING (1985) can easily be extended to state that $\boldsymbol{X}'\boldsymbol{A}_j\boldsymbol{X}$ has a non-central Wishart-distribution:

$$\boldsymbol{X}'\boldsymbol{A}_j\boldsymbol{X} \sim W_p(\boldsymbol{\Lambda}_j,\ \text{rank}\,(\boldsymbol{A}_j),\ \mathbf{B}_j'\boldsymbol{X}'\boldsymbol{\Lambda}_j^{-1}\boldsymbol{X}\mathbf{B}_j)$$

for all $\boldsymbol{j} \in T_1$, where $\mathbf{B}_{\boldsymbol{j}}' \boldsymbol{X}' \boldsymbol{\Lambda}_{\boldsymbol{j}}^{-1} \boldsymbol{X} \mathbf{B}_{\boldsymbol{j}}$ is the non-centrality parameter. To test the hypothesis $H_0 : \mathbf{B}_{\boldsymbol{j}} = \mathbf{0}$, an $\boldsymbol{i} \in T_0 - T_1$ with $\boldsymbol{\Lambda}_{\boldsymbol{i}} = \boldsymbol{\Lambda}_{\boldsymbol{j}}$ has to be selected. A spherical test of H_0 could then, for example, be based on

$$\frac{\operatorname{rank}(\boldsymbol{A}_{\boldsymbol{i}})}{\operatorname{rank}(\boldsymbol{A}_{\boldsymbol{j}})} \frac{\boldsymbol{d}' \boldsymbol{X}' \boldsymbol{A}_{\boldsymbol{j}} \boldsymbol{X} \boldsymbol{d}}{\boldsymbol{d}' \boldsymbol{X}' \boldsymbol{A}_{\boldsymbol{i}} \boldsymbol{X} \boldsymbol{d}} \overset{H_0}{\sim} F(\operatorname{rank}(\boldsymbol{A}_{\boldsymbol{j}}), \operatorname{rank}(\boldsymbol{A}_{\boldsymbol{i}}))\,,$$

where $\boldsymbol{d}$ is a function of $\boldsymbol{X}'(\boldsymbol{A}_{\boldsymbol{j}} + \boldsymbol{A}_{\boldsymbol{i}})\,\boldsymbol{X}$. It turns out that under the assumptions made in section 2, page 22f., namely that of "matching" covariance and mean structure ($T_1 \cap T_2 = \emptyset$), such an $\boldsymbol{i} \in T_0 - T_1$ always exists for each $\boldsymbol{j} \in T_1$ and that the technique amounts to a test of the hypothesis that the mean of some independent contrast from the data is zero. Since spherical tests of means have been treated elsewhere (LÄUTER, 1996; LÄUTER et al., 1996, 1999), this is not pursued any further here, but only the example is continued to demonstrate how this works in a simple case.

Example 3.3: [example 3.2 continued] To test the hypothesis $H_0 : \mathbf{B}_1 = \mathbf{0}$ of no time effect, $\boldsymbol{A}_1 = \left(\boldsymbol{I}_t - \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t'\right) \otimes \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'$ and $\boldsymbol{A}_{12} = \left(\boldsymbol{I}_t - \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t'\right) \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)$ are chosen, since $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_{12}$. With $\boldsymbol{d}$ calculated from $\boldsymbol{X}'(\boldsymbol{A}_1 + \boldsymbol{A}_{12})\,\boldsymbol{X} = \boldsymbol{X}'\left(\left(\boldsymbol{I}_t - \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t'\right) \otimes \boldsymbol{I}_s\right)\boldsymbol{X}$ by the iterated principal factor method or one of the proposals from LÄUTER et al. (1999), the test is based on

$$(s-1) \cdot \frac{\boldsymbol{d}' \boldsymbol{X}' \boldsymbol{A}_1 \boldsymbol{X} \boldsymbol{d}}{\boldsymbol{d}' \boldsymbol{X}' \boldsymbol{A}_{12} \boldsymbol{X} \boldsymbol{d}} \overset{H_0}{\sim} F(t-1, (t-1)\,(s-1))\,.$$

This procedure actually uses an ordinary F-test for the equality of $t-1$ means with s observations on each mean. It is based on $t-1$ independent time contrasts from each of the patients. □

4. Some Relations with Tests from other Contexts

This section is intended to show how some miscellaneous results obtained for certain, relatively simple multivariate linear models fit into the framework outlined in sections 2 and 3. The following will be based on the continued example 2.1 from the previous sections. As an additional restriction, it will be assumed that patients are only observed at $t = 2$ times.

4.1 *Testing independence of two potentially correlated samples*

Looking at $\boldsymbol{A}_2 = \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t' \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)$ and $\boldsymbol{A}_{12} = \left(\boldsymbol{I}_t - \frac{1}{t}\,\mathbf{1}_t\mathbf{1}_t'\right) \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)$ in case of just two times, it is immediately realized that the

"time-parts" of these matrices to the left of the Kronecker-product have roots $\frac{1}{\sqrt{2}}(1,1)'$ and $\frac{1}{\sqrt{2}}(1,-1)'$, respectively. Consequently, the test of $\boldsymbol{\Sigma}_2 = \mathbf{0}$ (see example 3.1) tests an equality-of-variance hypothesis for the sum $\boldsymbol{X}_+ = ((1,1) \otimes \boldsymbol{I}_s)\,\boldsymbol{X}$ and the difference $\boldsymbol{X}_- = ((1,-1) \otimes \boldsymbol{I}_s)\,\boldsymbol{X}$ of the observations from the two times. LÄUTER (1999) has pointed out this fact. In this case, (5) is the test statistic for the usual, well-known test of equality of variances for two independent normal samples (see, for example, LEHMANN, 1986, chapter 5.3) applied to the derived "score samples" $\boldsymbol{X}_+\boldsymbol{d}$ and $\boldsymbol{X}_-\boldsymbol{d}$. Example 3.2 gives a possible derivation of $\boldsymbol{d}$. It may be noted in passing, that for the $\boldsymbol{Q}$ given there, we have $\boldsymbol{Q} = \frac{1}{2}\left(\boldsymbol{X}'_+\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}'_s\right)\boldsymbol{X}_+ + \boldsymbol{X}'_-\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}'_s\right)\boldsymbol{X}_-\right)$. The test requires compound symmetry. In contrast to the univariate case ($p = 1$), where any 2×2-covariance matrix $\boldsymbol{\Sigma}$ with $\mathrm{Diag}\,(\Sigma) \propto \boldsymbol{I}_2$ is "automatically" compound symmetric, this is a restriction here, because it requires the off-diagonal block $\boldsymbol{\Sigma}_2$ to be symmetric.

If we are not willing to assume compound symmetry or if the variances at the first and second time are not the same, we can still test the hypothesis of independence between the two times by a spherical test of correlation between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, using $\boldsymbol{X}_1\boldsymbol{d}$ and $\boldsymbol{X}_2\boldsymbol{d}$. These tests have been described by LÄUTER et al. (1998), page 1979ff., and KROPF (2000), ch. 5.3.3. Since they do not exploit compound symmetry, they are less powerful in its presence than the variance component tests. In addition, it is unclear how they could be extended to the case of $t > 2$.

4.2 *Testing the compound-symmetry assumption*

For $t = 2$, $s > 2$, a test of the compound symmetry assumption is available. This test is easier to derive for the rearranged

$$(\boldsymbol{X}_1, \boldsymbol{X}_2) \sim N_{s\times 2p}\left(\mathbf{1}_s\left(\boldsymbol{\mu}'_1, \boldsymbol{\mu}'_2\right), \boldsymbol{I}_s \otimes \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right) \tag{8}$$

instead of $\boldsymbol{X}$. Using the definitions of $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ from section 4.1,

$$(\boldsymbol{X}_+, \boldsymbol{X}_-) \sim N_{s\times 2p}\left(\mathbf{1}_n\left(\boldsymbol{\mu}'_1 + \boldsymbol{\mu}'_2, \boldsymbol{\mu}'_1 - \boldsymbol{\mu}'_2\right),\right.$$
$$\left.\boldsymbol{I}_n \otimes \begin{pmatrix} \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12} + \boldsymbol{\Sigma}'_{12} \\ \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22} + \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}'_{12} & \boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12} - \boldsymbol{\Sigma}'_{12} \end{pmatrix}\right)$$

is obtained from (8). If compound symmetry holds, $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22}$ and $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{12}$, thus $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ are independent. Consequently, a spherical test of correlation between $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ can be used to check the hypothesis.

The derivation of spherical correlation tests has been described by LÄUTER et al. (1998) and is not fully repeated here, but some aspects are important for the calculation of the coefficients $\boldsymbol{D}$. Application of spherical tests for correlation amounts

to the application of a usual test for canonical correlation 0 to the scores $\boldsymbol{X}_+\boldsymbol{D}_+$ and $\boldsymbol{X}_-\boldsymbol{D}_-$, where $\boldsymbol{D}_+$ and $\boldsymbol{D}_-$ are coefficient matrices derived in accordance with theorem 2.1. The derivation is based on the conditional distribution of $\boldsymbol{X}_-$ given $\boldsymbol{X}_+$. Let $\boldsymbol{E}$, $\boldsymbol{EE}' = \boldsymbol{I}_{s-1}$, $\boldsymbol{E}'\boldsymbol{E} = \boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'$ be a matrix that eliminates the means. Conditional on $\boldsymbol{X}_+$, the matrices

$$\boldsymbol{H} = \boldsymbol{X}_-'\boldsymbol{E}'\boldsymbol{E}\boldsymbol{X}_+'\left(\boldsymbol{X}_+'\boldsymbol{E}'\boldsymbol{E}\boldsymbol{X}_+\right)^{-1}\boldsymbol{X}_+'\boldsymbol{E}'\boldsymbol{E}\boldsymbol{X}_-\,,$$

$$\boldsymbol{G} = \boldsymbol{X}_-'\boldsymbol{E}'\left(\boldsymbol{I}_{s-1} - \boldsymbol{E}\boldsymbol{X}_+\left(\boldsymbol{X}_+'\boldsymbol{E}'\boldsymbol{E}\boldsymbol{X}_+\right)^{-1}\boldsymbol{X}_+'\boldsymbol{E}'\right)\boldsymbol{E}\boldsymbol{X}_-$$

are independent with distributions

$$\boldsymbol{H}\,|\,\boldsymbol{X}_+ \overset{H_0}{\sim} W_p(2(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22}),\, p)\,,$$

$$\boldsymbol{G}\,|\,\boldsymbol{X}_+ \overset{H_0}{\sim} W_p(2(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{22}),\, n-p-1)\,.$$

Concerning the application of theorem 2.1, $\boldsymbol{H} + \boldsymbol{G} = \boldsymbol{X}_-'\left(\boldsymbol{I}_s - \frac{1}{s}\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_-$ is the "smallest" suitable basis for the calculation of score coefficients $\boldsymbol{D}_-$. If we choose rank $(\boldsymbol{D}_-) = q_- = 1$ and determine some $\boldsymbol{d}_-$ from that matrix, this leads to a test using a statistic $\dfrac{n-p-1}{p}\cdot\dfrac{\boldsymbol{d}_-'\boldsymbol{H}\boldsymbol{d}_-}{\boldsymbol{d}_-'\boldsymbol{G}\boldsymbol{d}_-} \overset{H_0}{\sim} F(p, n-p-1)$. This is, however, not recommendable. Note that $\boldsymbol{D}_+ = \boldsymbol{I}_p$ here, and that this is actually a test of multiple correlation zero between $\boldsymbol{X}_-\boldsymbol{d}_-$ and $\boldsymbol{X}_+$. Since the derivation could as well be made for $\boldsymbol{X}_+$ given $\boldsymbol{X}_-$ and since $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ are simple linear transformations of the same p variables, there seems to be no reason why one should treat $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$ differently.

As $\boldsymbol{X}_+$ is considered fixed, we may also use $\boldsymbol{X}_-'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_- + \boldsymbol{X}_+'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_+$ to calculate $\boldsymbol{D}_-$ and we may actually apply the same weights $\boldsymbol{D}_+ = \boldsymbol{D}_-$ to $\boldsymbol{X}_+$, too. Since

$$\frac{1}{2}\cdot\left(\boldsymbol{X}_-'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_- + \boldsymbol{X}_+'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_+\right)$$

$$= \boldsymbol{X}'\left(\boldsymbol{I}_2 \otimes \left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\right)\boldsymbol{X} = \boldsymbol{X}_1'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_1 + \boldsymbol{X}_2'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_2\,,$$

the latter method is consistent with the tests proposed in section 3 (see $\boldsymbol{Q}$ in example 3.2). This approach is intuitively much more appealing. In case of a one-factor model or other situations in which the scores capture an anticipated deviation from the null hypothesis, it is far more powerful than the method described before (this can be deduced from simulations by GLIMM, 1999, for spherical tests of correlation in comparison with the test of multiple correlation), whereas in other cases, the canonical correlation will be superior.

Thus, it is preferable to calculate just one coefficient matrix $\boldsymbol{D} = \boldsymbol{D}_+ = \boldsymbol{D}_-$ from $\boldsymbol{X}_1'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_1 + \boldsymbol{X}_2'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_2$ and apply it to both $\boldsymbol{X}_+$ and $\boldsymbol{X}_-$. As an example, suppose $\boldsymbol{d}$ is calculated by the iterated principal factor method from $\boldsymbol{X}_1'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_1 + \boldsymbol{X}_2'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_2$ with $q = 1$. We can then apply an ordinary test of correlation 0 to the scores $\boldsymbol{X}_+\boldsymbol{d}$ and $\boldsymbol{X}_-\boldsymbol{d}$. The Bravais-Pearson correlation coefficient is

$$r = \frac{\boldsymbol{d}'\boldsymbol{X}_-'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_+\boldsymbol{d}}{\sqrt{\boldsymbol{d}'\boldsymbol{X}_-'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_-\boldsymbol{d}\cdot\boldsymbol{d}'\boldsymbol{X}_+'\left(\boldsymbol{I}_s - \frac{1}{s}\,\mathbf{1}_s\mathbf{1}_s'\right)\boldsymbol{X}_+\boldsymbol{d}}}$$

with $(s-2)\cdot\frac{r^2}{1-r^2} \overset{H_0}{\sim} F(1, s-2)$.

Under the one-factor model, $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{12}'$ is a consequence of $\boldsymbol{\Sigma}_{12} = \sigma_2 \cdot \boldsymbol{\vartheta}\boldsymbol{\vartheta}'$. Given that the one-factor model holds, the test examines if the individual error matrix $\mathbf{K}$ remains unchanged in time.

As a final remark, it is noteworthy that theorem 2.1 allows us to calculate the coefficients $\boldsymbol{D}$ from $\boldsymbol{X}'\boldsymbol{X}$ (compare the final remark in section 3.1 on page 28). If that is done, the same scores $\boldsymbol{XD}$ can be used for every hypothesis test from sections 3 and 4. For the continued example 2.1, this means that the tests given in examples 3.2, 3.3 and section 4 can be applied with $\boldsymbol{d}$ calculated from $\sum_{i=1}^{t}\boldsymbol{X}_i'\boldsymbol{X}_i$ by the iterated principal factor method, say, instead of the various $\boldsymbol{d}$'s from different matrices given in those places.

4.3 *Simulations*

A small simulation experiment was run to get an impression of the power of the proposed tests. For this purpose, data corresponding to example 2.1 was simulated using SAS/IML with $s = 12$, $t = 2$, $p = 4$. Following example 3.2, it was assumed that the four variables in $\boldsymbol{X}$ arise from a latent variable y with mean $\mu_{(y)1} = 2$ at time 1, $\mu_{(y)2} = 4$ at time 2, and variance components $\sigma_0 = 3$, $\sigma_2 = 5$. The factor loadings $\boldsymbol{\vartheta}$ and the matrix $\mathbf{K}$ of individual error variances were randomly chosen in each of the replicates of the simulation experiment with $\boldsymbol{\vartheta}$ from an $N(2\cdot\mathbf{1}_p, \boldsymbol{I}_p)$-distribution and the diagonal elements κ_i of $\mathbf{K}$ as squares of independent realizations from an $N(2,1)$-distribution. The experiment was replicated 100 000 times. In each replication, the variance component test (example 3.2), the test of mean (example 3.3), a test of independence of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (section 4.1) and the test of the compound-symmetry assumption (section 4.2) were done.

Table 4.1
Average p-values for spherical tests from simulation with 100000 replicates

Test	PC-test/IPF method		ordinary PC-test		Wilks'Λ
Matrix	small	full	small	full	
var.comp.	0.0581	0.0580	0.0623	0.0658	0.2832
correlation	0.1094	0.1094	0.1170	0.1227	–
comp.symm.	0.5003	0.5005	0.5002	0.5002	–
mean	0.0698	0.0678	0.0714	0.0770	0.1715

The iterated principal factor (IPF) method described on page 27 was used to calculate coefficient vectors $\boldsymbol{d}$. In case of a negative estimate of κ_i, the corresponding value was set to 0.0001 and the cycle was resumed until the convergence criterion was met. Each of the tests was run in two versions, one using $\boldsymbol{X}'\boldsymbol{X}$, the other using the corresponding "smallest" matrix for the problem (e.g. $\boldsymbol{X}'(\boldsymbol{A}_2 + \boldsymbol{A}_{12})\,\boldsymbol{X}$ for the variance component test) as the basis for coefficient calculation from an eigenvalue problem like (7). In addition, the analogous two versions of the ordinary PC-test using Diag $(\boldsymbol{X}'\boldsymbol{X})$ or Diag $(\boldsymbol{X}'(\boldsymbol{A}_{i_1} + \boldsymbol{A}_{i_2})\,\boldsymbol{X})$, respectively, in (7) instead of $\hat{\mathbf{K}}$, were also calculated. Wilks' Λ tests for variance component $\boldsymbol{\Sigma}_2 = \mathbf{0}$ and for equality of means at the two times were included for comparison.

Table 4.1 shows the average p-values from this experiment. It is obvious that the spherical variance component tests are vastly superior to Wilks' Λ-test. The PC-test with principal factor iteration seems a little better than the ordinary PC test. Use of the "full" sums-of-product matrix $\boldsymbol{X}'\boldsymbol{X}$ instead of the "small" one does not seem to make any difference for the PC-test with principal factor iteration. This is not surprising in the light of the discussion in section 3.1, which shows that the expected values of these two matrices have the same eigenvectors. As was anticipated, the test of correlation between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is inferior to the variance component test in this situation, although it is still superior to Wilks'Λ test. Since we have compound symmetry in this simulation, the test of compound symmetry from section 4.2 just shows the expected result.

Finally, the test for equality of means at the two times reveals the same pattern as the test of the variance component: The spherical tests are superior to Wilks' Λ in this situation, the PC-test with principal factor iteration seems a little better than the ordinary PC-test.

Acknowledgments

The author is grateful to the two referees for their constructive comments that helped to improve the paper. Thanks are also due to Professor Muni Srivastava for his support of this work.

References

AHRENS, H. and LÄUTER, J., 1981: *Mehrdimensionale Varianzanalyse,* 2. Auflage. Akademie-Verlag, Berlin.

ANDERSON, T. W., 1984: *An Introduction to Multivariate Statistical Analysis,* 2nd ed. Wiley, New York.

BARTHOLOMEW, D. J. and KNOTT, M., 1999: *Latent Variable Models and Factor Analysis.* Arnold, London.

FANG, K.-T. and LIANG, J., 1999: Tests of spherical and elliptical symmetry. In: S. Kotz, C. Read, and D. L. Banks (eds.): *Encyclopaedia of Statistical Sciences, update volume* **3**. Wiley, New York, 686–691.

FANG, K.-T., LI, R., and LIANG, J., 1998: A multivariate version of Ghosh T_3-plot to detect non-multinormality. *Computational Statistics & Data Analysis* **28**, 371–386.

GLIMM, E., 1999: *Güte- und Optimalitätseigenschaften stabiler multivariater Verfahren.* Doctoral thesis, Mathematical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany.

GLIMM, E. and LÄUTER, J., 2000: On the admissibility of spherical multivariate tests. Submitted to *Journal of Multivariate Analysis.*

HOCKING, R. R., 1985: *The Analysis of Linear Models.* Brooks/Cole Publishing, Monterey, California.

KROPF, S., 2000: *Hochdimensionale multivariate Verfahren in der medizinischen Statistik.* Habilitation thesis, Medical Faculty, Otto von Guericke University Magdeburg, Magdeburg, Germany.

KROPF, S. and GLIMM, E., 1996: Stabilisierte Tests für multivariate Repeated-Measurement-Designs. In: Baur, M. P., Fimmers, R., and Blettner, M. (eds.): *Medizinische Informatik, Biometrie und Epidemiologie GMDS'96.* 41. Jahrestagung der GMDS. MMV Medizin Verlag, München, 350–354.

KROPF, S. and LÄUTER, J., 2000: Detection of pairwise correlations in a multivariate structure. *Biometrical Journal* **42**, 377–391.

KROPF, S., LÄUTER, J., and GLIMM, E., 1997: Stabilized multivariate tests – the inclusion of missing values. *Biometrical Journal* **39**, 149–169.

LÄUTER, J., 1992: *Stabile Multivariate Verfahren: Diskriminanzanalyse – Regressionsanalyse – Faktoranalyse.* Akademie-Verlag, Berlin.

LÄUTER, J., 1996: Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964–970.

LÄUTER, J., 1999: *personal communication.*

LÄUTER, J., GLIMM, E., and KROPF, S., 1996: New multivariate tests for data with an inherent structure. *Biometrical Journal* **38**, 5–23. Erratum: *Biometrical Journal* **40**, 1015.

LÄUTER, J., GLIMM, E., and KROPF, S., 1998: Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26**, 1972–1988.

LÄUTER, J., KROPF, S., and GLIMM, E., 1999: Exact stable multivariate tests for applications in clinical research. In: *1998 Proceedings of the Biopharmaceutical Section of the American Statistical Association.* American Statistical Association, 46–55.

LEHMANN, E. L., 1986: *Testing statistical hypotheses,* second edition. Wiley, New York.

LIANG, J. and FANG, K.-T., 2000: Some applications of Läuter's technique in constructing high-dimensional mean tests and goodness-of-fit tests for spherical symmetry. To appear in *Biometrical Journal.*

LIANG, J., LI, R., FANG, H., and FANG, K.-T., 2000: Testing multinormality based on low-dimensional projection. *Journal of Statistical Planning and Inference* **86**, 129–141.

RAO, C. R., 1973: *Linear Statistical Inference and Its Application,* 2nd ed. Wiley, New York.

EKKEHARD GLIMM
Department of Statistics
University of Toronto
100 St.George Street
Toronto, Ontario M5S 3G3
Canada
E-mail: glimm@utstat.utoronto.ca

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

COMMUN. STATIST.—SIMULA., 31(4), 589–604 (2002)

MULTIVARIATE TESTS OF NORMAL MEAN VECTORS WITH RESTRICTED ALTERNATIVES

Ekkehard Glimm,[1] Muni S. Srivastava,[2] and Jürgen Läuter[3]

[1]AICOS Technologies AG, 4057 Basel, Switzerland
E-mail: eglimm@aicos.com
[2]Department of Statistics, University of Toronto, Toronto M5S 3G3, Canada
[3]Institute of Biometry, University of Magdeburg, 39120 Magdeburg, Germany

ABSTRACT

In this paper, we consider tests for the hypothesis that the mean vector is zero against one-sided alternatives when the observation vectors are independently and identically distributed as normal with unknown covariance matrix. The exact null-distribution of the tests is derived. The tests generalize the centre-direction test proposed by Tang et al.[1] for known covariance. In addition, the modification is order- and scale-invariant. Power comparisons with some other tests are presented. It can be shown that the null distribution of the test statistic holds for data arising from any elliptical distribution, not just the normal distribution.

Key Words: Multivariate tests; Restricted alternatives; Order alternatives

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

1. INTRODUCTION

This paper deals with tests of the means in the statistical model

$$\mathbf{X} \sim N_{n\times p}\left(\mathbf{1}_n\boldsymbol{\mu}', \mathbf{I}_n \otimes \Sigma\right), \tag{1}$$

where $\otimes$ denotes the Kronecker product, $\mathbf{1}_n$ is a vector of n ones, $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{\mu}$ an unknown parameter vector and Σ a positive definite covariance matrix. The rows of the matrix $\mathbf{X}$ form a sample of n independent observations from a p-dimensional normal distribution with covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1,\ldots,p}$ and mean $\boldsymbol{\mu} = (\mu_i)_{i=1,\ldots,p}$. If no knowledge of $\boldsymbol{\mu}$ and Σ is assumed, the uniformly most powerful invariant test of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ is Hotelling's T^2 test (see e.g., Srivastava and Khatri,[2] Theorem 4.3.1). However, practical considerations often imply restrictions on either $\boldsymbol{\mu}$ or Σ or on both. Several authors have investigated tests of H_0 with various restrictions on the model (1) (Kudo,[3] Nüesch,[4] Schaafsma and Smid,[5] Perlman,[6] Robertson, Wright and Dykstra,[7] O'Brien,[8] Tang et al.,[1] Follmann,[9,10] Läuter et al.,[11–13] Wang and McDermott,[14] McDermott,[15] Srivastava et al.[16]).

A lot of attention has been given to tests of H_0 where the alternative is restricted to $A : \mu_i \geq 0$ for all $i = 1, \ldots, p$ with at least one strict inequality. Tests for such one-sided hypotheses are often needed in practice, for example in clinical trials, where sucess of a treatment is judged by the treatment's impact on several response features simultaneously. Kudo[3] and Nüesch[4] have given the likelihood-ratio test for this problem assuming Σ is known. Perlman[6] derived the likelihood-ratio test for unknown Σ. Wang and McDermott[14] derived a conditional likelihood ratio test by obtaining improved critical values from the conditional distribution of the likelihood-ratio test statistic given the total sums-of-products matrix $\mathbf{X}'\mathbf{X}$. Unfortunately, all of these tests are very cumbersome in application, since i) the calculation of the maximum likelihood estimate of $\boldsymbol{\mu}$ under the restriction $\mu_i \geq 0$ is difficult and may require iterative search in case of $p \geq 4$ and ii) the distributions of the test statistics are unknown and depend on Σ, even under H_0, or, in Wang and McDermott's case, on $\mathbf{X}'\mathbf{X}$. Therefore, several authors (Tang et al.,[1] Fraser et al.,[17] Follmann,[9,10] Srivastava et al.[16]) have looked for alternatives to the likelihood-ratio test.

In this paper, we follow an idea of Tang et al.[1] The hypothesis H_0 is tested by a statistic depending on the original data matrix $\mathbf{X}$ via a transformation $\mathbf{X}\mathbf{A}'$. Tang et al.[1] developed their method for the case of known Σ. Section 2 outlines the motivation behind Tang's approach and explains how it can be modified to be applicable for the case of unknown Σ.

Section 3 gives a modified method that is scale- and order-invariant. In Section 4, the null distribution and power of the resulting tests are investigated.

2. MOTIVATION OF ORTHOGONAL CONTRAST TESTS

Tang et al.[1] have proposed to test the hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ vs. $A : \boldsymbol{\mu} \geq \mathbf{0}$ by means of contrasts $z_i = \sqrt{n}\mathbf{a}_i'\bar{\mathbf{x}}$, $i = 1, \ldots, p$, where $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}_n$ and the coefficient vectors $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)'$ fulfill $\mathbf{A}'\mathbf{A} = \Sigma^{-1}$. This leads to $\mathbf{X}\mathbf{A}' \sim N_{n\times p}\big(\mathbf{1}_n\boldsymbol{\mu}'\mathbf{A}', \mathbf{I}_n \otimes \mathbf{I}_p\big)$, i.e., the contrasts $z_i = \sqrt{n}\mathbf{a}_i'\bar{\mathbf{x}}$ are independent with variance 1 and mean 0 under H_0. We may combine the contrasts in various ways to test the null hypothesis. Tang et al.[1] propose to use $\sum_i \max_i(z_i, 0)^2$. They also derived the null distribution of this statistic, which is a special case of the chi-bar-squared distibution. As Tang et al.[1] noted, $\mathbf{A}$ is not unique and its choice affects the power of the ensuing test.

If Σ is unknown, one might consider replacing Σ by its usual estimate. However, several authors have shown by simulation that this approach leads to extremely liberal tests (see for example Reitmeir and Wassmer[18]). Yet, the idea of producing independent contrasts can be applied to $\mathbf{G} = \mathbf{X}'\mathbf{X}$ instead of Σ. This results in some root $\mathbf{A}$ of $\mathbf{G}^{-1}$, i.e., an $\mathbf{A}$ that fulfills $\mathbf{A}'\mathbf{A} = \mathbf{G}^{-1}$. Läuter et al.[12] have pointed out that $\mathbf{X}\mathbf{A}'$ is distributed according to the so-called left-spherical uniform $n \times p$-distribution (Fang and Zhang,[19] Chapter 3) under H_0.

Tests of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ can be based on the components of

$$\mathbf{u} = (u_i) = \sqrt{n}\mathbf{A}\bar{\mathbf{x}} \tag{2}$$

in the same way as they are based on z_i in the case of known Σ. For example, Läuter et al.[13] have proposed to apply Tang's statistic, i.e., reject $H_0 : \boldsymbol{\mu} = \mathbf{0}$, if

$$\tilde{u}^2 = \sum_{i=1}^{p}(\max(0, u_i))^2 \tag{3}$$

is larger than a critical value. However, they only considered one certain choice of $\mathbf{A}$ and did not derive the distribution of $\tilde{u}^2$.

In order to devise powerful tests from these premises, we have to make a choice of $\mathbf{A}$. This problem is considered in the following section.

MARCEL DEKKER, INC. • 270 MADISON AVENUE • NEW YORK, NY 10016

3. CHOICE OF THE ROOT OF G

3.1. Centre-Direction Coefficients

Tang et al.[1] propose to choose $\mathbf{A}$ in such a way that the so-called centre direction of the p column vectors of $\mathbf{A}$ coincides with the centre direction of the positive orthant. The general aim is to replace the "original" coefficient vectors $\mathbf{e}_1, \ldots, \mathbf{e}_p$, $(\mathbf{e}_1, \ldots, \mathbf{e}_p) = \mathbf{I}_p$, which are the edges of the alternative space $\{\boldsymbol{\mu} \in \mathbf{R}^p : \mu_i \geq 0\}$, by $\mathbf{A}$ with $\mathbf{A}\Sigma\mathbf{A}' = \mathbf{I}_p$, such that $\mathbf{A}$ is "as close as possible" to $\mathbf{I}_p$. The centre direction $\mathbf{c}_\mathbf{M}$ of a $p \times p$-matrix $\mathbf{M}$ is defined as the p-dimensional vector (of length 1) that has the same angle with each of the column vectors of $\mathbf{M}$, i.e., $\big(\mathrm{Diag}(\mathbf{M}'\mathbf{M})\big)^{-1/2}\mathbf{M}'\mathbf{c}_\mathbf{M} \propto \mathbf{1}_p$. The above mentioned centre direction condition from Tang et al.[1] requests that $\mathbf{c}_\mathbf{A} = \frac{1}{\sqrt{p}}\mathbf{1}_p$. Hence, in addition to

$$\mathbf{A}'\mathbf{A} = \mathbf{G}^{-1}, \tag{4}$$

$\mathbf{A}$ has to fulfill

$$\mathbf{A}'^{-1}\big(\mathrm{Diag}(\mathbf{A}'\mathbf{A})\big)^{1/2}\mathbf{1}_p \propto \mathbf{1}_p. \tag{5}$$

This additional restriction still does not make the choice of $\mathbf{A}$ unique if $p \geq 3$. Tang et al.[1] give an algorithm that leads to an arbitrary solution of (4) and (5). Since this algorithm gives different results for different orderings of columns in $\mathbf{X}$, they propose an ordering of the columns of Σ^{-1} (the equivalent here is an ordering of $\mathbf{G}^{-1}$) before the application of the algorithm. This removes dependence on order, but since the ordering is based on the magnitude of the variance estimates of the variables in $\mathbf{X}$, it introduces scale-dependence.

In this paper, we propose to obtain $\mathbf{A}$ by a procedure that yields a scale- and order-invariant solution:

First, $\mathbf{U}$ with $\mathbf{U}'\mathbf{U} = \mathbf{G}^{-1}$ is obtained by any method for the calculation of a root of a positive definite matrix, for example by the well-known Cholesky-composition. Let $\mathbf{Q}_1 = (\mathbf{q}_1, \ldots, \mathbf{q}_p)$ be any orthogonal matrix with $\mathbf{q}_1 \propto \mathbf{U}'^{-1}\big[\mathrm{Diag}(\mathbf{G}^{-1})\big]^{1/2}\mathbf{1}_p$ and $\mathbf{Q}_2 = (\mathbf{q}_1^*, \ldots, \mathbf{q}_p^*)$ be any orthogonal matrix with $\mathbf{q}_1^* \propto \mathbf{1}_p$. Then $\mathbf{A}_0 = \mathbf{Q}_2\mathbf{Q}_1'\mathbf{U}$ fulfills (4) and (5). This approach is due to Tang et al.[1] Their solution is obtained, if $\mathbf{U}$ is from Cholesky-decomposition, $\mathbf{Q}_1$ is calculated by Gram-Schmidt-orthogonalization of $\mathbf{q}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$, $\mathbf{Q}_2$ by Gram-Schmidt-orthogonalization of $\mathbf{q}_1^*, \mathbf{e}_2, \ldots, \mathbf{e}_p$.

Now, let us consider the partition

$$\mathbf{Q}_2'\mathbf{A}_0\big(\mathrm{Diag}(\mathbf{A}_0'\mathbf{A}_0)\big)^{-1/2}\ \mathbf{Q}_2=\mathbf{Q}_1'\mathbf{U}\big(\mathrm{Diag}(\mathbf{G}^{-1})\big)^{-1/2}\ \mathbf{Q}_2=\begin{pmatrix} a_{11} & \mathbf{a}_{12}' \\ \mathbf{a}_{(0)21} & \mathbf{A}_{(0)22}\end{pmatrix}$$

and let $\mathbf{E}\Lambda^{1/2}\mathbf{E}'$ be from the eigenvalue decomposition $\mathbf{A}_{(0)22}'\mathbf{A}_{(0)22}=\mathbf{E}\Lambda\mathbf{E}'$, where Λ is the diagonal matrix of eigenvalues and $\mathbf{E}$ the matrix of corresponding eigenvectors. Then

$$\mathbf{A}=\mathbf{Q}_2\begin{pmatrix} a_{11} & \mathbf{a}_{12}' \\ \mathbf{E}\Lambda^{1/2}\mathbf{E}'\mathbf{A}_{(0)22}^{-1}\mathbf{a}_{(0)21} & \mathbf{E}\Lambda^{1/2}\mathbf{E}'\end{pmatrix}\ \mathbf{Q}_2'\big(\mathrm{Diag}(\mathbf{A}_0'\mathbf{A}_0)\big)^{1/2}. \tag{6}$$

This solution is unique with probability 1. It does not depend on the choice of $\mathbf{Q}_1$ and $\mathbf{Q}_2$.

$\mathbf{A}$ from (6) is also the matrix that maximizes $\mathrm{tr}\Big(\mathbf{A}\big(\mathrm{Diag}(\mathbf{A}'\mathbf{A})\big)^{-1/2}\Big)$ subject to the restrictions (4) and (5). Among all coefficients with the "right" centre direction, these are the ones that maximize the sum of the cosines with the edges $\mathbf{e}_i$ of the positive orthant. In this sense, $\mathbf{A}$ is closest to $\mathbf{I}_p$. This assertion is proved in Proof 1 of Appendix A. Proof 2 of Appendix A shows that the approach is order- and scale-invariant.

3.2. Other Coefficients

The centre direction is just one possible criterion for defining "closeness" of $\mathbf{I}_p$ and $\mathbf{A}$. Matrices $\mathbf{A}$ of coefficients that fulfill (4) but not (5) also might be taken into consideration. One obvious choice for $\mathbf{A}$ would be the symmetric, positive definite root of $\mathbf{G}^{-1}$ without the restriction of the centre direction. This choice maximizes $\mathrm{tr}(\mathbf{A})$ as well as $\mathrm{tr}(\mathbf{AG})$. It can be obtained as $\mathbf{E}\Lambda^{\frac{1}{2}}\mathbf{E}'$ from the usual eigenvalue decomposition $\mathbf{G}^{-1}=\mathbf{E}\Lambda\mathbf{E}'$.

Since this approach does not give a scale-invariant solution, it will usually be better to use

$$\mathbf{A}=\mathbf{E}\Lambda^{1/2}\mathbf{E}'\big(\mathrm{Diag}(\mathbf{G}^{-1})\big)^{1/2} \tag{7}$$

where the eigenvalue decomposition

$$\big(\mathrm{Diag}(\mathbf{G}^{-1})\big)^{-1/2}\mathbf{G}^{-1}\big(\mathrm{Diag}(\mathbf{G}^{-1})\big)^{-1/2}=\mathbf{E}\Lambda\mathbf{E}' \tag{8}$$

is used. This choice maximizes $\mathrm{tr}(\mathbf{A}\big(\mathrm{Diag}(\mathbf{G}^{-1})\big)^{-\frac{1}{2}})$ and is scale-invariant. In comparison to the method described in the previous subsection, the same criterion is maximized without the restriction of a given centre direction.

Scale-invariance is also reached by standardizing $\mathbf{X}$ to $\mathbf{X}(\mathrm{Diag}(\mathbf{G}))^{-1/2}$ and obtaining $\mathbf{E}\Lambda^{\frac{1}{2}}\mathbf{E}'$ from the eigenvalue decomposition of the inverse of the correlation matrix

$$(\mathrm{Diag}(\mathbf{G}))^{1/2}\mathbf{G}^{-1}(\mathrm{Diag}(\mathbf{G}))^{1/2} = \mathbf{E}\Lambda\mathbf{E}'. \tag{9}$$

In that case, the test statistic would be

$$\sqrt{n}\mathbf{E}\Lambda^{1/2}\mathbf{E}'(\mathrm{Diag}(\mathbf{G}))^{-1/2}\bar{\mathbf{x}}. \tag{10}$$

Läuter et al.[13] have considered the so-called left-symmetric root of $\mathbf{G}$, which is defined as the matrix $\mathbf{B}$ with maximum product $\prod_{i=1}^{p} b_{ii}$ of its diagonal elements b_{ii} among all matrices that fulfill $\mathbf{B}'\mathbf{B} = \mathbf{G}$. A matrix that fulfills (4) is obtained from setting $\mathbf{A} = \mathbf{B}'^{-1}$. The left-symmetric root is scale-invariant and sign-invariant in the sense that by changing the sign of a variable, only the sign of the corresponding column in $\mathbf{B}$ is affected. Läuter et al.[13] give a simple iterative algorithm that can be used to calculate the left-symmetric root. The other proposals from this section are also easy to implement on computer with any package that provides eigenvalue decomposition, such as SAS/IML.

4. NULL DISTRIBUTIONS AND POWER OF ORTHOGONAL CONTRAST TESTS

4.1. The Null Distribution of $\tilde{u}^2$

Theorem 1. *For $n \geq p+1$ and any* $\mathbf{A}$ *that fulfills* (4), *the distribution of* $\tilde{u}^2$ *from* (3) *has the pdf*

$$f_{\tilde{u}^2}(r) = \frac{1}{2^p}\sum_{k=0}^{p-1}\binom{p}{k}\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{p-k}{2}\right)\Gamma\left(\frac{n-p+k}{2}\right)}\cdot r^{\frac{p-k}{2}-1}(1-r)^{\frac{n-p+k}{2}-1}, \quad r \in (0,1]$$

$$P(\tilde{u}^2 = 0) = \frac{1}{2^p} \tag{11}$$

The proof of this theorem is given in Appendix A, Proof 3. The theorem shows, that the distribution of $\tilde{u}^2$ is a mixture of Beta-distributions. The quantiles of the distribution of $\tilde{u}^2$ can be obtained by a simple

Taylor-series expansion: To a first-order approximation, we have

$$\beta_\alpha \approx \beta_{\alpha,0} + \frac{1-\alpha-F_{\vec{u}^2}(\beta_{\alpha,0})}{f_{\vec{u}^2}(\beta_{\alpha,0})}, \tag{12}$$

where β_α is the $1-\alpha$-quantile of the distribution of $\vec{u}^2$, $F_{\vec{u}^2}(\cdot)$ is its distribution function, $F_{\vec{u}^2}(\beta_\alpha) = 1-\alpha$, and $\beta_{\alpha,0}$ is an approximation of β_α. Many statistics software packages, like SAS, provide quantiles of the Beta-distributions, such that for given $\beta_{\alpha,0}$, (12) can easily be calculated. A corresponding SAS-program is given in Glimm and Srivastava.[20]

In any practical application, it is even easier to obtain the p-value from the observed value $\vec{u}_0^2$ of $\vec{u}^2$, since many statistics software packages provide the distribution functions $F_{\frac{p-k}{2},\frac{n-p+k}{2}}(\cdot)$ of $B(\frac{p-k}{2}, \frac{n-p+k}{2})$-distributions. It is obvious from (11) that the p-value is

$$1 - \frac{1}{2^p} - \frac{1}{2^p}\sum_{k=0}^{p-1}\binom{p}{k} F_{\frac{p-k}{2},\frac{n-p+k}{2}}(\vec{u}_0^2), \quad \vec{u}_0^2 > 0.$$

4.2. Simulations

In this subsection, empirical powers of the orthogonal contrast tests applying statistic (3) with the root $\mathbf{A}$ of $\mathbf{G}^{-1}$ calculated by

1. (7) and (8) (test "std." in Tables 1–4) and by
2. (6) (test "CD_{max}" in Tables 1–4)

are compared with those of the M-test by Srivastava et al.[16] and Wang and McDermott's conditional likelihood-ratio test. The conditional likelihood-ratio test uses Perlman's likelihood-ratio statistic

$$\frac{n\widehat{\boldsymbol{\mu}}'\mathbf{W}^{-1}\widehat{\boldsymbol{\mu}}}{1+n(\overline{\mathbf{x}}-\widehat{\boldsymbol{\mu}})'\mathbf{W}^{-1}(\overline{\mathbf{x}}-\widehat{\boldsymbol{\mu}})}, \tag{13}$$

where $\widehat{\boldsymbol{\mu}}$ is the maximum likelihood estimate of $\boldsymbol{\mu}$ from (1) under the restriction $A: \mu_i \geq 0$ and $\mathbf{W} = \mathbf{X}'(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n')\mathbf{X}$. Wang and McDermott[14] derive the conditional distribution of (13) given the observed $\mathbf{X}'\mathbf{X}$. They also provide an algorithm for obtaining the p-value of the conditional likelihood-ratio test by numerical integration. The maximum likelihood estimate $\widehat{\boldsymbol{\mu}}$ can be obtained by an iterative algorithm provided by Wollan and Dykstra.[21]

Table 1. Empirical Power of Multivariate Tests for Positive Mean, $p = 3$, $n = 17$

ρ	δ^2	$\boldsymbol{\mu}$	Std.	CD_{max}	M, bootstr.	CLRT
−0.25	0.3	$(0,0,0.5)'$	0.519	0.521	0.506	0.504
	1.5	$(0.5,0.5,0.5)'$	0.995	0.995	0.992	0.994
0	0.25	$(0,0,0.5)'$	0.410	0.410	0.393	0.401
	0.75	$(0.5,0.5,0.5)'$	0.904	0.905	0.860	0.875
0.25	0.278	$(0,0,0.5)'$	0.407	0.402	0.423	0.407
	0.5	$(0.5,0.5,0.5)'$	0.776	0.776	0.723	0.701
0.5	0.375	$(0,0,0.5)'$	0.471	0.459	0.491	0.506
	0.375	$(0.5,0.5,0.5)'$	0.665	0.665	0.574	0.564
0.75	0.7	$(0,0,0.5)'$	0.680	0.640	0.754	0.759
	0.3	$(0.5,0.5,0.5)'$	0.577	0.577	0.456	0.448

ρ: constant correlation between any two variables.
δ^2: Mahalanobis distance $\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}$.
std.: statistic (3) with scale-invariant root.
CD_{max}: statistic (3) with centre-direction contrasts + maximum trace.
M, bootstr.: M-test with critical value from parametric bootstrap.
CLRT: conditional likelihood-ratio test by Wang and McDermott.[14]

The M-test uses the statistic $M = \max(t_0, t_1, \ldots, t_p)$ with $t_i = \sqrt{n}\mathbf{a}_i'\bar{\mathbf{x}}/\sqrt{\mathbf{a}_i'\mathbf{W}\mathbf{a}_i/(n-1)}$, where $\mathbf{a}_1, \ldots, \mathbf{a}_p$ are the column vectors of $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{a}_0' = (\sqrt{g_{11}}, \ldots, \sqrt{g_{pp}})(\mathbf{X}'\mathbf{X})^{-1}$; g_{ii} being the i-th diagonal element of $\mathbf{X}'\mathbf{X}$. Hence, this test compares the direction of the observed mean vector with that of several vectors corresponding to various departures from the null hypothesis. Critical values for this test are obtained from a parametric bootstrap.

The simulation results are given in Tables 1–4. Tables 1–3 use the same situations as Srivastava et al.,[16] Tables 2–5. For convenience, the covariance matrices used in Table 3 are reproduced in Appendix B. The situation investigated in Table 1 has first been studied by Wang and McDermott.[14]

Empirical power values for the orthogonal contrast tests 1. and 2. are based on 100 000 replications. Simulation results for coefficients from the three other methods described in subsection 3.2 were very similar to those for test 1, whereas for the centre-direction coefficients according to Tang's original proposals they were close, but predominantly slightly inferior to those for test 2. Hence, the corresponding empirical powers are not reported. For $p = 2$, all scale-invariant tests from Section 3 are the same. In general, the differences in power between all of the considered tests were

Table 2. Empirical Power of Multivariate Tests for Positive Mean, $p=4$, $n=40$

ρ	$\boldsymbol{\mu}$	Std.	CD_{max}	M, bootstr.	CLRT
−0.3	$(1,1,1,1)'$	0.846	0.847	0.898	0.902
	$(1,0,0,0)'$	0.836	0.837	0.855	0.863
	$(1,2,0,0)'$	0.843	0.844	0.881	0.876
	$(1,2,3,4)'$	0.846	0.846	0.901	0.891
	$(1,1.1,1.2,1.3)'$	0.848	0.848	0.893	0.891
	$(1,1.2,1.4,1.6)'$	0.848	0.848	0.901	0.890
0	$(1,1,1,1)'$	0.849	0.849	0.801	0.831
	$(1,0,0,0)'$	0.765	0.763	0.754	0.765
	$(1,2,0,0)'$	0.797	0.800	0.720	0.780
	$(1,2,3,4)'$	0.841	0.842	0.763	0.821
	$(1,1.1,1.2,1.3)'$	0.850	0.851	0.812	0.829
	$(1,1.2,1.4,1.6)'$	0.848	0.848	0.803	0.821
0.5	$(1,1,1,1)'$	0.848	0.848	0.789	0.769
	$(1,0,0,0)'$	0.674	0.661	0.750	0.706
	$(1,2,0,0)'$	0.690	0.690	0.603	0.718
	$(1,2,3,4)'$	0.800	0.804	0.594	0.763
	$(1,1.1,1.2,1.3)'$	0.845	0.845	0.776	0.768
	$(1,1.2,1.4,1.6)'$	0.844	0.844	0.755	0.770
0.9	$(1,1,1,1)'$	0.848	0.848	0.796	0.719
	$(1,0,0,0)'$	0.590	0.568	0.734	0.673
	$(1,2,0,0)'$	0.551	0.544	0.560	0.680
	$(1,2,3,4)'$	0.638	0.644	0.470	0.710
	$(1,1.1,1.2,1.3)'$	0.832	0.833	0.677	0.724
	$(1,1.2,1.4,1.6)'$	0.794	0.798	0.574	0.726

Means $\boldsymbol{\mu}$ are standardized such that δ^2=0.25. Otherwise, legend of Table 1 applies.

always small for this situation. Therefore we have not included these results either.

The empirical powers for Wang and McDermott's conditional likelihood-ratio test are taken from Srivastava et al.[16] Critical values for this test depend on $\mathbf{X}'\mathbf{X}$, such that they have to be obtained as a part of every application. For the M-test, critical values depend on Σ. Calculation of the empirical power of the M-tests in Tables 1–4 requires bootstrap sampling within each simulation step. Due to this, only 3000 replications were done. Consequently, these results are somewhat less reliable than the other entries in the tables. To gain insight into this, simulations were also performed for the null case. The empirical powers for the M-tests with the bootstrap were

Table 3. Empirical Power of Multivariate Tests for Positive Mean, $p = 4$, $n = 40$

Σ	$\boldsymbol{\mu}$	Std.	CD_{max}	M, bootstr.
Σ_1	$(1,1,1,1)'$	0.842	0.837	0.733
	$(1,0,0,0)'$	0.710	0.666	0.768
	$(1,2,0,0)'$	0.716	0.745	0.594
	$(1,2,3,0)'$	0.759	0.746	0.626
	$(1,2,3,4)'$	0.803	0.800	0.692
	$(1,1.2,1.4,1.6)'$	0.833	0.826	0.706
Σ_2	$(1,1,1,1)'$	0.842	0.836	0.753
	$(1,0,0,0)'$	0.672	0.646	0.742
	$(1,2,0,0)'$	0.763	0.703	0.637
	$(1,2,3,0)'$	0.690	0.655	0.639
	$(1,2,3,4)'$	0.798	0.807	0.626
	$(1,1.2,1.4,1.6)'$	0.838	0.836	0.745
Σ_3	$(1,1,1,1)'$	0.840	0.839	0.669
	$(1,0,0,0)'$	0.673	0.624	0.747
	$(1,2,0,0)'$	0.739	0.695	0.697
	$(1,2,3,0)'$	0.711	0.758	0.498
	$(1,2,3,4)'$	0.812	0.815	0.517
	$(1,1.2,1.4,1.6)'$	0.836	0.839	0.633
Σ_4	$(1,1,1,1)'$	0.845	0.846	0.742
	$(1,0,0,0)'$	0.628	0.641	0.729
	$(1,2,0,0)'$	0.655	0.675	0.544
	$(1,2,3,0)'$	0.716	0.720	0.533
	$(1,2,3,4)'$	0.785	0.779	0.588
	$(1,1.2,1.4,1.6)'$	0.845	0.845	0.759

Σ: covariance matrices from Srivastava et al.,[16] Table 4 (see Appendix B). Otherwise, legend of Table 2 applies.

between 0.040 and 0.068 for $\alpha = 0.05$, whereas for all other tests they were between 0.049 and 0.051.

The tables show that the tests based on the statistic (3) compare favourably with the M-test. Only on the edges of the alternative space, they have lower power than the M-test. In the interior of the alternative space, they are superior and they outperform the M-tests in most situations given in Table 1–4 without having severe power deficiencies anywhere. The power differences between the orthogonal contrast tests using statistic (3) are small.

Table 4. Empirical Power of Multivariate Tests for Positive Mean, $p = 6$, $n = 17$

ρ	$\boldsymbol{\mu}$	Std.	CD_{max}	M, bootstr.
0	$(1,\ldots,1)'$	0.464	0.471	0.321
	$(1,1,1,0,0,0)'$	0.357	0.376	0.211
	$(1,0,\ldots,0)'$	0.273	0.277	0.237
	$(1,\ldots,6)'$	0.429	0.441	0.264
0.2	$(1,\ldots,1)'$	0.469	0.473	0.285
	$(1,1,1,0,0,0)'$	0.297	0.313	0.133
	$(1,0,\ldots,0)'$	0.232	0.230	0.211
	$(1,\ldots,6)'$	0.398	0.411	0.190
0.5	$(1,\ldots,1)'$	0.473	0.477	0.310
	$(1,1,1,0,0,0)'$	0.228	0.239	0.109
	$(1,0,\ldots,0)'$	0.201	0.196	0.211
	$(1,\ldots,6)'$	0.334	0.345	0.155
0.9	$(1,\ldots,1)'$	0.473	0.476	0.320
	$(1,1,1,0,0,0)'$	0.149	0.153	0.083
	$(1,0,\ldots,0)'$	0.172	0.161	0.198
	$(1,\ldots,6)'$	0.197	0.203	0.095

Legend of Table 2 applies.

5. CONCLUSIONS

The main purpose of this paper is to improve on the idea by Tang et al.[1] of centre-direction coefficients. The additional requirement of a maximum trace subject to the centre-direction removes the ambiguity in their calculation and facilitates invariance with respect to the order of the variables as well as with respect to their scale.

The method, but also Tang's original approach as well as some related techniques for obtaining orthogonal contrasts can be adapted to the case of unknown Σ utilizing the total sums-of-products matrix $\mathbf{X}'\mathbf{X}$. The resulting tests have null distributions that are much easier to obtain than that of previously developed tests. They also perform well in terms of power. Comparing the orthogonal contrast methods from subsection 3.1 and 3.2 with each other, the differences in power appear marginal. In accord with intuition, there are weak indications that the centre-direction tests perform better if $\boldsymbol{\mu}$ is close to the centre of the positive orthant, whereas the tests from subsection 3.2 are superior if $\boldsymbol{\mu}$ is close to the edges of the orthant. However, in general the performances are very similar. In the light of this

conclusion, the test based on (9), (10) and the one based on (7), (8) have the advantage on being easiest to implement on a computer.

APPENDIX A

Proof 1. All matrices $\mathbf{A}^*$ that fulfill (4) and (5) are of the form $\mathbf{A}^* = \mathbf{Q}\mathbf{A}_0$ with $\mathbf{Q} = \mathbf{Q}_2 \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{Q}_{p-1} \end{pmatrix} \mathbf{Q}_2'$, where $\mathbf{Q}_{p-1}$ is any orthogonal $(p-1) \times (p-1)$-matrix. We have $\mathbf{A}_0'\mathbf{Q}'\mathbf{Q}\mathbf{A}_0 = \mathbf{A}_0'\mathbf{A}_0 = \mathbf{G}^{-1}$ for all $\mathbf{Q}$ and

$$\operatorname{tr}(\mathbf{Q}\mathbf{A}_0\big(\operatorname{Diag}(\mathbf{G}^{-1})\big)^{-1/2}) = tr\left(\begin{pmatrix} 1 & 0 \\ 0 & \mathbf{Q}_{p-1} \end{pmatrix} \mathbf{Q}_2'\mathbf{A}_0\big(\operatorname{Diag}(\mathbf{G}^{-1})\big)^{-1/2}\mathbf{Q}_2\right).$$

Thus, $\operatorname{tr}(\mathbf{Q}\mathbf{A}_0(\operatorname{Diag}(\mathbf{G}^{-1}))^{-1/2}) = tr\begin{pmatrix} a_{11} & \mathbf{a}_{12}' \\ \mathbf{Q}_{p-1}\mathbf{a}_{(0)21} & \mathbf{Q}_{p-1}\mathbf{A}_{(0)22}) \end{pmatrix}$. Note that a_{11} and $\mathbf{a}_{12}$ are the same for every matrix that fulfills (4) and (5).

Obviously, $\operatorname{tr}(\mathbf{Q}\mathbf{A}_0(\operatorname{Diag}(\mathbf{G}^{-1}))^{-1/2})$ is maximized, if $\operatorname{tr}(\mathbf{Q}_{p-1}\mathbf{A}_{(0)22})$ is at a maximum. This maximum is attained for the symmetric, positive definite root $\mathbf{E}\Lambda^{1/2}\mathbf{E}'$ from the eigenvalue decomposition of $\mathbf{A}_{(0)22}'\mathbf{A}_{(0)22}$ (cf. Ahrens and Läuter,[22] (2.57)). The corresponding orthogonal transformation that transforms $\mathbf{A}_{(0)22}$ into $\mathbf{E}\Lambda^{1/2}\mathbf{E}'$ is $\mathbf{Q}_{p-1} = \mathbf{E}\Lambda^{1/2}\mathbf{E}'\mathbf{A}_{(0)22}^{-1}$. Consequently,

$$\mathbf{Q}_2'\mathbf{A}\big(\operatorname{Diag}(\mathbf{G}^{-1})\big)^{-1/2}\mathbf{Q}_2 = \begin{pmatrix} a_{11} & \mathbf{a}_{12}' \\ \mathbf{E}\Lambda^{1/2}\mathbf{E}'\mathbf{A}_{(0)22}^{-1}\mathbf{a}_{(0)21} & \mathbf{E}\Lambda^{1/2}\mathbf{E}' \end{pmatrix}$$

provides the $\mathbf{A}$ that maximizes $\operatorname{tr}(\mathbf{Q}\mathbf{A}_0(\operatorname{Diag}(\mathbf{G}^{-1}))^{-1/2})$ subject to the conditions. Since $\mathbf{G}$ has p distinct eigenvalues with probability 1, the root $\mathbf{E}\Lambda^{1/2}\mathbf{E}'$ is unique with probability 1. This shows that $\mathbf{A}$ is also uniquely determined with probability 1.

Proof 2. For $\mathbf{XD} \sim N_{n\times p}(\mathbf{1}_n\boldsymbol{\mu}'\mathbf{D}, \mathbf{D}'\mathbf{D})$, where $\mathbf{D}$ is a fixed positive definite diagonal matrix, the corresponding total sums-of-products matrix is $\mathbf{G}_D = \mathbf{D}'\mathbf{X}'\mathbf{X}\mathbf{D} = \mathbf{D}'\mathbf{G}\mathbf{D}$ with $\mathbf{A}_D = \mathbf{A}\mathbf{D}^{-1}$, $\bar{\mathbf{x}}_D' = \frac{1}{n}\mathbf{1}_n'\mathbf{X}\mathbf{D}$. We thus have $\sqrt{n}\mathbf{A}_D\bar{\mathbf{x}}_D = \sqrt{n}\mathbf{A}\bar{\mathbf{x}}$, so that the test statistic is unaffected by the scale of measurement of the individual variables. An analoguous argument regarding order-invariance applies, if $\mathbf{D}$ is a permutation matrix.

Proof 3. Under H_0, we have the two independent statistics

$$\sqrt{n}\bar{\mathbf{x}} \sim N(\mathbf{0}, \Sigma), \quad \mathbf{W} = \mathbf{X}'\mathbf{X} - n\bar{\mathbf{x}}\bar{\mathbf{x}}' \sim W_p(\Sigma, n-1).$$

The joint pdf of $\sqrt{n}\bar{\mathbf{x}}$ and $\mathbf{W}$ is proportional to

$$|\mathbf{W}|^{\frac{n-p}{2}-1}\exp\left(\operatorname{tr}\left(-\frac{1}{2}\Sigma^{-1}(\mathbf{W}+n\bar{\mathbf{x}}\bar{\mathbf{x}}')\right)\right).$$

Thus, the joint pdf of $\sqrt{n}\bar{\mathbf{x}}$ and $\mathbf{X}'\mathbf{X}=\mathbf{W}+n\bar{\mathbf{x}}\bar{\mathbf{x}}'$ is proportional to

$$\left|\mathbf{X}'\mathbf{X}\right|^{\frac{n-p}{2}-1}\cdot\left(1-n\bar{\mathbf{x}}'(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}}\right)^{\frac{n-p}{2}-1}\exp\left(\operatorname{tr}\left(-\frac{1}{2}\Sigma^{-1}\mathbf{X}'\mathbf{X}\right)\right),$$

since $|\mathbf{X}\mathbf{X}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'|=|\mathbf{X}\mathbf{X}'|\cdot(1-n\bar{\mathbf{x}}'(\mathbf{X}'\mathbf{X})^{-1}\bar{\mathbf{x}})$. Now, $\mathbf{A}'\mathbf{A}=(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{u}=\sqrt{n}\mathbf{A}\bar{\mathbf{x}}$. The Jacobian of the transformation from $\sqrt{n}\bar{\mathbf{x}}$ to $\mathbf{u}$ is $|\mathbf{A}'\mathbf{A}|^{-1/2}=|\mathbf{X}'\mathbf{X}|^{1/2}$, so that the joint pdf of $\mathbf{u}$ and $\mathbf{X}'\mathbf{X}$ is proportional to

$$\left|\mathbf{X}'\mathbf{X}\right|^{\frac{n-p-1}{2}}\cdot\left(1-\mathbf{u}'\mathbf{u}\right)^{\frac{n-p}{2}-1}\exp\left(\operatorname{tr}\left(-\frac{1}{2}\Sigma^{-1}\mathbf{X}'\mathbf{X}\right)\right),$$

and hence the pdf of $\mathbf{u}$ is proportional to

$$\left(1-\mathbf{u}'\mathbf{u}\right)^{\frac{n-p}{2}-1},\quad \mathbf{u}'\mathbf{u}\le 1. \tag{14}$$

Suppose now, the vector $\mathbf{u}$ is partitioned into $\mathbf{u}'=\left(\mathbf{u}_1'\ \mathbf{u}_2'\right)$, where $\mathbf{u}_1=(u_{i1})$ is a $p-k$-dimensional vector, $k=0,\ldots,p-1$. Let $\mathbf{u}_2=(u_{i2})$ and

$$\mathbf{v}=\frac{\mathbf{u}_2}{\sqrt{1-\mathbf{u}_1'\mathbf{u}_1}}. \tag{15}$$

The Jacobian of the transformation from $\mathbf{u}_2$ to $\mathbf{v}=(v_i)$ is $(1-\mathbf{u}_1'\mathbf{u}_1)^{k/2}$, so it follows from (14) that the joint pdf of $\mathbf{u}_1$ and $\mathbf{v}$ is proportional to

$$(1-\mathbf{u}_1'\mathbf{u}_1)^{\frac{n-p}{2}-1+\frac{k}{2}}\cdot(1-\mathbf{v}'\mathbf{v})^{\frac{n-p}{2}-1},\quad \mathbf{u}_1'\mathbf{u}_1\le 1,\ \mathbf{v}'\mathbf{v}\le 1. \tag{16}$$

Hence, the pdf of $\mathbf{u}_1$ is proportional to

$$(1-\mathbf{u}_1'\mathbf{u}_1)^{\frac{n-p+k}{2}-1},\quad \mathbf{u}_1'\mathbf{u}_1\le 1. \tag{17}$$

Since $\mathbf{u}$ has a spherical distribution, its distribution function is invariant to permutations of its components. Furthermore, it can be seen from (15) and (16) that the marginal distribution of $\mathbf{u}_1$ and its conditional distribution given $u_{i2}\le 0$ are the same. Finally, the additional condition $u_{i1}\ge 0$ changes the density (17) only by the multiplicative constant 2^{p-k}. Thus, (17) also gives the conditional pdf of the positive components of $\mathbf{u}^*=(u_1^*,\ldots,u_p^*)$ with $u_i^*=\max(0,u_i)$ under the condition that k components u_i are smaller than zero.

From (17), the conditional density of $\mathbf{u}^{*\prime}\mathbf{u}^{*}$ given $K = k$, $k = 0, \ldots, p-1$, is

$$f_{\overline{u}^2|K}(r|k) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{p-k}{2}\right)\Gamma\left(\frac{n-p+k}{2}\right)} \cdot r^{\frac{p-k}{2}-1}(1-r)^{\frac{n-p+k}{2}-1}, \quad r \in (0,1]$$

by Lemma 3.2.3 of Srivastava and Khatri.[2] Note that this is the density of the Beta-distribution $B(\frac{p-k}{2}, \frac{n-p+k}{2})$. Since $f_{\overline{u}^2}(r) = \sum_{k=0}^{p} P(K=k) \cdot f_{\overline{u}^2|K}(r|k)$, (11) follows from $P(K=k) = \binom{p}{k}/2^p$. □

Remark. It can be shown that Theorem 1 holds for any $\mathbf{X}$ whose pdf is given by $f(\Sigma^{-1}\mathbf{X}'\mathbf{X})$.

APPENDIX B

Covariance matrices used in Table 3

$$\Sigma_1 = \begin{pmatrix} 478.51 & 249.36 & -12.65 & 90.13 \\ 249.36 & 528.65 & 145.91 & 159.80 \\ -12.65 & 145.91 & 414.47 & 93.64 \\ 90.13 & 159.80 & 93.64 & 189.47 \end{pmatrix},$$

$$\Sigma_2 = \begin{pmatrix} 545.78 & 213.78 & 316.44 & 295.11 \\ 213.78 & 330.88 & 113.96 & 120.80 \\ 316.44 & 113.96 & 1049.45 & 698.39 \\ 295.11 & 120.80 & 698.39 & 753.03 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 13.6 & 1.6 & 6.7 & -3.7 \\ 1.6 & 4.0 & 1.9 & 3.0 \\ 6.7 & 1.9 & 7.9 & 4.1 \\ -3.7 & 3.0 & 4.1 & 32.4 \end{pmatrix}, \quad \Sigma_4 = \begin{pmatrix} 1076 & 698 & 708 & 789 \\ 698 & 1105 & 597 & 868 \\ 708 & 597 & 1056 & 712 \\ 789 & 868 & 712 & 1922 \end{pmatrix}$$

ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada. The first author's research was done during

a postdoctoral year at the University of Toronto. The authors are grateful to the referee whose valuable comments helped to improve this paper.

REFERENCES

1. Tang, D.-I., Gnecco, C.; Geller, N. An Approximate Likelihood Ratio Test for a Normal Mean Vector with Nonnegative Components with Application to Clinical Trials. Biometrika **1989**, *76*, 577–583.
2. Srivastava, M.S.; Khatri, C.G. *An Introduction to Multivariate Statistics*; Elsevier: New York, 1979.
3. Kudo, A. A Multivariate Analogue of the One-Sided Test. Biometrika **1963**, *50*, 403–418.
4. Nüesch, P. On the Problem of Testing Location in Multivariate Problems for Restricted Alternatives. Annals of Mathematical Statistics **1966**, *37*, 113–119.
5. Schaafsma, W.; Smid, L.J. Most Stringent Somewhere most Powerful Tests Against Alternatives Restricted by a Number of Inequalities. Annals of Mathematical Statistics **1966**, *37*, 1161–1172.
6. Perlman, M.D. One-Sided Testing Problems in Multivariate Analysis. Annals of Mathematical Statistics **1969**, *40*, 549–567.
7. Robertson, T.; Wright, F.T.; Dykstra, R.L. *Order Restricted Statistical Inference*; Wiley: New York, 1988.
8. O'Brien, P.C. Procedures for Comparing Samples with Multiple Endpoints. Biometrics **1984**, *40*, 1079–1087.
9. Follmann, D. Multivariate Tests for Multiple Endpoints in Clinical Trials. Statistics in Medicine **1995**, *14*, 1163–1175.
10. Follmann, D. A Simple Multivariate Test for One-sided Alternatives. Journal of the American Statistical Association **1996**, *91*, 854–861.
11. Läuter, J.; Glimm, E.; Kropf, S. New Multivariate Tests for Data with an Inherent Structure. Biometrical Journal **1996**, *38*, 5–23.
12. Läuter, J.; Glimm, E.; Kropf, S. Multivariate Tests Based on Left-spherically Distributed Linear Scores. Annals of Statistics **1998**, *26*, 1972–1988.
13. Läuter, J.; Kropf, S.; Glimm, E. Exact Stable Multivariate Tests for Applications in Clinical Research. In *Proceedings of the Biopharmaceutical Section of the American Statistical Association*, 1998; 46–55.
14. Wang, Y.; McDermott, M.P. Conditional Likelihood Ratio Test for a Normal Mean Vector. Journal of the American Statistical Association **1998**, *93*, 380–386.

15. McDermott, M.P. Generalized Orthogonal Contrast Tests for Homogeneity of Ordered Means. The Canadian Journal of Statistics **1999**, *27*, 457–470.
16. Srivastava, M.S.; Hirotsu, C.; Aoki, S.; Glimm, E. Multivariate One-Sided Tests. In *Data Analysis from Statistical Foundations—A Festschrift in Honour of the 75th Birthday of D.A.S. Fraser*, Nova Science Publishers, 2001; 387–401.
17. Fraser, D.A.S.; Guttman, I.; Srivastava, M.S. Inference for Treatment and Error in Multivariate Analysis. Biometrika **1991**, *78*, 565–572.
18. Reitmeir, P.; Wassmer, G. One-Sided Multiple Endpoint Testing in Two-Sample Comparisons. Communications in Statistics B: Computation and Simulation **1996**, *25*, 99–117.
19. Fang, K.-T., Zhang, Y.-T. *Generalized Multivariate Analysis*; Springer, Berlin and Academic Science Press: Beijing, 1990.
20. Glimm, E.; Srivastava, M.S. *Technical Report No. 2012: Multivariate Tests of Normal Mean Vectors with Restricted Alternatives*. Technical Report Series, Department of Statistics, University of Toronto, 2000.
21. Wollan, P.C.; Dykstra, R.L. Minimizing Linear Inequality Constrained Mahalanobis Distances. Journal of the Royal Statistical Society B: Applied Statistics **1987**, *36*, 234–240.
22. Ahrens, H.; Läuter, J. *Mehrdimensionale Varianzanalyse*, Akademie-Verlag: Berlin, 1981.

Directional multivariate tests rejecting null and negative effects in all variables

Ekkehard Glimm[*,1] and **Jürgen Läuter**[2]

[1] Novartis Pharma AG, Novartis Campus, 4056 Basel, Switzerland
[2] Otto-von-Guericke-Universität Magdeburg, Mittelstr. 2/151, 39114 Magdeburg, Germany

Received 10 October 2009, revised 19 April 2010, accepted 23 May 2010

This paper suggests two directional multivariate tests that aim at establishing superiority of a treatment over a control in at least one of several endpoints that are assumed to have a multivariate normal distribution. One of these tests is a one-sided, scale-invariant version of the classical Hotelling T^2-test. The other is based on a summary score with weights derived from the data. Both tests overcome an important shortcoming of previous "one-sided" multivariate suggestions, namely that the null hypothesis was restricted to a single point in the multidimensional parameter space. The derivation of the tests is supplemented by simulations investigating their performance and by the application in an osteoporosis trial.

Key words: Directional alternatives; Multivariate tests.

Supporting Information for this article is available from the author or on the WWW under http://dx.doi.org/10.1002/bimj.200900254

1 Background and motivation

This paper deals with the problem of testing whether at least one of many variables in a multivariate distribution has a mean which is larger than zero. This problem arises, for example, in clinical trials where treatment success is assessed by many endpoints and it is desired to establish that at least one of them shows a positive response to the treatment. The problem has received considerable attention in recent years (Dunnett and Tamhane, 1992; Tamhane *et al*., 1996; Cai and Sarkar, 2006; Röhmel *et al*., 2006; Chuang-Stein *et al*., 2007). Some of the papers consider situations in which the correlation structure of the endpoints is either completely known, or has a sparse structure described by very few parameters, or is otherwise known to be restricted (e.g. assuming that all pairwise correlations are positive). We do not discuss such situations in this paper. Excluding these methods, the topic can also be approached from a multiple testing perspective by considering univariate tests for the individual endpoints and combining them on the basis of the closed test procedure (Marcus *et al*. 1976) using Bonferroni's inequality. Undoubtedly, this approach has many advantages. Namely, it is simple to implement and if there are few variables, the power loss relative to methods that assume the knowledge of the correlation between endpoints (such as, e.g. Dunnett's test, Dunnett 1955) is small. If the primary interest is in the investigation and interpretation of single variables in isolation, these methods are most appropriate. They are less appropriate and—due to the inherent conservatism of the Bonferroni adjustment—less powerful, if a deviation from the null hypothesis of no treatment effect, say, manifests itself in moderate elevations of the values of several

*Corresponding author: e-mail: ekkehard.Glimm@Novartis.com, phone: +41-61-3240173, Fax: +41-61-3243039

variables without a single dominant one. In such cases, multivariate statistical methods are more appropriate.

Traditionally, multivariate statistical inference has focused on invariant methods. For example, Hotelling's famous T^2-test is the uniformly most powerful invariant test for the mean of a multivariate normal distribution (cf. Anderson 2003). Invariant tests are characterized by the assumption that only the distance of the true parameter value from the null hypothesis, not its direction, are relevant. The advances of bioinformatics with its huge data sets of correlated variables, e.g. in gene expression data, have recently brought about a revival of research into multivariate statistics, in particular in the case of sample sizes n that are smaller than the number of variables p. Older research on this topic (Box 1954, Dempster 1958) has recently been expanded upon by several authors (Srivastava and Fujikoshi 2006, Srivastava and Du 2008) and generalized, for example, to non-parametric statistics (Munzel and Brunner 2000, Oja and Randles 2004, Bathke and Harrar 2008). These suggestions are no longer invariant. However, they are so due to technical limitations, not by purpose. In contrast, our focus in this paper is on purposefully directional multivariate tests. Furthermore, while some of the methods we explore in the sequel can be applied in the case of $n < p$, this is only a secondary aspect of the investigations. Rather, potential applications for the suggested methods will usually have $n > p$ and will also have few variables, maybe up to 10 at most.

The investigations of this paper were in part inspired by a phase II clinical trial in osteoporosis. In this trial, several endpoints measured treatment effects on systemic aspects (such as pain relief and flexibility of joints) and others physiologic aspects (such as joint space narrowing and cartilage volume). It was of course hoped that a benefit could be established regarding both of these aspects, but it was unclear if this would be the case and how strong a benefit (if existent) would manifest itself in the various endpoints. In Section 6, we will discuss the analysis of data from such a trial. For confidentiality reasons, the numbers in that section are not from the real trial.

For simplicity, we introduce our suggestions for the one-sample case. The two-sample case is a straightforward extension (see Section 4). In terms of an application, we can think of the response being a difference between a post-treatment and a baseline measurement in multiple endpoints. Assume that we observe p response variables on each of n individuals. The data are arranged in an $n \times p$ matrix $\mathbf{X}$. A row of $\mathbf{X}$ represents the p responses of an individual; hence, rows are assumed to be stochastically independent and follow a p-dimensional normal distribution with unknown mean vector $\boldsymbol{\mu} = (\mu_i)_{i=1,\ldots,p}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. If there is no treatment effect on any of the endpoints, we have $\boldsymbol{\mu} = \mathbf{0}$.

Hotelling's T^2-test tests the hypothesis $H_0 : \boldsymbol{\mu} = \mathbf{0}$ against the general, "non-directional" alternative $A : \boldsymbol{\mu} \neq \mathbf{0}$. Its power only depends on the Mahalanobis distance $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ of $\boldsymbol{\mu}$ from zero. For practical applications with "directional" questions, this is often an unsuitable property, as positive and negative deviations from the null hypothesis are treated equally. Clinical trial applications of multivariate statistics are often faced with the problem of appropriately generalizing the concept of the univariate *one-sided* test. This has led to the derivation of numerous "directional" multivariate tests. All of these suggestions aim at a restricted alternative hypothesis, some of them explicitly stating the alternative, others doing so only implicitly (e.g. O'Brien 1984, Läuter 1996). Most notably, Kudo (1963), Nüesch (1966) and Perlman (1969) have derived the likelihood-ratio (LR) test of H_0 against the "one-sided" alternative $A: \boldsymbol{\mu} > \mathbf{0}$. Other restricted alternatives have also been considered. Silvapulle and Sen (2004) provide an overview.

Unfortunately, as has been pointed out by Silvapulle (1997), Perlman and Wu (2004) and Röhmel *et al.* (2006), the LR test has some serious drawbacks regarding its practical application. First, it is computationally very demanding. More seriously, the test has an intuitively unappealing property: It can lead to rejection of the null hypothesis in favor of the alternative, if all observed estimates $\bar{x}_i$ of μ_i are negative. This might happen because H_0 is restricted to a single point in space, such that the situation $\mu_i < 0$ is excluded *a priori* from consideration. Silvapulle (1997) gives a nice illustration of this shortcoming which also affects many suggestions that have been made to overcome the com-

putational complications of the LR test (e.g. Schaafsma and Smid 1966, Tang *et al.* 1989, Tang *et al.* 1993, Glimm *et al.* 2002).

In this paper, we suggest two multivariate tests that strictly keep α for the entire negative orthant. Thus, they allow to claim a statistically significant positive effect in at least one of the p response variables. In Sections 2 and 3, respectively, these two tests are introduced. After briefly discussing the two-sample case in Section 4, the power of the suggestions is compared via simulation in Section 5. In Section 6, the application to data from the osteoporosis trial is presented and discussed.

2 Directional Hotelling test

Follmann (Follmann 1995, Follmann 1996) suggested an alternative to the LR test which is particularly easy to implement. The method converts Hotelling's T^2-test into a directional test by requiring the extra condition $\bar{\mathbf{x}}' \cdot \mathbf{1}_p \geq 0$ for rejection, where $\mathbf{1}_p$ is a vector of p ones and $\bar{\mathbf{x}}$ is the usual least-squares estimate of $\boldsymbol{\mu}$. Hotelling's T^2-test accepts $H_0 : \boldsymbol{\mu} = \mathbf{0}$ if $\mathbf{0}$ lies in an ellipsoid with center $\bar{\mathbf{x}}$. By introducing the extra condition, this acceptance region is modified to a half-space and a half-ellipsoid. The T^2-test is performed at level 2α, i.e. with a contracted ellipsoid, to maintain the preassigned test level α.

Obviously, this approach avoids the case where H_0 is rejected with $\bar{x}_i < 0$ for all $i = 1, \ldots, p$. However, the issue is not entirely resolved: One cannot conclude from the rejection of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ that every "shifted" hypothesis $H_0^{\boldsymbol{\mu}} : \mu_i \leq 0$, for all $i = 1, \ldots, p$ where μ_i are the elements of $\boldsymbol{\mu}$, can also be rejected. In addition, Follmann's test is not scale-invariant.

We now give a modification of Follmann's test which (i) renders it scale-invariant and (ii) uses the extended null hypothesis $H_0^{\text{orth}} : \mu_i \leq 0$ for $i = 1, \ldots, p$. Condition (ii) demands that the test level α is kept for each fixed vector $\boldsymbol{\mu}$ of the negative orthant. A weaker version of this test considers the null hypothesis $H_0^{\text{corn}} : u_i < \mu_i \leq 0$ for $i = 1, \ldots, p$ with given fixed values of u_i.

If H_0^{orth} is rejected, we can conclude that not all variables have zero or negative mean values, i.e. in at least one of the p variables there is a positive response. However, the unique identification of such a "positive variable" would demand further testing steps, for example, the application of the closure principle by Marcus *et al.* (1976).

Let

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}'_{(1)} \\ \vdots \\ \mathbf{x}'_{(n)} \end{pmatrix} \sim N_{n \times p}(\mathbf{1}_n \boldsymbol{\mu}', \mathbf{I}_n \otimes \Sigma) \tag{1}$$

be the $n \times p$ matrix of n individuals, each having observations from p endpoints with means $\boldsymbol{\mu} = (\mu_i)_{i=1,\ldots,p}$ and common positive-definite covariance matrix $\boldsymbol{\Sigma}$, where $n > p$. For the sake of convenience, notation does not distinguish between random variables and their realizations.

The usual least-squares estimates are $\bar{\mathbf{x}} = (\bar{x}_i)_{i=1,\ldots,p} = (1/n) \cdot \sum_{j=1}^{n} \mathbf{x}_{(j)}$ for $\boldsymbol{\mu}$ and $\mathbf{S} = (1/(n-1))\mathbf{G} = (1/(n-1)) \sum_{j=1}^{n} (\mathbf{x}_{(j)} - \overline{\mathbf{x}}) (\mathbf{x}_{(j)} - \overline{\mathbf{x}})'$ for $\boldsymbol{\Sigma}$. A minimum-volume $(1-\alpha)$ confidence region for $\boldsymbol{\mu}$ is given by the ellipsoid around $\bar{\mathbf{x}}$

$$C_{1-\alpha}(\bar{\mathbf{x}}, \mathbf{G}) = \{\boldsymbol{\mu}_0 \text{ with } \frac{(n-p)n}{p} (\boldsymbol{\mu}_0 - \bar{\mathbf{x}})' \mathbf{G}^{-1} (\boldsymbol{\mu}_0 - \bar{\mathbf{x}}) < F_{1-\alpha}(p, n-p)\} \tag{2}$$

where $F_{1-\alpha}(p, n-p)$ is the $(1-\alpha)$ quantile of the F-distribution with p and $n-p$ degrees of freedom. Hotelling's T^2-test rejects $H_0^{\boldsymbol{\mu}}$ if and only if $\boldsymbol{\mu} \notin C_{1-\alpha}(\bar{\mathbf{x}}, \mathbf{G})$.

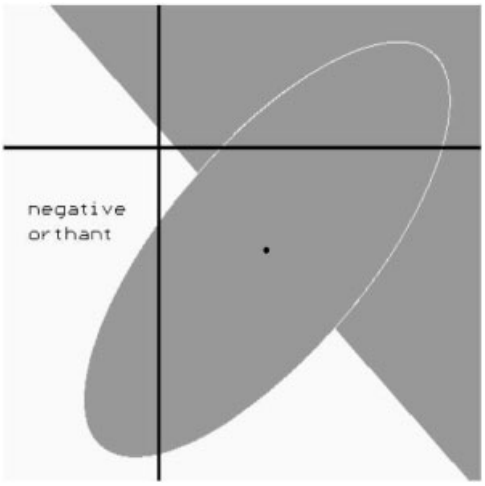


Figure 1 Confidence region corresponding to the directional Hotelling test for H_0^{orth}(Procedure I). The test fails significance, because the ellipse intersects the negative orthant ($p = 2$).

To test the extended null hypothesis H_0^{orth}, we suggest the modified directed confidence region

$$C_{1-2\alpha}(\bar{\mathbf{x}}, \mathbf{G}) \cup \left\{ \boldsymbol{\mu}_0 \text{ with } \sum_{i=1}^{p} \frac{\mu_{0i} - \bar{x}_i}{\sqrt{g_{ii}}} > 0 \right\}, \tag{3}$$

where g_{ij}, $i, j = 1,\ldots,p$, are the elements of the sums-of-products matrix $\mathbf{G}$. This way, a scale-invariant test arises. Each of the p endpoints is standardized with its corresponding standard deviation. Region (3) consists of a half-space and a half-ellipsoid (Fig. 1). The corresponding multivariate test is given by

Procedure I: H_0^{orth} is rejected if no vector $\boldsymbol{\mu}$ of the negative orthant lies in confidence region (3).

The probability that $\boldsymbol{\mu}$ is not in the directed confidence region (3) is exactly α because

(i) the probability of $\boldsymbol{\mu} \notin C_{1-2\alpha}(\bar{\mathbf{x}}, \mathbf{G})$ is 2α (for $\alpha \leq 0.5$),
(ii) the probability of $\boldsymbol{\mu} \notin \{\boldsymbol{\mu}_0 \text{ with } \sum_{i=1}^{p} (\mu_{0i} - \bar{x}_i)/\sqrt{g_{ii}} > 0\}$ is 0.5,
(iii) the boundary line of the half-space $\{\boldsymbol{\mu}_0 \text{ with } \sum_{i=1}^{p} (\mu_{0i} - \bar{x}_i)/\sqrt{g_{ii}} > 0\}$ goes through the center $\bar{\mathbf{x}}$ of the ellipsoid such that $C_{1-2\alpha}(\bar{\mathbf{x}}, \mathbf{G})$ is cut into two halves with probability mass $(1-2\alpha)/2$. Consequently, the region excluded by (3) is $0.5-(1-2\alpha)/2 = \alpha$.

This reasoning is valid for any dimension p. (ii) holds because the matrix $\mathbf{G}$ is stochastically independent of the mean vector $\bar{\mathbf{x}}$. The "multiple" rejection condition for H_0^{orth}(i.e. requiring that (3) excludes *all* $\boldsymbol{\mu}$ with $\mu_i \leq 0$) results in a further reduced significance level for each fixed $\boldsymbol{\mu}$ in the negative orthant. Hence, the test of H_0^{orth} always keeps the significance level α.

In order to reject H_0^{orth}, procedure **I** requires that $\bar{\mathbf{x}} \neq 0$ and $\sum_{i=1}^{p} \bar{x}_i/\sqrt{g_{ii}} \geq 0$ hold, such that $\bar{\mathbf{x}}$ must not be in the negative orthant. In the following, it is assumed that these conditions are met. In addition, we need to check that

$$\min_{\boldsymbol{\mu} \in \text{neg. orthant}} \frac{(n-p)n}{p} (\boldsymbol{\mu} - \bar{\mathbf{x}})' \mathbf{G}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) \geq F_{1-2\alpha}(p, n-p). \tag{4}$$

Since the quadratic form in this expression is a convex function of $\boldsymbol{\mu}$, its unrestricted unique minimum is at $\boldsymbol{\mu} = \bar{\mathbf{x}}$ and its values are monotonously increasing in all directions if one moves away

from this minimum. Furthermore, convexity implies that the restricted minimum value in (4) must be on the boundary of the negative orthant, i.e. at least one element μ_i must be 0. In addition, this restricted minimum is unique. Consequently, we can find the minimum of the quadratic form within the negative orthant by repeating the following steps:

(i) Fix some μ_i to be 0.
(ii) Obtain the minimum of $(n-p)n/p(\boldsymbol{\mu}-\bar{\mathbf{x}})'\mathbf{G}^{-1}(\boldsymbol{\mu}-\overline{\mathbf{x}})$ with this restriction. This may result in some or all of the non-null μ_i's being positive.

We have to do this for all 2^p-1 possible combinations of zeros in places of μ_i's. The one ($\boldsymbol{\mu}^*$, say) that provides the minimum value among the solutions with all components $\mu_i \leq 0$ is the minimum sought in (4). In special cases, this may be $\boldsymbol{\mu}^* = \mathbf{0}$. The p-value of the test is the solution α^* of

$$\frac{(n-p)n}{p}(\boldsymbol{\mu}^*-\bar{\mathbf{x}})'\mathbf{G}^{-1}(\boldsymbol{\mu}^*-\bar{\mathbf{x}}) = F_{1-2\alpha^*}(p, n-p).$$

Obviously, we reject H_0^{orth} if $\alpha^* \leq \alpha$.

Each of the single minimization problems is easy to solve. If the p variables are partitioned into two subsets corresponding to

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \bar{\mathbf{x}} = \begin{pmatrix} \bar{\mathbf{x}}_1 \\ \bar{\mathbf{x}}_2 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix},$$

then the minimum of the quadratic form subject to $\boldsymbol{\mu}_2 = \mathbf{0}$ is given by $\boldsymbol{\mu}_1^* = \bar{\mathbf{x}}_1 - \mathbf{G}_{12}\mathbf{G}_{22}^{-1}\bar{\mathbf{x}}_2$. This result is obtained from well-known properties of the conditional normal distribution (e.g. Mardia *et al.* (1979), Chapter 3).

It follows that in the special case of $p = 2$, H_0^{orth} can be rejected if $\bar{\mathbf{x}} \neq \mathbf{0}$, $\bar{x}_1/\sqrt{g_{11}} + \bar{x}_2/\sqrt{g_{22}} \geq 0$ and none of the three points

$$\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\mu} = \begin{pmatrix} \mathbf{0} \\ \bar{\mathbf{x}}_2 - \mathbf{g_{21}}\mathbf{g_{11}^{-1}}\,\bar{\mathbf{x}}_1 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \bar{\mathbf{x}}_1 - \mathbf{g_{12}}\mathbf{g_{22}^{-1}}\,\bar{\mathbf{x}}_2 \\ \mathbf{0} \end{pmatrix}$$

is in the negative orthant as well as in the $(1-2\alpha)$ ellipse.

In comparison with Follmann's test, our proposal extends the test's acceptance region by certain points $\boldsymbol{\mu}$ in the intersection of ellipsoid and negative orthant (see Fig. 1). Naturally, this is only possible at the expense of some power. In some cases, the weaker null hypothesis H_0^{corn} mentioned above can be applied. Namely, if we know in advance that there are lower limits u_i for the parameters μ_i, we only have to check whether a vector $\boldsymbol{\mu}$ of the corresponding negative corner lies in the directed confidence region (3). Thus the power of the test can be increased.

3 Standardized sum test for H_0^{orth}

Läuter and co-workers (Läuter 1996, Läuter *et al.* 1996, Läuter *et al.* 1998) introduced the concept of spherical multivariate tests. These are exact multivariate tests of $H_0 : \boldsymbol{\mu} = \mathbf{0}$ based on low-dimensional scores calculated from the observed data. Their main advantage is that they can be used with arbitrarily high dimension p. In particular, p may be larger than the sample size n. The tests are not affine-invariant. Just like all the tests discussed in the introduction, however, these tests cannot be considered as tests of H_0^{orth}. With them it is also possible that for given data $\mathbf{X}$, H_0 is rejected, but $H_0^{\boldsymbol{\mu}_0} : \boldsymbol{\mu} = \boldsymbol{\mu}_0$is not for some $\boldsymbol{\mu}_0$ of the negative orthant.

Here, we will consider only one class of spherical tests, the so-called standardized sum (SS) tests. We assume that the covariance matrix $\mathbf{\Sigma}$ has positive diagonal elements $\sigma_{11},\ldots,\sigma_{pp}$. Positive definiteness of $\mathbf{\Sigma}$ is not required. The usual SS test in the one-sided version rejects H_0 if

$$t^0 = \sqrt{n-1}\frac{\sqrt{n}\bar{\mathbf{x}}'\mathbf{d}^0}{\sqrt{\mathbf{d}^{0\prime}\mathbf{G}\mathbf{d}^0}} \geq t_{1-\alpha}(n-1) \tag{5}$$

where $t_{1-\alpha}\ (n-1)$ is the $(1-\alpha)$ quantile of the t-distribution with $n-1$ degrees of freedom and $\mathbf{d}^0 = \left(1/(\sqrt{g_{ii}+n\,\bar{x}_i^2})\right)_{i=1,\ldots,p}$. The test is scale-invariant and has high power if all variables have roughly the same positive deviation from the null hypothesis in their respective scale and roughly equal correlations with each other (Läuter *et al.* 1996).

The corresponding test for $H_0^{\boldsymbol{\mu}_0} : \boldsymbol{\mu} = \boldsymbol{\mu}_0$is

$$t = \sqrt{n-1}\frac{\sqrt{n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{d}}{\sqrt{\mathbf{d}'\mathbf{G}\mathbf{d}}} \geq t_{1-\alpha}(n-1) \tag{6}$$

with $\mathbf{d} = \left(1/\sqrt{g_{ii}+n(\bar{x}_i-\mu_{0i})^2}\right)_{i=1,\ldots,p}$. Unfortunately, t does not generally increase if the parameters μ_{0i} decrease. Hence, rejection of H_0 does not translate into rejection of $H_0^{\boldsymbol{\mu}_0}$ for all $\boldsymbol{\mu}_0$ of the negative orthant.

Additional conditions are necessary to establish rejection of H_0^{orth}. In the following, a corresponding modification of test (5) will be derived. Let $\xi_1,\ldots,\xi_p$ be angles defined for $g_{ii}>0$ by

$$0<\xi_i<\pi, \quad \sin\xi_i = \sqrt{g_{ii}}d_i, \quad \cos\xi_i = \sqrt{n}(\bar{x}_i - \mu_{0i})d_i \tag{7}$$

for $i=1,\ldots,p$. Then the test (6) can be written as:

$$t = \frac{\sqrt{n-1}\sum_{i=1}^p \sqrt{n}(x_i-\mu_{0i})d_i}{\sqrt{\sum_{i=1}^p\sum_{h=1}^p d_h\sqrt{g_{hh}}r_{hi}\sqrt{g_{ii}}d_i}} = \frac{\sqrt{n-1}\sum_{i=1}^p cos\ \xi_i}{\sqrt{\sum_{i=1}^p\sum_{h=1}^p sin\ \xi_h r_{hi} sin\ \xi_i}} \geq t_{1-\alpha}(n-1). \tag{8}$$

Here, $r_{hi} = g_{hi}/\sqrt{g_{hh}g_{ii}}$ denotes the correlation coefficients from the residual matrix $\mathbf{G}$. Writing $\mathbf{R} = (r_{hi})_{h,i=1,\ldots,p}$, $\cos\xi = (\cos\xi_i)_{i=1,\ldots,p}$ and $\sin\xi = (\sin\xi_i)_{i=1,\ldots,p}$, (8) becomes

$$t = \sqrt{n-1}\frac{(\cos\xi)'1_p}{\sqrt{(\sin\xi)'\,\mathbf{R}\sin\xi}} \geq t_{1-\alpha}(n-1). \tag{9}$$

It is important to note that the parameters μ_{0i} are contained in the angles ξ_i, but do not appear otherwise. The test of H_0^{orth} is significant if inequality (9) holds for all $\boldsymbol{\mu}_0$ of the negative orthant.

As can be seen from (7), $\cot\xi_i = cos\ \xi_i/sin\ \xi_i$ increases and ξ_i decreases for fixed $\mathbf{G}$ if $\bar{x}_i - \mu_{0i}$ increases. Hence, the numerator $(\cos\xi)'\mathbf{1}_p$ of the t-ratio increases if $\boldsymbol{\mu}_0$ is moved from $\mathbf{0}$ into the negative orthant.

In contrast, the denominator can decrease or increase with such a move of $\boldsymbol{\mu}_0$, depending on the specific values of $\mathbf{R}$ and $\bar{\mathbf{x}}$. If all r_{hi} and all $\bar{x}_i$ are non-negative in an application, then $\xi_i \leq \pi/2$ holds such that $(\sin\xi)'\mathbf{R}\sin\xi$ decreases as μ_i decreases. Therefore, the rejection of $\boldsymbol{\mu}_0 = \mathbf{0}$ also implies that the rejection of all $\boldsymbol{\mu}_0$ in the negative orthant and the test of H_0^{orth} is finished. If, however, negative values r_{hi} or negative $\bar{x}_i$ arise, additional checks are necessary.

Let ξ^0 denote the vector of angles ξ_i for $\boldsymbol{\mu}_0 = \mathbf{0}$ and let ξ^{0+} be the modified vector obtained by replacing all components of ξ^0 which are larger than $\pi/2$ by $\pi/2$. Likewise, let $\mathbf{R}^+$ be the matrix obtained from $\mathbf{R}$ by replacing all negative elements with 0. Then we have the inequality

$$(\sin\xi)'\mathbf{R}\sin\xi \leq (\sin\xi^{0^+})'\mathbf{R}^+\sin\xi^{0^+}$$

in the negative orthant. Consequently,

$$t^{\text{orth}} = \sqrt{n-1}\frac{(\cos \xi^0)'1_p}{\sqrt{(\sin \xi^{0^+})'\mathbf{R}^+ \sin \xi^{0^+}}} \geq t_{1-\alpha}(n-1). \tag{10}$$

is a sufficient condition for the rejection of $H_0^{\text{orth}}(\alpha < 0.5)$.
To state the test procedure, we rewrite (10) without the angles ξ_i:
Procedure IIa: Reject H_0^{orth} if

$$t^{\text{orth}} = \sqrt{n-1}\frac{\sqrt{n}\bar{\mathbf{x}}'\mathbf{d}^0}{\sqrt{\mathbf{d}^{0^+\prime}\mathbf{G}^+\mathbf{d}^{0^+}}} \geq t_{1-\alpha}(n-1). \tag{11}$$

This provides a conservative test ($\alpha < 0.5$) of H_0^{orth}. Here, $\mathbf{d}^{0^+}$ is based on $\mathbf{d}^0$ with all $d_i^0 = 1/\sqrt{g_{ii}+n\bar{x}_i^2}$ replaced by $1/\sqrt{g_{ii}}$ if $\bar{x}_i < 0$ and $\mathbf{G}^+$ is $\mathbf{G}$ with all negative elements replaced by 0. This is the desired sharpening modification of the usual SS test (5).

A more conservative simplification of this test is given by
Procedure IIb: Reject H_0^{orth} if

$$\tilde{t}^{\text{orth}} = \sqrt{n-1}\frac{(\cos \xi^0)'\mathbf{1}_p}{\sqrt{\mathbf{1}_p'\,\mathbf{R}^+\mathbf{1}_p}} = \sqrt{n-1}\frac{\sqrt{n}\bar{\mathbf{x}}'\mathbf{d}^0}{\sqrt{\mathbf{1}_p'\,\mathbf{R}^+\mathbf{1}_p}} \geq t_{1-\alpha}(n-1). \tag{12}$$

Röhmel *et al.* (2006) have done a more detailed investigation of the case $p = 2$. They show that in order to establish significance with level α controlled in the whole negative orthant, it is sufficient to check the validity of the inequality (6) only for $\boldsymbol{\mu}_0 = \mathbf{0}$ and additionally, if $g_{12} < 0$, for the "vertices"

$$\boldsymbol{\mu}_0 = \begin{pmatrix} -\infty \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ -\infty \end{pmatrix}. \tag{13}$$

This results in the following modification:

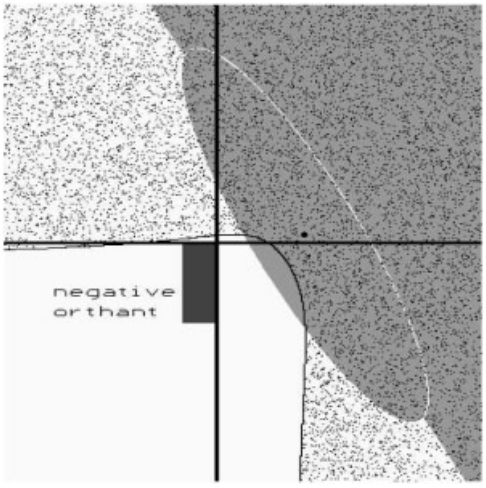


Figure 2 One-sided confidence region of the SS test (dotted area) and of the directional Hotelling test (light-grey half-plane and half-ellipse). Rectangle (dark-grey) corresponding to a restricted null hypothesis H_0^{corn}.

Procedure IIc: ($p = 2$ only) Reject H_0^{orth} if (5) holds and either $g_{12} \geq 0$, or (in case of $g_{12} < 0$)

$$\begin{aligned} t &= \sqrt{n-1}\,\frac{1+\sqrt{n}\,\bar{x}_2\,d_2^0}{\sqrt{g_{22}}d_2^0} \geq t_{1-\alpha}(n-1) \quad \text{and} \\ t &= \sqrt{n-1}\,\frac{\sqrt{n}\,\bar{x}_1\,d_1^0+1}{\sqrt{g_{11}}d_1^0} \geq t_{1-\alpha}(n-1). \end{aligned} \tag{14}$$

Procedure IIc thus provides an "optimal testing rule" in case of $p = 2$. In case of a restricted negative orthant, Procedure IIc can be modified to testing $H_0 : \boldsymbol{\mu}_0 = \mathbf{0}$ with (5) and additionally, if $g_{12} < 0$, the vertices

$$\boldsymbol{\mu}_0 = \begin{pmatrix} u_1 \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_0 = \begin{pmatrix} 0 \\ u_2 \end{pmatrix} \tag{15}$$

instead of (13).

Figure 2 shows the one-sided confidence regions and, correspondingly, the directional rejection regions of the SS test and the directional Hotelling test from Section 2 for an example with $p = 2$, $n = 5$. In this example, the covariance g_{12} is negative, and the SS test has a non-monotone behaviour: The small rectangle in the corner of the negative orthant belongs to the rejection region of the SS test so that the corresponding restricted null hypothesis H_0^{corn} is rejected. In contrast, H_0^{orth} cannot be rejected because parts of the whole negative orthant are intersecting with the SS test confidence region. The one-sided confidence region of the Hotelling test, a half-plane and a half-ellipse, does not intersect with the negative orthant. Therefore, the one-sided Hotelling test rejects H_0.

The monotonicity investigations performed here are "pointwise" for fixed values of $\bar{\mathbf{x}}$ and $\mathbf{G}$, not taking into account the multivariate distribution at hand. It is conceivable that the one-sided SS test (5) for $H_0 : \boldsymbol{\mu} = \mathbf{0}$ keeps the α-level when applied to normally distributed data without any modification for any hypothesis $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ with $\boldsymbol{\mu}_0$ in the negative orthant. This question has not yet been settled.

It is also worth noting that in practical applications with few variables and large sample sizes, the additional checks discussed in this section will very rarely play a role. If all correlations between variables are positive, no additional check is necessary. In the simulations done by Röhmel *et al.* (2006) for $p = 2$, there was no case where the additional vertex check (13) was necessary due to a non-monotonicity of the test statistic (5). As Röhmel *et al.* (2006) also discuss, the reason for this is that such monotonicities can only arise in cases with strong negative correlations and one variable having a very large effect.

4 The two-sample case

The previous sections have implicitly covered the two-sample case as well. Suppose there are n_k observations $\mathbf{x}_{(jk)} \sim N_p(\boldsymbol{\mu}_k,\ \boldsymbol{\Sigma})$ in groups $k = 1,\ 2$ and inference is concerned with tests of $H_0^{orth} : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is in the negative orthant, and with corresponding confidence regions for $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. All methods presented in Sections 2 and 3 are based on the complete, sufficient statistics $\bar{\mathbf{x}}$ and $\mathbf{G}$. In the two-sample case, these are essentially the same, if we re-define $\bar{\mathbf{x}}$ as $\bar{\mathbf{x}} := \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ and $\mathbf{G}$ as $\mathbf{G} : = \mathbf{G}_1 + \mathbf{G}_2$, where $\bar{\mathbf{x}}_k = \frac{1}{n_k}\Sigma_{j=1}^{n_k}\mathbf{x}_{(jk)}$ and $\mathbf{G}_k = \Sigma_{j=1}^{n_k}(\mathbf{x}_{(jk)} - \bar{\mathbf{x}}_k)\,(\mathbf{x}_{(jk)} - \bar{\mathbf{x}}_k)'$. The two-sample case is thus handled by using these re-definitions in Sections 2 and 3. The only other modifications necessary are a change of a constant in the total sums-of-products matrix and a change in the denominator degrees of freedom of F- and t-statistics, respectively. Regarding the former, $\mathbf{G} + n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'$ has to be replaced by $\mathbf{G}_1 + \mathbf{G}_2 + (n_1 n_2/(n_1 + n2))(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\mu}_0)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 - \boldsymbol{\mu}_0)'$ in other words, n is replaced by $n_1 n_2/(n_1 + n_2)$. Regarding degrees of freedom, $n-1$ needs to be changed to

n_1+n_2-2. The minor modifications this requires in the previous sections can be summarized as follows:

(i) In Section 2, formula (2), $(n-p)n/p$ has to be changed to $((n_1+n_2-1-p)/p)n_1n_2/(n_1+n_2)$ and the numerator degrees of freedom of the F-quantile from $n-p$ to n_1+n_2-1-p in addition to re-definition of $\bar{\mathbf{x}}$ and $\mathbf{G}$.
(ii) In Section 3, "n" has to be replaced with $n_1n_2/(n_1+n_2)$ and "$n-1$" with n_1+n_2-2 in addition to re-definition of $\bar{\mathbf{x}}$ and $\mathbf{G}$.

5 Simulation results

This section presents results obtained from comparing procedures I and II from Sections 2 and 3 by simulation. As mentioned previously, the procedures are not primarily intended for "huge" dimensions as they occur, for example, in microarray analyses with thousands of variables. First of all, the directional Hotelling test from Section 2 requires $n > p$. The standardized sum test from Section 3 does not require this, but directional hypotheses like H_0^{orth} are rarely relevant in applications with very many variables. The spherical multivariate tests introduced by Läuter and co-workers (Läuter 1996, Läuter et al. 1996, Läuter et al. 1998) are appropriate for large p.

We have tried to set up a few simulation scenarios as a compromise between situations that are likely in practical applications and cases that highlight specific properties of the suggested methods. Thus, we are restricting our attention to cases where all correlations between variables are positive and where the direction of deviations from the null hypothesis is the same in all variables. The results are summarized in the following Tables 1–5. All values are the results of 100 000 simulation runs.

Tables 1 and 2 investigate the power of the suggested tests in the two-sample case with $p=4$ variables. In these two tables, the Mahalanobis distance $\Delta^2=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ is set to a fixed value. Table 1 has the results for equally correlated, equally informative variables. Here and subsequently, we use the term "informative" to indicate "distance from the null hypothesis", e.g. in this case, all variables contribute equally to the Mahalanobis-distance from $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2=\mathbf{0}$ (they are "equally far apart" from the null hypothesis). As expected, the SS test performs very well in this situation. For the low sample size of $n_1=n_2=6$, the power of the directional Hotelling test suffers from instability problems that typically occur with the ordinary Hotelling test as well if the dimension is large in comparison with the sample size.

Table 2 has two equally informative variables and two that simply represent additional "noise" (no group differences and no correlation with each other and the two informative variables). In this

Table 1 Power of two-sample tests, $p=4$, $\alpha=0.05$, $\boldsymbol{\mu}_1-\boldsymbol{\mu}_2 \propto \mathbf{1}_p$, $\Sigma=(1-\rho)\cdot\mathbf{I}_p+\rho\cdot\mathbf{1}_p\mathbf{1}_p$

ρ	$n_1=n_2=6$, $\Delta^2=4$			$n_1=n_2=20$, $\Delta^2=1$		
	Direct. T^2	SS proc. IIa	SS proc. IIb	Direct. T^2	SS proc. IIa	SS proc. IIb
0	0.645	0.892	0.847	0.762	0.903	0.896
0.1	0.628	0.919	0.883	0.755	0.919	0.913
0.2	0.616	0.931	0.897	0.749	0.926	0.920
0.4	0.589	0.941	0.904	0.731	0.925	0.919
0.6	0.554	0.942	0.897	0.716	0.928	0.920
0.9	0.466	0.941	0.883	0.657	0.927	0.918

Table 2 Power of two-sample tests, $p = 4$, $\alpha = 0.05$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \propto (1\ 1\ 0\ 0)'$, correlation ρ between $\mathbf{x}_1$ and $\mathbf{x}_2$, 0 otherwise.

ρ	$n_1 = n_2 = 8$, $\Delta^2 = 4$			$n_1 = n_2 = 20$, $\Delta^2 = 1$		
	Direct. T^2	SS proc. IIa	SS proc. IIb	Direct. T^2	SS proc. IIa	SS proc. IIb
0	0.797	0.684	0.596	0.701	0.626	0.602
0.1	0.795	0.709	0.615	0.698	0.649	0.625
0.2	0.797	0.733	0.636	0.702	0.673	0.647
0.4	0.794	0.763	0.655	0.703	0.703	0.675
0.6	0.792	0.786	0.666	0.696	0.731	0.700
0.9	0.773	0.813	0.675	0.678	0.754	0.720

Table 3 Rejection probabilities of two-sample tests under H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $p = 4$, $= 0.05$, $n_1 = n_2 = 20$.

ρ	balanced covariances (as in Table 1)			Unbalanced covariances (as in Table 2)		
	Direct. T^2	SS proc. IIa	SS proc. IIb	Direct. T^2	SS proc. IIa	SS proc. IIb
0	0.025	0.035	0.033	0.026	0.035	0.033
0.1	0.022	0.044	0.041	0.025	0.036	0.033
0.2	0.019	0.048	0.044	0.024	0.037	0.034
0.4	0.014	0.050	0.046	0.023	0.038	0.035
0.6	0.011	0.049	0.044	0.021	0.039	0.035
0.9	0.005	0.049	0.044	0.018	0.040	0.036

Table 4 Power of two-sample tests, $p = 2$, $\alpha = 0.05$, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = (2\ 0)'$, $\boldsymbol{\Sigma} = \mathbf{I}_2$.

$n_1 = n_2$	Direct. T^2	SS (5)	SS proc. IIc	SS proc. IIa	SS proc. IIb
2	0.097	0.148	0.101	0.077	0
3	0.333	0.260	0.244	0.190	0.036
4	0.555	0.374	0.370	0.305	0.146
5	0.716	0.470	0.469	0.414	0.255
6	0.827	0.561	0.561	0.507	0.358
7	0.896	0.638	0.638	0.592	0.451
8	0.939	0.704	0.704	0.665	0.535
9	0.966	0.762	0.762	0.726	0.610
10	0.980	0.809	0.809	0.777	0.676

case, the SS test and the directional Hotelling test are similar in their performance with the SS test having slight advantages with highly correlated variables and the directional Hotelling performing a little better when correlations are low.

Table 3 shows the probability of rejection for the covariances from Tables 1 and 2 if $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ is true. The nominal level α is 5%, but all tests investigated here have to keep α for the composite hypothesis H_0^{orth}, not just H_0, so it is no surprise that α is not exhausted at $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. The tests are conservative in the sense that they do not exhaust the α-level anywhere. However, when correlations are high, the SS test procedure IIa comes very close to doing so. In contrast, the rejection prob-

Table 5 Rejection probability of two-sample tests under H_0: $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$, $\boldsymbol{\Sigma} = \mathbf{I}_2$, $p = 2$, $\alpha = 0.05$.

$n_1 = n_2$	Direct. T^2	SS (5)	SS proc. IIc	SS proc. IIa	SS proc. IIb
2	0.027	0.050	0.020	0.012	0
3	0.034	0.050	0.046	0.024	0.003
4	0.037	0.050	0.050	0.030	0.011
5	0.038	0.051	0.051	0.034	0.018
6	0.038	0.050	0.050	0.035	0.022
7	0.039	0.050	0.050	0.037	0.026
8	0.039	0.050	0.050	0.040	0.029
9	0.039	0.049	0.049	0.040	0.031
10	0.040	0.049	0.049	0.042	0.034

ability of the directional Hotelling test decreases with increasing positive correlations between the variables.

Table 4 shows how much power is lost by the orthant-related modifications of the directional SS test and the corresponding simplifications discussed in Section 3. The special case of $p = 2$ with one informative variable and one uncorrelated uninformative variable is considered. For $p = 2$, the "optimal" procedure IIc is available. This rule has less power than the original SS test (5) alone for extremely small sample sizes $n_i \leq 5$, but test (5) might not keep the level α for the entire negative orthant. For still very moderate sample sizes of $n_i > 5$, our simulations did not find any power loss due to the additional vertex checks. This is in line with Röhmel et al. (2006). In addition, the table gives the power of the more conservative directional test procedures IIa and IIb. Note that in any practical application, one would be allowed to do the simplest test procedure IIb first, if it is not significant, try procedure IIa and if this does not yield significance either, in case of $p = 2$, finally try procedure IIc. The purpose of Table 4 is not a power comparison between SS- and directional T^2 test. It is clear that in the situation of one informative and one uninformative variable, the SS test is inferior. The simulation results of the directed T^2 test reflect this.

In analogy to Table 3, Table 5 shows the rejection probabilities under H_0 in case of two independent variables with equal variance.

6 Application in a clinical trial

In a phase II clinical trial on an osteoporosis drug with two treatment groups (treatment and control), it was initially unclear whether the benefit of a new treatment over standard treatment would primarily be

(i) physiologic improvement of the knee, measured by joint space width (JSW) in mm,
(ii) better pain relief, measured by a pain score, or
(iii) better functional ability, measured by a function score.

Thus, the focus of this phase II trial was to establish a benefit in at least one of these indicators. A future phase III trial would then focus on the most promising variables.

For the corresponding tests, a level of $\alpha = 0.05$ was selected. The trial was performed with 32 patients per group. The three endpoints were investigated as change from baseline after 3 months of treatment. In all three variables, positive values indicate an improvement. It was expected that the treatment would yield better results in all three endpoints and that all three endpoints would be positively

Table 6 Observed means and covariances in the osteoporosis trial.

	Means			Covariance
Treatment	JSW	Pain score	Function score	
New	0.43	12.1	63.6	$\begin{pmatrix} 0.38 & 17.0 & 43 \\ 17.0 & 2763 & 3257 \\ 43 & 3257 & 12042 \end{pmatrix}$
Control	0.08	14.4	83.0	$\begin{pmatrix} 0.17 & 8.4 & 20 \\ 8.4 & 2752 & 2043 \\ 20 & 2043 & 7572 \end{pmatrix}$
Means : difference Covariance : pooled estimate	0.35	−2.4	−19.4	$\begin{pmatrix} 0.27 & 12.7 & 32 \\ 12.7 & 2758 & 2650 \\ 32 & 2758 & 9807 \end{pmatrix}$

correlated. However, it was suspected that trial duration might be too short for the pain and the function scores, resulting in large variability of them as well as in a lack of positive treatment effect.

Table 6 shows the results of the trial. It is clear that the three endpoints are not on the same scale. Only JSW produced a result in accord with expectations. Contrary to expectations, the results of the two scores turned out to be worse on average in the treatment group than in the control group. The pooled estimate of the correlation between the three endpoints was

$$\begin{pmatrix} 1 & 0.46 & 0.61 \\ 0.46 & 1 & 0.51 \\ 0.61 & 0.51 & 1 \end{pmatrix}.$$

Let $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ denote the treatment and the control mean, respectively, $\mathbf{S}$ the pooled covariance estimate, $\mathbf{G} = (n_1+n_2-2)\mathbf{S}$, and $\bar{\mathbf{x}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$the mean difference between treatment and control. The T^2-statistic for $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ yields a value of

$$\frac{(n_1 + n_2 - p - 1)}{p(n_1 + n_2 - 1)} \frac{n_1 n_2}{n_1 + n_2} \bar{\mathbf{x}}' \mathbf{S}^{-1} \bar{\mathbf{x}} = 5.37$$

which is larger than the critical value $F_{1-2\alpha}\ (p,\ n_1+n_2-p-1) = 2.18$. The corresponding p-value is 0.0024. Regarding the required additional checks, we have $\bar{\mathbf{x}} \neq 0$ and $\sum_{i=1}^{p} \bar{x}_i / \sqrt{g_{ii}} = 0.422 \geq 0$. Follmann's criterion $\bar{\mathbf{x}}' \cdot \mathbf{1}_p \geq 0$ fails here due to the different scales of the endpoints.

Finally, investigation of $C_{1-2\alpha}(\bar{\mathbf{x}}, \mathbf{G})$ reveals that it does not intersect with the negative orthant. Consequently, we can conclude that the new treatment is superior to the standard treatment in at least one of the three endpoints. The minimal value of the quadratic form in the negative orthant is 2.27. It is attained at $\boldsymbol{\mu}_0 = (0,\ -18.4,\ -59.5)'$. The corresponding p-value of the directed test is 0.0447.

As an additional aspect of the application of the methodology presented here, we must of course verify that the treatment does not cause harm in one of the endpoints. Here, this was covered by separate non-inferiority tests on all three endpoints. These are not discussed in this paper. We note, however, that as a consequence, the directional multivariate tests could have been applied with the "weaker" null hypothesis H_0^{corn} using the non-inferiority margins as u_i's.

The SS test is primarily designed to have high power against alternatives where all variables have approximately the same deviation from the null hypothesis in their respective scales. Thus, it is no surprise that it does not work well in this application. The usual SS test (5) yields a t^0 value of 0.639 here. This corresponds to a p-value of 0.2625. The modified SS test procedure IIa is almost identical with a t^{orth} value of 0.638 and a p-value of 0.2629.

7 Discussion

This paper suggests two new multivariate tests for establishing that at least one of several endpoints in a clinical trial shows a beneficial treatment effect. It therefore fills a gap in existing multivariate test approaches, since these only consider a "single point" null hypothesis (like $H_0 : \boldsymbol{\mu} = 0$) which allows no claim about other undesired parameter constellations (like all $\mu_i < 0$).

In comparison with multiple testing approaches (like the Bonferroni method), the new tests have most power if the treatment effect is roughly equally strong in all variables, for example, if all variables are subject to an underlying common treatment effect. If the treatment effect is not "evenly spread" across all variables in this way, but rather there is a single variable with a strong treatment effect, then multiple testing procedures are superior. The SS test in particular has good power if all variables are equally far away from the null hypothesis and have equal pairwise correlations. This is well known and investigated for multivariate methods in general (Srivastava 2005). Since in this respect, the methods suggested here are no different from other multivariate methods, we did not do extensive simulations of these aspects, but rather concentrated on investigating the price to be paid for extending the multivariate test decision to the negative orthant. Especially for the SS test, this price is very small and in the vast majority of concrete examples, there will be no difference between the unmodified and the modified versions of this test.

In clinical trial applications, a significant result of the new tests allows to conclude that in at least one endpoint there is a *beneficial* treatment effect, and not just an effect. The limits of this interpretational extension should be acknowledged. Since a significant result does not rule out the possibility of a harmful treatment effect in some endpoints, the new suggestions are not appropriate for confirmatory clinical trials which require a positive effect on all co-primary endpoints. Nevertheless, we believe that the extended conclusions facilitated by the new suggestions are of real, practically relevant value in earlier phases of clinical development when there still is a number of candidate endpoints.

Acknowledgements The authors thank two anonymous referees for their insightful comments that helped in improving the paper.

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

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd edn). Wiley, New York.

Bathke, A. C. and Harrar, S. W. (2008). Nonparametric methods in multivariate factorial designs for large number of factor levels. *Journal of Statistical Planning and Inference* **138**, 588–610.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics* **25**, 290–302.

Cai, G. and Sarkar, S. K. (2006). Modified Simes' critical values under positive dependence. *Journal of Statistical Planning and Inference* **136**, 4129–4146.

Chuang-Stein, C., Stryszak, P., Dmitrienko, A. and Offen, W. (2007). Challenge of multiple co-primary endpoints: A new approach. *Statistics in Medicine* **26**, 1181–1192.
Dempster, A. P. (1958). A high dimensional two sample significance test. *Annals of Mathematical Statistics* **29**, 995–1010.
Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* **50**, 1096–1121.
Dunnett, C. W. and Tamhane, A. C. (1992). A step-up multiple test procedure. *Journal of the American Statistical Association* **87**, 162–170.
Follmann, D. (1995). Multivariate Tests for multiple endpoints in clinical trials. *Statistics in Medicine* **14**, 1163–1175.
Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association* **91**, 854–861.
Glimm, E., Srivastava, M. S. and Läuter, J. (2002). Multivariate tests of normal mean vectors with restricted alternatives. *Communications in Statistics—Simulation and Computation* **31**, 589–604.
Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* **50**, 403–418.
Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics* **52**, 964–970.
Läuter, J., Glimm, E. and Kropf, S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal* **38**, 5–23.
Läuter, J., Glimm, E. and Kropf, S. (1998). Multivariate tests based on left-spherically distributed linear scores. *Annals of Statistics* **26**, 1972–1988.
Marcus, R., Peritz, E., Gabriel and K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.
Mardia, K. V., Kent, J. T., Bibby and M. J. (1979). *Multivariate Analysis*. Academic Press, London.
Munzel, U. and Brunner, E. (2000). Nonparametric methods in multivariate factorial designs. *Journal of Statistical Planning and Inference* **88**, 117–132.
Nüesch, P. (1966). On the problem of testing location in multivariate problems for restricted alternatives. *Annals of Mathematical Statistics* **37**, 113–119.
O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–1087.
Oja, H. and Randles, R. H. (2004). Multivariate nonparametric tests. *Statistical Science* **19**, 598–605.
Perlman, M. D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics* **40**, 549–567.
Perlman, M. D. and Wu, L. (2004). A note on one-sided tests with multiple endpoints. *Biometrics* **60**, 276–280.
Röhmel, J., Benda, N., Gerlinger, C. and Läuter, J. (2006). On testing simultaneously non-inferiority in two multiple primary endpoints and superiority in at least one of them. *Biometrical Journal* **39**, 1–18.
Schaafsma, W. and Smid, L. J. (1966). Most stringent somewhere most powerful tests against alternatives restricted by a number of inequalities. *Annals of Mathematical Statistics* **37**, 1161–1172.
Silvapulle, M. J. (1997). A curious example involving the likelihood ratio test against one-sided alternatives. *The American Statistician* **51**, 178–181.
Silvapulle, M. J. and Sen, P. K. (2004). *Constrained Statistical Inference*. Wiley, New York.
Srivastava, M. S. (2005). *Methods of Multivariate Statistics*. Wiley, New York.
Srivastava, M. S. and Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *Journal of Multivariate Analysis* **97**, 1927–1940.
Srivastava, M. S. and Du, M. (2008). A test for the mean vector with fewer observations than the dimension. *Journal of Multivariate Analysis* **99**, 386–402.
Tamhane, A. C., Hochberg, Y. and Dunnett, C. W. (1996). Multiple test procedures for dose finding. *Biometrics* **52**, 21–37.
Tang, D. -I., Geller, N. and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics* **49**, 23–30.
Tang, D. -I., Gnecco, C., Geller, N. (1989). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika* **76**, 577–583.

Statistica Neerlandica (2005) Vol. 59, nr. 3, pp. 298–312

Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate

Jürgen Läuter*

Otto von Guericke University Magdeburg and Interdisciplinary Centre for Bioinformatics, University of Leipzig, Mittelstr. 2/151, D-39114 Magdeburg, Germany

Ekkehard Glimm†

AICOS Technologies, Efringerstrasse 32, CH-4057 Basel, Switzerland

Markus Eszlinger‡

III. Medical Department, University of Leipzig, Philipp-Rosenthal-Str. 27, D-04103 Leipzig, Germany

Two multiple procedures for the detection of relevant sets of variables in a high-dimensional problem are suggested. Multivariate tests for significance are combined with the search for interpretable multivariate structures. Thus, groups of highly correlated variables are investigated. The emphasis lies on managing the huge number of possible subsets, for example, in gene expression analysis. The first procedure is based on parametric spherical tests and an order relation of the subsets. The second procedure is a non-parametric method utilizing Westfall–Young principles.

Key Words and Phrases: multiple procedure, multivariate analysis, model choice, spherical test, permutation test, gene expression analysis.

1 Introduction

The problem of a high dimension

Willem Schaafsma was among the first who recognized the difficulties of multivariate analysis connected with a high dimension and with small samples of the data. He compared the statistician's situation with "sailing between Scylla and Charybdis": The utilization of all potential variables – Charybdis – leads to a

*juergen.laeuter@medizin.uni-magdeburg.de
†eglimm@aicos.com
‡markus.eszlinger@medizin.uni-leipzig.de

poorer performance of the procedures than the use of some appropriate subset of variables. On the other hand, selection of a promising subset on the basis of the data often results in an illusion – Scylla – because a good choice of variables is not attained. One has to avoid Charybdis without coming too close to Scylla (SCHAAFSMA and VAN VARK, 1979, p. 117). Schaafsma together with his co-authors has explored many paths to overcome these calamities. For example, he suggested multivariate tests for ordered alternatives (SCHAAFSMA and SMID, 1966), thus narrowing the focus of attention in the huge multivariate space on a relevant subspace. He developed a lot of ingenious mathematical methods to determine a subset of variables, as small as possible, with a power, as high as possible. Thus, the deficiencies of the multivariate procedures due to overfitting phenomena can be contained.

Spherical tests

We have followed yet another strategy in the last ten years (LÄUTER, 1996; LÄUTER, GLIMM and KROPF, 1996, 1998). We intended to extract the full information from the data, without a selection of variables.

For the one-sample testing problem, we will assume n independent p-dimensional normally distributed data vectors

$$\boldsymbol{x}'_{(j)} = (x_{j1} \quad \ldots \quad x_{jp}) \sim \mathrm{N}_p(\boldsymbol{\mu}', \boldsymbol{\Sigma}) \quad (j = 1, \ldots, n; n \geq 2; p \geq 1). \tag{1}$$

The null hypothesis to be tested is $H_0 : \boldsymbol{\mu}' = \boldsymbol{0}$. Then the beta statistic

$$\mathrm{B} = \frac{\boldsymbol{d}'\bar{\boldsymbol{x}}n\bar{\boldsymbol{x}}'\boldsymbol{d}}{\boldsymbol{d}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{d}} \sim \mathrm{B}\left(\frac{1}{2}, \frac{n-1}{2}\right) \quad (\text{under } H_0) \tag{2}$$

is used, where $\boldsymbol{X}$ is the $n \times p$ data matrix of the n rows $\boldsymbol{x}'_{(1)}, \ldots, \boldsymbol{x}'_{(n)}$, $\bar{\boldsymbol{x}}' = \frac{1}{n}\Sigma_{j=1}^{n}\boldsymbol{x}'_{(j)}$ is the mean vector, and $\boldsymbol{d}$ is a p-dimensional weight vector, which can data-dependently be represented as a unique function of the total sums of products matrix $\boldsymbol{X}'\boldsymbol{X}$. The regularity condition $\boldsymbol{d}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{d} > 0$ must hold with probability 1. The covariance matrix $\boldsymbol{\Sigma}$ is an unknown $p \times p$ positive definite matrix.

One of the major advantages of this test is that it can handle very large dimensions p, even if the sample size n is small. In any case, a score vector $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{d}$ is calculated as a combination of the p variables, and this vector can exactly be analyzed by the well-known one-dimensional beta test $\mathrm{B} = \frac{n\bar{z}^2}{\boldsymbol{z}'\boldsymbol{z}} \geq \mathrm{B}_{1-\alpha}(\frac{1}{2}, \frac{n-1}{2})$. The method does not suffer from instability and overfitting. We call it a "spherical test", because the theory of the spherical matrix distributions (FANG and ZHANG, 1990) serves as the basis of the proof.

The above statements concerning the beta test (2) also remain correct if only the conditional spherical distribution of $\boldsymbol{X}$ for a given fixed value $\boldsymbol{X}'\boldsymbol{X}$ is considered. Instead of the B distribution, the F distribution can also be used:

$$F = \frac{\boldsymbol{d}'\bar{\boldsymbol{x}}n\bar{\boldsymbol{x}}'\boldsymbol{d}}{\boldsymbol{d}'(\boldsymbol{X}'\boldsymbol{X} - \bar{\boldsymbol{x}}n\bar{\boldsymbol{x}}')\boldsymbol{d}/(n-1)} \sim F(1, n-1) \quad (\text{under } H_0). \tag{3}$$

A widely applied possibility is to determine the weight vector $\boldsymbol{d}$ as the first eigenvector of the $p \times p$ eigenvalue problem

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{d} = \boldsymbol{d}\lambda, \quad \boldsymbol{d}'\boldsymbol{d} = 1. \tag{4}$$

Then, the score vector $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{d} = \sum_{i=1}^{p} \boldsymbol{x}_i d_i$ provides the values of the first principal component of the data matrix $\boldsymbol{X} = (\boldsymbol{x}_1 \ldots \boldsymbol{x}_p)$, where the column vectors $\boldsymbol{x}_i$ ($i = 1,\ldots, p$) correspond to the p variables. This choice of the weight vector $\boldsymbol{d}$ is most appropriate, if the variables are highly correlated with each other (see, for example, simulation results in KROPF, 2000). If the dimension p is larger than the sample size n, the $n \times n$ dual eigenvalue problem

$$\boldsymbol{X}\boldsymbol{X}'\boldsymbol{u} = \boldsymbol{u}\lambda, \quad \boldsymbol{u}'\boldsymbol{u} = 1 \tag{5}$$

should be used instead of (4). In this case, the beta test (2) is replaced by the equivalent formulas $\bar{u} = \frac{1}{n}\boldsymbol{I}_n'\boldsymbol{u}$, $\mathrm{B} = n\bar{u}^2$, where $\boldsymbol{I}_n$ is the vector consisting of n ones.

The principle of spherical tests is applicable to all linear models. Selection of variables can also be carried out in the framework of this general method, because one may define $\boldsymbol{d}$ as a vector of ones and zeros, of course, as a function of $\boldsymbol{X}'\boldsymbol{X}$.

Multiple test procedures

Multiple procedures searching for the variables i with mean values $\mu_i \neq 0$ can also be developed on this basis. Unfortunately, some other, well-known procedures of multiple testing fail if the dimension p is large and the variables have strong dependencies: The Bonferroni–Holm method (HOLM, 1979) works with extremely small critical values for the smallest P values. A closed test procedure (MARCUS, PERITZ and GABRIEL, 1976) cannot be applied in the usual way because the number of intersection hypotheses is too high. A procedure with a priori ordered hypotheses (BAUER *et al.*, 1998) would demand that a priori information on the ranks of the variables is available.

To overcome these difficulties, KROPF (2000) has proposed a Procedure 1 which is based on the method of spherical tests. The p variables i are sorted according to the sums of squares $\boldsymbol{x}_i'\boldsymbol{x}_i = \sum_{j=1}^{n} x_{ji}^2$ (in the diagonal of $\boldsymbol{X}'\boldsymbol{X}$) by decreasing values. The procedure keeps the familywise type I error rate in the strong sense (FWER), that is, if m_0 is the actual subset of variables i with mean values $\mu_i = 0$, then some relevant variables within m_0, so-called "false relevances", can occur with probability α, at most. Here, the null set m_0 may be any unknown subset of the full set $\{1, \ldots, p\}$ of variables, α is the prespecified multiple significance level. Procedure 1 is particularly effective if the p variables have similar variances σ_{ii} (in the diagonal of $\boldsymbol{\Sigma}$). This can, for example, be fulfilled if the p variables are repeated measures of the same or similar responses in time or space. Then, because $\boldsymbol{x}_i'\boldsymbol{x}_i$ has the expectation

$n(\sigma_{ii} + \mu_i^2)$, the largest sums of squares $\boldsymbol{x}_i'\boldsymbol{x}_i$ $(i = 1, \ldots, p)$ preferably correspond to the variables with the largest values of μ_i^2.

Procedure 1 (KROPF, 2000; KROPF and LÄUTER, 2002):

- Sort the p variables by decreasing values of $\boldsymbol{x}_i'\boldsymbol{x}_i$.
- Carry out the univariate beta tests $B_i = \frac{n\bar{x}_i^2}{\boldsymbol{x}_i'\boldsymbol{x}_i} \geq B_{1-\alpha}\left(\frac{1}{2}, \frac{n-1}{2}\right)$ in the order obtained, as long as significances result. Thus, the relevant variables are stepwisely found. Stop the process at the first non-significant test if present.

In the procedure, the single beta tests are used without any adjustment, at the significance level α. KROPF and LÄUTER (2002) have reported examples from gene expression analysis and computer simulations in which Procedure 1 surpassed the methods by HOLM (1979) and WESTFALL and YOUNG (1993).

HOMMEL and KROPF (2005) suggested a generalization of Kropf's method which allows a continuation of the beta-test steps up to the mth non-significant result. However, then the single tests have to be carried out at the adjusted level α/m. The fixed positive integer m must be given in advance.

Another generalization has been proposed by WESTFALL, KROPF and FINOS (2004). The procedure is based on the Bonferroni–Holm method, but the sums of squares $\boldsymbol{x}_i'\boldsymbol{x}_i$ are used for determining additional weights. Thus, the order of the variables is no longer exclusively based on $\boldsymbol{x}_i'\boldsymbol{x}_i$. The univariate P values of the tests are exploited to modify the order.

In this paper, we are dealing with procedures of multiple testing that refer to sets of variables. The objective is to find sets of variables for which some deviation from the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{0}$, can be proved. Besides, these sets should enable an interpretation of the multivariate structure and the relation between the variables. In our applications, sets of correlated variables are determined. Both parametric tests and permutation tests are considered. The application of the procedures is illustrated by examples from gene expression analysis. An extension of the methods to multisample problems concludes the paper.

2 A multiple procedure for searching sets of variables on the basis of parametric multivariate tests

The general principle

Generalizing Procedure 1, Läuter has proposed a Procedure 2 that searches for relevant subsets within the full set $\{1, \ldots, p\}$ of variables (KROPF and LÄUTER, 2002). Here, we will continue and extend these considerations with some more concrete explanations. We concentrate on methods for managing the multitude of subsets.

Assume a fixed set M of non-empty subsets $m = m_1, m_2, m_3, \ldots$ of ordered variables. Each subset has a unique representation $m = \{i_1, \ldots, i_s\}$ by s different integer indices of the interval from 1 to p $(s \geq 1)$. Two subsets are considered as

different if they consist of different index values or if only the succession of indices is different in both subsets. The subsets are also called "models". The set M of models can be defined, for example, by all non-empty index sequences without repetitions.

We start from a random $n \times p$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1 \quad \ldots \quad \boldsymbol{x}_p)$ that has a multivariate distribution with the expectation $E(\boldsymbol{X}) = \boldsymbol{1}_n\boldsymbol{\mu}' = \boldsymbol{1}_n(\mu_1 \quad \ldots \quad \mu_p)$. Additionally, we assume that the n rows $\boldsymbol{x}'_{(1)}, \ldots, \boldsymbol{x}'_{(n)}$ are stochastically independent and have the normal distribution $\mathrm{N}_p(\boldsymbol{\mu}', \boldsymbol{\Sigma})$ each, with an unknown positive definite covariance matrix $\boldsymbol{\Sigma}$ ($n \geq 2$, $p \geq 1$). For any model $m = \{i_1, \ldots, i_s\} \in \mathrm{M}$ and the corresponding $n \times s$ marginal distribution $\boldsymbol{X}_m = (\boldsymbol{x}_{i_1} \ldots \boldsymbol{x}_{i_s})$, a test statistic $F_m = F_m(\boldsymbol{X}_m)$ with the following property must exist: The conditional marginal test $F_m \geq k_m$, for a given value of $\boldsymbol{X}'_m\boldsymbol{X}_m$, keeps the level of significance α, that is, $\Pr(F_m \geq k_m | \boldsymbol{X}'_m\boldsymbol{X}_m) \leq \alpha$ if $\mu_{i_1} = 0, \ldots, \mu_{i_s} = 0$. The use of the total covariances is necessary, because a general theorem on spherical tests (Läuter, Glimm and Kropf, 1998) is to be applied.

Furthermore, a data-dependent (conditional) order relation of the models has to be established: Given the total sums of products matrix $\boldsymbol{X}'\boldsymbol{X}$, an order $m_{(1)} \succ m_{(2)} \succ m_{(3)} \succ \ldots$ of the models $m \in \mathrm{M}$ must uniquely be defined with probability 1. More precisely, the relation $m_g \succ m_h$ or $m_h \succ m_g$ for any two models m_g und m_h is already uniquely determined if only the submatrices of $\boldsymbol{X}'\boldsymbol{X}$ corresponding to the two models, $\boldsymbol{X}'_{m_g}\boldsymbol{X}_{m_g}$, $\boldsymbol{X}'_{m_h}\boldsymbol{X}_{m_h}$ and $\boldsymbol{X}'_{m_g}\boldsymbol{X}_{m_h}$, are given.

The following Procedure 2 serves the purpose of finding models $m = \{i_1, \ldots, i_s\}$ consisting not only of null variables, that is, $m \not\subseteq m_0$ with m_0 being the subset with mean values $\mu_i = 0$. However, the relation $\mu_i \neq 0$ need not be fulfilled for every index $i = i_1, \ldots, i_s$ of such a model. The familywise error rate (FWER) is kept again in this procedure: In the series of all obtained "relevant models", some "false-relevant models", that is, models consisting only of variables with $\mu_i = 0$, may appear with the probability α, at most.

Procedure 2:

- Sort the models $m \in \mathrm{M}$ according to $\boldsymbol{X}'\boldsymbol{X}$: $m_{(1)} \succ m_{(2)} \succ m_{(3)} \succ \cdots$
- Carry out the tests $F_m \geq k_m$ in the order obtained, as long as significances result. Thus, the relevant models m are stepwisely found. Stop the process at the first non-significant test if present.

The proof of the FWER property of Procedure 2 runs on the same line as the corresponding proof presented by Kropf and Läuter (2002). The main idea is that a false relevance can only arise through a significance $F_m \geq k_m$ with a model $m \subseteq m_0$, uniquely determined by the submatrix $\boldsymbol{X}'_{m_0}\boldsymbol{X}_{m_0}$ of the null variables. Then, since $\boldsymbol{X}$ is normally distributed, the matrix $\boldsymbol{X}_m$, given $\boldsymbol{X}'_{m_0}\boldsymbol{X}_{m_0}$, has a conditional spherical distribution which coincides with the conditional distribution of $\boldsymbol{X}_m$, given $\boldsymbol{X}'_m\boldsymbol{X}_m$ (see Fang and Zhang, 1990, Theorem 3.1.1). Hence, F_m provides an exact conditional test, that is, a false relevance is possible with probability α, at most.

A special implementation of Procedure 2

Let M denote the full set of all possible models m represented by non-empty index sequences $\{i_1, \ldots, i_s\}$ without repetitions. To sort these models, we first define so-called essential and unessential models based on the matrix $\boldsymbol{X}'\boldsymbol{X}$. Subsequently, only the essential models will play a role. If c is a fixed given value with $0 \leq c \leq 1$, then a model $m = \{i_1, \ldots, i_s\}$ is considered "essential" if

- $\boldsymbol{x}'_{i_2}\boldsymbol{x}_{i_2} \geq \boldsymbol{x}'_{i_3}\boldsymbol{x}_{i_3} \geq \cdots \geq \boldsymbol{x}'_{i_s}\boldsymbol{x}_{i_s}$ is valid for $s \geq 3$,
- the conditions

$$r^2_{i_1 i} = \frac{(\boldsymbol{x}'_{i_1}\boldsymbol{x}_i)^2}{\boldsymbol{x}'_{i_1}\boldsymbol{x}_{i_1} \cdot \boldsymbol{x}'_i\boldsymbol{x}_i} \geq c \quad (i = i_2, \ldots, i_s) \tag{6}$$

  on the squared correlation coefficients are satisfied for $s \geq 2$.

Here, the variable i_1 is called "pivot variable". All models m that are not essential in the sense of this definition are denoted as "unessential". In the special case of $c = 1$, only the models of size $s = 1$, $\{1\}$ to $\{p\}$, are essential models (with probability 1), as in Procedure 1.

Now, the order relation of the models corresponding to the general principle is defined in three steps:

- At first, all essential models get precedence over all unessential models.
- Then, the essential models are sorted by decreasing values of $\boldsymbol{x}'_{i_1}\boldsymbol{x}_{i_1} \cdot s$.
- Finally, in the groups of essential models with the same pivot variable i_1 and the same size s, a sorting according to the lexical principle based on $\boldsymbol{x}'_i\boldsymbol{x}_i$ is carried out. The variables with large values $\boldsymbol{x}'_i\boldsymbol{x}_i$ precede the variables with small values (see the example below).

In the essential models, variables with large sums of squares are preferred over variables with small sums of squares. This enables us to suppress weak variables within the random noise of the measuring process and to find variables with strong mean-value deviations from zero. Many researchers have observed that such strategies work well in practice (see, for example, HASTIE, TIBSHIRANI and FRIEDMAN, 2001, Section 3.4). The sorting of the unessential models is not specified in detail, because these models are not included in the testing steps of the procedure. As test statistic $F_m = F_m(\boldsymbol{X}_m)$ for an essential model m, we will use the spherical beta test (2) applied to $\boldsymbol{X}_m$. The s variables of model m are compressed to a one-dimensional score. To obtain the weight vector $\boldsymbol{d}$, the eigenvalue problem (4) is solved for $\boldsymbol{X}'_m\boldsymbol{X}_m$.

In this implementation, models with high total correlations are searched. Note that high total correlations imply similar mean values of the variables. Models with many variables (s large) get precedence over models with few variables (s small). This is motivated by the applications to gene expression analysis we have in mind. The interest is focused on groups of similarly behaving co-regulated genes rather than on

isolated genes. Submodels of more comprehensive models are analyzed only if the latter have already been identified as relevant.

An example from gene expression analysis

The multiple procedures are demonstrated by gene chip data recorded on 14 patients with thyroid diseases (M. Eszlinger). The expression patterns of 12 625 genes (Affymetrix GeneChip U95Av2) have been measured in tissue samples of cold thyroid nodules and the normal surroundings. The differences between the logarithmic expression values of nodular and surrounding tissue are analyzed. In the following, we consider only a subset of 148 genes that belong to different signaling cascades that had been identified by independent biological research. It is likely that these signaling cascades are involved in the development of cold thyroid nodules. We intend to find genes that are significantly differentially expressed between the nodular tissue and the normal surrounding tissue. Furthermore, we want to discover statistical relationships between the genes. Such multivariate considerations can give insight into biological pathways and can provide interesting sets of co-regulated genes.

The matrix $\boldsymbol{X}$ has $n = 14$ rows and $p = 148$ columns. If $\alpha = 0.05$, the univariate tests at the level α/p according to the Bonferroni method yield two relevant genes: $i = 36$ (1675_at) and $i = 42$ (1731_at). Unfortunately, Procedure 1 of this paper does not yield a relevant gene. In the sequence of genes corresponding to the sums of squares $\boldsymbol{x}_i'\boldsymbol{x}_i$, gene $i = 64$ (2070_i_at) is first. However, this gene is not significant at level α.

Applying the special implementation of Procedure 2 described above, several relevant gene sets are found. Thus, the difference between the thyroid nodules and the surrounding is proved, and at the same time structural knowledge on the genes is acquired. For $\alpha = 0.05$ and the correlation bound $c = 0.50$, that is, $\text{abs}(r_{i_1 i}) \geq \sqrt{0.50} = 0.7071$, the following relevant gene sets are obtained:

$$
\begin{array}{llll}
m = \{42, 30, 52, 40\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 109.63, & \text{B} = 0.6810, & P \text{ value} = 0.0002, \\
m = \{30, 42, 82\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 94.54, & \text{B} = 0.6646, & P \text{ value} = 0.0002, \\
m = \{95, 138, 100\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 90.51, & \text{B} = 0.3074, & P \text{ value} = 0.0320, \\
m = \{42, 30, 52\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 82.23, & \text{B} = 0.6762, & P \text{ value} = 0.0002, \\
m = \{42, 30, 40\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 82.23, & \text{B} = 0.6732, & P \text{ value} = 0.0002, \\
m = \{42, 52, 40\}, & \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 82.23, & \text{B} = 0.6251, & P \text{ value} = 0.0005.
\end{array}
$$

The procedure ends with

$$m = \{64\}, \quad \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot s = 65.40, \quad \text{B} = 0.0000, \quad P \text{ value} = 0.9896.$$

It is seen that the series of relevant sets begins with the "large" set

$$\{42(1731\_\text{at}), 30(1591\_\text{at}), 52(1879\_\text{at}), 40(1709\_\text{g\_at})\}.$$

Its three "smaller" subsets

$$\{42, 30, 52\}, \{42, 30, 40\}, \{42, 52, 40\}$$

with lexical order of the variables i_2 and i_3 (corresponding to the inequality $\boldsymbol{x}_{30}'\boldsymbol{x}_{30} > \boldsymbol{x}_{52}'\boldsymbol{x}_{52} > \boldsymbol{x}_{40}'\boldsymbol{x}_{40}$) are at the end.

Some relations are interesting: In this example, there are 6.9×10^{258} non-empty subsets of genes without repetitions. From these sets, 512 are essential subsets (with respect to the correlations) and, as we see, six are relevant subsets (with respect to the significance level $\alpha = 0.05$).

Some further proposals of implementation

There are many possibilities to restrict the diversity of models within the general framework provided by Procedure 2. It is important that any additional conditions must be based on $\boldsymbol{X}'\boldsymbol{X}$ or the corresponding submatrices, respectively.

If the total number of variables provided for the models is to be reduced, then one can set the additional condition $\boldsymbol{x}_i'\boldsymbol{x}_i \geq a$, where a has a fixed given value. The above correlation condition $r_{i_1 i}^2 \geq c$ can be strengthend by the positivity condition $r_{i_1 i} > 0$ $(i = i_2, \ldots, i_s)$. Besides, a relative bound for the sums of squares of the variables $i = i_2, \ldots, i_s$ can be specified: $\boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1} \cdot b \leq \boldsymbol{x}_i'\boldsymbol{x}_i \leq \boldsymbol{x}_{i_1}'\boldsymbol{x}_{i_1}$, where b has a fixed given value with $0 \leq b < 1$.

The order relation of the essential models m may also be defined, for example, by sorting for decreasing first eigenvalues λ of the matrices $\boldsymbol{X}_m'\boldsymbol{X}_m$. Instead of the eigenvalue problem (4), the eigenvalue problem $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{d} = \mathrm{Diag}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{d}\lambda$, $\boldsymbol{d}'\mathrm{Diag}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{d} = 1$, which yields a scale-invariant score $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{d}$, can be used for the determination of $\boldsymbol{d}$ and λ.

3 A multiple procedure for searching sets of variables on the basis of the Westfall–Young permutation strategy

The general principle

As in Section 2, we start from a fixed set M of non-empty models $m = \{i_1, \ldots, i_s\}$. The random $n \times p$ data matrix $\boldsymbol{X}$ consists of n independent rows $\boldsymbol{x}_{(j)}'$ that are symmetrically distributed with respect to $\boldsymbol{\mu}'$:

$$-(\boldsymbol{x}_{(j)}' - \boldsymbol{\mu}') \stackrel{d}{=} \boldsymbol{x}_{(j)}' - \boldsymbol{\mu}' \quad (j = 1, \ldots, n;\ n \geq 1;\ p \geq 1). \tag{7}$$

This assumption is fulfilled, for example, if $\boldsymbol{x}_{(j)}' \sim \mathrm{N}_p(\boldsymbol{\mu}', \boldsymbol{\Sigma})$. For each model $m = \{i_1, \ldots, i_s\}$ with the submatrix $\boldsymbol{X}_m = (\boldsymbol{x}_{i_1} \ \ldots \ \boldsymbol{x}_{i_s})$ of $\boldsymbol{X}$, a test statistic $F_m = F_m(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_s})$ must be given. Without loss of generality, it is assumed that large values of F_m correspond to the deviation from the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{0}$.

The one-sample permutation test is constructed by means of sign commutations of the rows of $\boldsymbol{X}$. A convenient way of denoting these sign commutations uses the 2^n diagonal matrices $\boldsymbol{E}^*$ with $+1$ or -1 in the diagonal. Then, we obtain the transformed matrices $\boldsymbol{X}^* = \boldsymbol{E}^*\boldsymbol{X}$. Under the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{0}$, all these matrices $\boldsymbol{X}^*$ are equally likely values of the random variable $\boldsymbol{X}$. Correspondingly, the test-statistic values $F_m^* = F_m(\boldsymbol{x}_{i_1}^*, \ldots, \boldsymbol{x}_{i_s}^*)$ of a model m are equally likely and, consequently, so are the maximum values from all models m, $F^* = \max_{m \in \mathrm{M}} F_m^*$.

Therefore, the maximum value $F = \max_{m \in \mathrm{M}} F_m$ derived from the original data will lie in the upper α tail of the empirical distribution of all maximum values F^* only with probability α if the null hypothesis is true:

$$\Pr(\#(F \leq F^*) \leq 2^n \alpha) \leq \alpha. \tag{8}$$

Thus, to perform the exact test for the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{0}$ at level α, we have to check the condition

$$\#(F \leq F^*) \leq 2^n \alpha. \tag{9}$$

If the inequality $F = \max_{m \in \mathrm{M}} F_m \leq \max_{m \in \mathrm{M}} F_m^* = F^*$ holds true for $2^n\alpha$ commutations at most, the null hypothesis is rejected. This is the permutation strategy by WESTFALL and YOUNG (1993) based on the maximum of all test-statistic values. In the literature, it is referred to as the "max$T$ method" (YANG and SPEED, 2003).

In this way, a multiple procedure can also be obtained. The condition for rejecting the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{0}$ by model m is

$$\#(F_m \leq F^*) \leq 2^n \alpha. \tag{10}$$

Then, the rejection of some model m will occur with probability α, at most, under $\boldsymbol{\mu}' = \boldsymbol{0}$.

Furthermore, the FWER is still kept if any nonempty set m_0 of null variables is supposed. In all cases, the relevance condition (10) is applied to a model m. It is clear from the above consideration that, at most with probability α, some model $m \subseteq m_0$ will satisfy the condition $\#(F_m \leq F_0^*) \leq 2^n\alpha$, where $F_0^* = \max_{m \subseteq m_0} F_m^*$ is the maximum value of all models m contained in the null set m_0. Then, because of $F_0^* \leq F^*$, condition (10) is fulfilled with an even lesser probability. The empirical distribution of the occurring values F^* dominates the real distribution of the values F_0^* given in m_0.

In the following multiple procedure, the 2^n fixed ± 1 diagonal matrices $\boldsymbol{E}^*$ may also be replaced by r random ± 1 diagonal matrices. If a number $r \geq 2$ is prespecified, we may set $\boldsymbol{E}_1^* = \boldsymbol{I}_n$ (identity matrix) and obtain $\boldsymbol{E}_2^*, \ldots, \boldsymbol{E}_r^*$ by generating independent random diagonal matrices with uniform distribution on all 2^n different diagonal patterns, that is, "random sampling with replacement" is applied. Such random sign commutations are important if a large sample size n makes it impossible to go through all 2^n commutations.

Procedure 3:
r fixed or random sign commutation matrices $\boldsymbol{E}^*$, as defined in the previous paragraphs, are supposed. For each of the r transformed matrices $\boldsymbol{X}^* = \boldsymbol{E}^*\boldsymbol{X}$ (including the given matrix $\boldsymbol{X}$), the maximum test-statistic value $F^* = \max_{m \in \mathrm{M}} F_m^* = \max_{m \in \mathrm{M}} F_m(\boldsymbol{x}_{i_1}^*, \ldots, \boldsymbol{x}_{i_s}^*)$ of all models $m \in M$ is calculated. Then a model m is relevant if the number of commutations fulfilling $F_m = F_m(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_s}) \leq F^*$ is less than or equal to $r\alpha$. The corresponding "adjusted" P value is $P = \frac{\#(F_m \leq F^*)}{r}$.

An equivalent relevance condition is based on the increasing sequence of the values F^* arising from the commutations: $F^{(1)}, F^{(2)}, \ldots, F^{(r)}$. We determine the $(1 - \alpha)$ quantile $F^*_{1-\alpha} = F^{(k)}$, where k is the smallest integer greater than or equal to $r(1 - \alpha)$. Then, the relevance condition for model m can be written as $F_m > F^*_{1-\alpha}$.

A special implementation of Procedure 3

In analogy to the implementation of Procedure 2, we will start from the set M of all non-empty models $m = \{i_1, \ldots, i_s\}$ with different ordered indices $i_1, \ldots, i_s$. The distinction between essential and unessential models from the former implementation is also maintained. Provided that a fixed value c with $0 \leq c \leq 1$ is given, the class of models with correlation coefficients $r^2_{i_1 i} \geq c$ $(i = i_2, \ldots, i_s)$ is again considered. We define the test statistic F_m for model m by

$$F_m(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_s}) = \begin{cases} \mathrm{B}_{i_1} + \cdots + \mathrm{B}_{i_s} & \text{if } m \text{ is an essential model,} \\ 0 & \text{if } m \text{ is an unessential model.} \end{cases} \tag{11}$$

Here, B_i is the univariate beta statistic $\mathrm{B}_i = \frac{n\bar{x}_i^2}{\boldsymbol{x}_i'\boldsymbol{x}_i}$ $(i = 1, \ldots, p)$.

It is an advantage of this implementation that the essentiality conditions are not affected by the commutations $\boldsymbol{X}^* = \boldsymbol{E}^*\boldsymbol{X}$, that is, $F^*_m = F_m(\boldsymbol{x}^*_{i_1}, \ldots, \boldsymbol{x}^*_{i_1}) = 0$ if and only if $F_m = F_m(\boldsymbol{x}_{i_1}, \ldots, \boldsymbol{x}_{i_1}) = 0$. Another advantage lies in the existence of a "supermodel" for each pivot variable i_1, which surpasses all essential models for i_1. The supermodel M_{i_1} consists of variable i_1 and additionally all variables i with $r^2_{i_1 i} \geq c$. The order relationship $\boldsymbol{x}'_{i_2}\boldsymbol{x}_{i_2} \geq \boldsymbol{x}'_{i_3}\boldsymbol{x}_{i_3} \geq \cdots \geq \boldsymbol{x}'_{i_s}\boldsymbol{x}_{i_s}$ must hold. Then, the inequalities $F_m \leq F_{M_{i_1}}$ and $F^*_m \leq F^*_{M_{i_1}}$ are valid for each essential model $m = \{i_1, \ldots, i_s\}$, and $F^* = \max_{m \in \mathrm{M}} F^*_m = \max_{i_1 = 1, \ldots, p} F^*_{M_{i_1}}$ holds. Thus, the computational expense of Procedure 3 is much reduced; only the p supermodels are needed to determine the distribution of F^*.

Continuation of the example from the gene expression analysis

This implementation of Procedure 3 is applied to the data by M. Eszlinger. We set $\alpha = 0.05$, $c = 0.50$ and use $r = 8192$ random commutations. Then the maximum test-statistic values F^* lie in the interval from 0.29 to 3.64. The critical value is $F^*_{1-\alpha} = 1.71$. The following three supermodels turn out to be relevant:

$$\begin{aligned} M_{36} &= \{36, 27, 40, 63\}, \\ F_{M_{36}} &= \mathrm{B}_{36} + \mathrm{B}_{27} + \mathrm{B}_{40} + \mathrm{B}_{63} = 0.70 + 0.41 + 0.54 + 0.26 = 1.91, \\ M_{40} &= \{40, 42, 36\}, \\ F_{M_{40}} &= \mathrm{B}_{40} + \mathrm{B}_{42} + \mathrm{B}_{36} = 0.54 + 0.67 + 0.70 = 1.91, \\ M_{42} &= \{42, 30, 52, 40\}, \\ F_{M_{42}} &= \mathrm{B}_{42} + \mathrm{B}_{30} + \mathrm{B}_{52} + \mathrm{B}_{40} = 0.67 + 0.50 + 0.32 + 0.54 = 2.03. \end{aligned}$$

In this case, a reduction of a supermodel to a smaller essential model is not possible without losing the property of relevance. In every case, the critical limit 1.71 must be

attained if relevance is desired. The model M_{42} was already recognized as a relevant one in the parametric procedure of Section 2.

The results change if the condition of positive correlation between the variables $r_{i_1 i} > 0$ is added. This means that we will search sets of such genes that are all activated or all deactivated in the thyroid nodules against the surrounding. Then the values F^* run from 0.26 to 2.99. The critical value is $F^*_{1-\alpha} = 1.50$. We find the following three relevant supermodels:

$$
\begin{aligned}
&M_{36} = \{36, 27, 40, 63\},\\
&F_{M_{36}} = \mathrm{B}_{36} + \mathrm{B}_{27} + \mathrm{B}_{40} + \mathrm{B}_{63} = 0.70 + 0.41 + 0.54 + 0.26 = 1.91,\\
&M_{87} = \{87, 94, 39, 105\},\\
&F_{M_{87}} = \mathrm{B}_{87} + \mathrm{B}_{94} + \mathrm{B}_{39} + \mathrm{B}_{105} = 0.37 + 0.32 + 0.53 + 0.44 = 1.66,\\
&M_{94} = \{94, 39, 87, 105\},\\
&F_{M_{94}} = \mathrm{B}_{94} + \mathrm{B}_{39} + \mathrm{B}_{87} + \mathrm{B}_{105} = 0.32 + 0.53 + 0.37 + 0.44 = 1.66.
\end{aligned}
$$

Here, we are able to a find a relevant submodel of model M_{36}:

$$m = \{36, 27, 40\}, \quad F_m = \mathrm{B}_{36} + \mathrm{B}_{27} + \mathrm{B}_{40} = 0.70 + 0.41 + 0.54 = 1.65.$$

This reduced model is still relevant because $F_m > 1.50$. It can be seen that the procedure provides two relevant models that consist of the same variables: M_{87} and M_{94}. However, the relevant sets M_{36} and M_{87} (or M_{94}) have no common variables. These sets correspond to different "factors".

Further proposals

Procedure 3 can be applied in many different ways. All modifications mentioned in Section 2 are again possible. In addition, many further variations are available.

If rank numbers are substituted for the observed values, the influence of special distributions in the data can be decreased. For each variable separately, the absolute values of the given data are replaced by the corresponding rank numbers, but the given signs of the data must remain unchanged.

In the framework of the special implementation of Procedure 3, the test statistic F_m can be defined by means of an arbitrary univariate function $f(\boldsymbol{x})$ with non-negative values:

$$F_m(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s) = \begin{cases} f(\boldsymbol{x}_1) + \ldots + f(\boldsymbol{x}_s) & \text{if } m \text{ is an essential model,} \\ 0 & \text{if } m \text{ is an unessential model.} \end{cases} \tag{12}$$

For example, the truncated univariate beta statistic

$$f(\boldsymbol{x}) = \begin{cases} \mathrm{B} & \text{if } B \geq \mathrm{B}_0, \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

or the corresponding 0–1 statistic

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } B \geq \mathrm{B}_0, \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

can be applied, where $B = n\bar{x}^2/\boldsymbol{x}'\boldsymbol{x}$ and B_0 is a fixed truncation point (for example, $B_0 = B_{1-\alpha}(1/2, (n-1)/2)$, the quantile of the univariate test). A scale-dependent function f is also possible, for example $f(\boldsymbol{x}) = \bar{x}^2$.

More general, the essentiality conditions (see (6)) may also be affected by the commutations, for example, if the within-sample sums of products matrix $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})$ is used for the determination of the correlation coefficients $r_{i_1 i}$ instead of the total sums of products matrix $\boldsymbol{X}'\boldsymbol{X}$. The additional restriction $\bar{x}_i > 0$ $(i = i_1, \ldots, i_s)$ or only $\bar{x}_{i_1}\bar{x}_i > 0$ $(i = i_2, \ldots, i_s)$ could also be included. Of course, these extensions increase the computational burden.

In all cases, Procedure 3 keeps the familywise error rate α in the strong sense. Any restricting conditions, for example "subset pivotality", do not play a role (WESTFALL and YOUNG, 1993; DUDOIT, SHAFFER and BOLDRICK, 2002; DUDOIT, VAN DER LAAN and POLLARD, 2003).

4 Application of more general rotation tests instead of permutation tests

In the case that the sample size n is small, only few different commutations represented by corresponding ± 1 diagonal matrices $\boldsymbol{E}^*$ are available. This impairs the applicability of the sign-commutation strategy in the procedure by Westfall and Young. However, the more comprehensive class of $n \times n$ orthogonal matrices $\boldsymbol{E}^*$ may then be used. LANGSRUD (2004) has also treated such "rotation tests".

Assume an $n \times p$ data matrix $\boldsymbol{X}$ that is left-spherically distributed with regard to the mean-value matrix $\mathbf{M} = \boldsymbol{1}_n\boldsymbol{\mu}'$, that is,

$$\boldsymbol{C}'(\boldsymbol{X} - \boldsymbol{1}_n\boldsymbol{\mu}') \stackrel{d}{=} \boldsymbol{X} - \boldsymbol{1}_n\boldsymbol{\mu}' \tag{15}$$

for each fixed $n \times n$ orthogonal matrix $\boldsymbol{C}$ (FANG and ZHANG, 1990). This is fulfilled, in particular, if $\boldsymbol{X}$ consists of n independent rows $\boldsymbol{x}'_{(j)} \sim N_p(\boldsymbol{\mu}', \boldsymbol{\Sigma})$.

In the following, Procedure 3 is again taken as the basis, but arbitrary orthogonal transformations $\boldsymbol{X}^* = \boldsymbol{E}^{*\prime}\boldsymbol{X}$ are applied. We set $\boldsymbol{E}_1^* = \boldsymbol{I}_n$ and generate $\boldsymbol{E}_2^*, \ldots, \boldsymbol{E}_r^*$ as independent random matrices each having the $n \times n$ spherical standard distribution. An $n \times q$ matrix $\boldsymbol{U}$ has the spherical standard distribution if it is left-spherically distributed according to (15) with expectation $\boldsymbol{\mu}' = \boldsymbol{0}$ and consists of orthogonal column vectors, that is, $\boldsymbol{C}'\boldsymbol{U} \stackrel{d}{=} \boldsymbol{U}$ for each orthogonal $\boldsymbol{C}$ and $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_q$ must hold.

Under the distributional assumption (15) with $\boldsymbol{\mu}' = \boldsymbol{0}$, for fixed values of the matrices $\boldsymbol{E}^* = \boldsymbol{E}_1^*, \ldots, \boldsymbol{E}_r^*$ and for a given $\boldsymbol{X}$, all the "rotated" matrices $\boldsymbol{X}^* = \boldsymbol{E}^{*\prime}\boldsymbol{X}$ are $\boldsymbol{X}$ values with the same probability. If the null hypothesis is fulfilled only on the subset m_0, then this property is valid only for the columns of $\boldsymbol{X}^*$ corresponding to m_0. This is the mathematical background of this method. All considerations of Section 3 can be transferred to general orthogonal rotations (with the exception that the rank transformation of the data is not allowed), because $\boldsymbol{X}^{*\prime}\boldsymbol{X}^* = \boldsymbol{X}'\boldsymbol{X}$. In the case of the

special implementation of Procedure 3, the "essential models" need not be determined anew for each rotation. It is sufficient for the computer program to generate $n \times 1$ spherically distributed vectors $\boldsymbol{e}^*$ with $\boldsymbol{e}^{*\prime}\boldsymbol{e}^* = 1$ and then to calculate the beta statistics $\mathrm{B}_i^* = (\boldsymbol{e}^{*\prime}\boldsymbol{x}_i)^2/\boldsymbol{x}_i'\boldsymbol{x}_i$ $(i = 1, \ldots, p)$.

5 Comparison of the mean vectors of two groups

The methods treated in Sections 2 to 4 can easily be applied to the comparison of the means $\boldsymbol{\mu}^{(1)\prime}$ and $\boldsymbol{\mu}^{(2)\prime}$ from two samples. We consider the two samples of p-dimensional observations

$$\boldsymbol{x}_{(j)}^{(1)'} = (x_{j1}^{(1)} \ldots x_{jp}^{(1)}) \quad (j = 1, \ldots, n^{(1)}), \tag{16}$$

$$\boldsymbol{x}_{(j)}^{(2)'} = (x_{j1}^{(2)} \ldots x_{jp}^{(2)}) \quad (j = 1, \ldots, n^{(2)}). \tag{17}$$

All vectors $\boldsymbol{x}_{(j)}^{(1)\prime} - \boldsymbol{\mu}^{(1)\prime}(j = 1, \ldots, n^{(1)})$ and $\boldsymbol{x}_{(j)}^{(2)\prime} - \boldsymbol{\mu}^{(2)\prime}(j = 1, \ldots, n^{(2)})$ are supposed to be independently and identically distributed.

We want to find variables and sets of variables i for which the elements $\mu_i^{(1)}$ and $\mu_i^{(2)}$ of $\boldsymbol{\mu}^{(1)\prime}$ and $\boldsymbol{\mu}^{(2)\prime}$ are different. Let $\boldsymbol{X} = (\boldsymbol{x}_1 \ldots \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{pmatrix}$ denote the $n \times p$ data matrix consisting of $n^{(1)}$ vectors of group 1 and $n^{(2)}$ vectors of group 2 $(n = n^{(1)} + n^{(2)})$.

Instead of the matrix $\boldsymbol{X}'\boldsymbol{X}$ in the former presentation, $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})$ is now used as the "total sums of products matrix", where $\bar{\boldsymbol{X}} = \boldsymbol{1}_n\bar{\boldsymbol{x}}' = \boldsymbol{1}_n\frac{1}{n}(n^{(1)}\bar{\boldsymbol{x}}^{(1)\prime} + n^{(2)}\bar{\boldsymbol{x}}^{(2)\prime})$ is the matrix of the total mean values. For testing the null hypothesis $H_0 : \boldsymbol{\mu}^{(1)\prime} = \boldsymbol{\mu}^{(2)\prime}$ in the case of the normal distribution of $\boldsymbol{X}$, the beta statistic

$$\mathrm{B} = \frac{\boldsymbol{d}'(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})a(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})'\boldsymbol{d}}{\boldsymbol{d}'(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{d}} \sim \mathrm{B}\left(\frac{1}{2}, \frac{n-2}{2}\right) \quad (\text{under } H_0) \tag{18}$$

is available, with a vector $\boldsymbol{d}$ being a function of $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})$ and $a = n^{(1)}n^{(2)}/n$. The corresponding univariate version of this test is $\mathrm{B}_i = \frac{a(\bar{x}_i^{(1)} - \bar{x}_i^{(2)})^2}{(\boldsymbol{x}_i - \boldsymbol{1}_n\bar{x}_i)'(\boldsymbol{x}_i - \boldsymbol{1}_n\bar{x}_i)} = \frac{a(\bar{x}_i^{(1)} - \bar{x}_i^{(2)})^2}{\boldsymbol{x}_i'\boldsymbol{x}_i - n\bar{x}_i^2}$. It is necessary for Procedure 1 and 3. The eigenvalue problems (4) and (5) are replaced by the corresponding eigenvalue problems

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})\boldsymbol{d} = \boldsymbol{d}\lambda, \quad \boldsymbol{d}'\boldsymbol{d} = 1, \tag{19}$$

$$(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})'\begin{pmatrix} \boldsymbol{v}^{(1)} \\ \boldsymbol{v}^{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{v}^{(1)} \\ \boldsymbol{v}^{(2)} \end{pmatrix}\lambda, \quad \begin{pmatrix} \boldsymbol{v}^{(1)\prime} & \boldsymbol{v}^{(2)\prime} \end{pmatrix}\begin{pmatrix} \boldsymbol{v}^{(1)} \\ \boldsymbol{v}^{(2)} \end{pmatrix} = 1. \tag{20}$$

Then test (18) becomes

$$\mathrm{B} = a(\bar{v}^{(1)} - \bar{v}^{(2)})^2 \quad \text{with} \quad \bar{v}^{(1)} = \frac{1}{n^{(1)}}\boldsymbol{1}'_{n^{(1)}}\boldsymbol{v}^{(1)}, \bar{v}^{(2)} = \frac{1}{n^{(2)}}\boldsymbol{1}'_{n^{(2)}}\boldsymbol{v}^{(2)}. \tag{21}$$

The "essential models" and their order relation are defined by means of $(\boldsymbol{X} - \bar{\boldsymbol{X}})'(\boldsymbol{X} - \bar{\boldsymbol{X}})$. Instead of the sign commutations in Section 3, permutations of the rows of matrix $\boldsymbol{X}$ are applied.

The method of random orthogonal transformations from Section 4 can also be transferred to the case of two samples. If we suppose $\boldsymbol{X} \sim \mathrm{N}_{n\times p}(\boldsymbol{1}_n\boldsymbol{\mu}', \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$ under the null hypothesis $\boldsymbol{\mu}^{(1)\prime} = \boldsymbol{\mu}^{(2)\prime} = \boldsymbol{\mu}'$, then the matrix $\boldsymbol{X}$ is transformed to mean $\boldsymbol{0}$ by a fixed $n \times (n-1)$ matrix $\boldsymbol{A}$ with $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_{n-1}$ and $\boldsymbol{A}'\boldsymbol{1}_n = \boldsymbol{0}$ (for example, by a Helmert matrix): $\boldsymbol{Y} = \boldsymbol{A}'\boldsymbol{X} \sim \mathrm{N}_{(n-1)\times p}(\boldsymbol{0}, \boldsymbol{I}_{n-1} \otimes \boldsymbol{\Sigma})$. From this, if independent random $(n-1) \times (n-1)$ spherically standard distributed matrices $\boldsymbol{E}^*$ are used, the rotated matrices $\boldsymbol{Y}^* = \boldsymbol{E}^{*\prime}\boldsymbol{Y}$ arise, which can be utilized in Procedure 3. In the computer program, $(n-1) \times 1$ spherically distributed vectors $\boldsymbol{e}^*$ with $\boldsymbol{e}^{*\prime}\boldsymbol{e}^* = 1$ are sufficient to simulate the random variation of the beta statistics, that is, $\mathrm{B}_i^* = \frac{(\boldsymbol{e}^{*\prime}\boldsymbol{y}_i)^2}{\boldsymbol{y}_i'\boldsymbol{y}_i} = \frac{(\boldsymbol{e}^{*\prime}\boldsymbol{y}_i)^2}{(\boldsymbol{x}_i - \boldsymbol{1}_n\bar{x}_i)'(\boldsymbol{x}_i - \boldsymbol{1}_n\bar{x}_i)} (i = 1, \ldots, p)$.

6 Conclusion

In this paper, we are suggesting a new solution to an old problem in multivariate statistics: how to find the essential among many variables. The methods we are using consist of two steps. First, sets of variables, called "models", are defined by lumping together variables that behave similarly, for example, based on their mutual correlation. Then, principles from multiple testing theory are applied to detect the relevant sets while keeping the familywise type I error rate. Thus, the probability of false identification of relevant genes is strictly controlled. The strategies proposed can be applied even in situations with a very high dimension because they can be implemented in ways that entail a relatively low computational burden. The authors have written computer programs which can be used to analyze several thousands of variables in gene expression analysis. Of course, a critical assessment of the results by a scientist from the application field is necessary.

Each set of variables obtained by the algorithms of this paper has a "one-factorial" structure, because it is defined by conditions on the correlatedness to a pivot variable. Then, the different relevant sets (or the scores derived from them) can provide a basis for further multivariate investigations, for example, in discriminant or regression analysis.

Acknowledgements

The authors are grateful to the referee for his useful comments that helped to improve this paper. Thanks are also due to Professor Ton Steerneman and Professor Herold Dehling for organizing a special issue of *Statistica Neerlandica* in honour of Willem Schaafsma.

References

Bauer, P., J. Röhmel, W. Maurer and L. A. Hothorn (1998), Testing strategies in multiple-dose experiments including active control, *Statistics in Medicine* **17**, 2133–2146.

Dudoit, S., J. P. Shaffer and J. C. Boldrick (2002), *Multiple hypothesis testing in microarray experiments*, The Berkeley Electronic Press, Year 2002, Paper 110.

Dudoit, S., M. J. van der Laan and K. S. Pollard (2003), *Multiple testing: Part I, single-step procedures for control of general type I error rates*, The Berkeley Electronic Press, Year 2003, Paper 138.

Fang, K.-T. and Y.-T. Zhang (1990), *Generalized multivariate analysis*, Springer-Verlag, Berlin.

Hastie, T., R. Tibshirani and J. Friedman (2001), *The elements of statistical learning*, Springer-Verlag, New York.

Holm, S. (1979), A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. and S. Kropf (2005), Tests for differentiation in gene expression using a data-driven order or weights for hypotheses, *Biometrical Journal*, accepted for publication.

Kropf, S. (2000), *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*, Shaker, Aachen.

Kropf, S. and J. Läuter (2002), Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data, *Biometrical Journal* **44**, 789–800.

Langsrud, Ø. (2004), Rotation tests, *Statistics and Computing*, submitted for publication.

Läuter, J. (1996), Exact t and F tests for analyzing studies with multiple endpoints, *Biometrics* **52**, 964–970.

Läuter, J., E. Glimm and S. Kropf (1996), New multivariate tests for data with an inherent structure, *Biometrical Journal* **38**, 5–23, Erratum: *Biometrical Journal* **40**, 1015.

Läuter, J., E. Glimm and S. Kropf (1998), Multivariate tests based on left-spherically distributed linear scores, *The Annals of Statistics* **26**, 1972–1988, Correction: *The Annals of Statistics* **27**, 1441.

Marcus, R., E. Peritz and K. R. Gabriel (1976), On closed testing procedures with special reference to ordered analysis of variance, *Biometrika* **63**, 655–660.

Schaafsma, W. and L. J. Smid (1966), Most stringent somewhere most powerful tests against alternatives restricted by a number of inequalities, *Annals of Mathematical Statistics* **37**, 1161–1172.

Schaafsma, W. and G. N. van Vark (1979), Classification and discrimination problems with applications, Part IIa, *Statistica Neerlandica* **33**, 91–126.

Westfall, P. H., S. Kropf and L. Finos (2004), Weighted FWE-controlling methods in high-dimensional situations, in: Y. Benjamini, F. Bretz and S. K. Sarkar (eds.), *Recent developments in multiple comparison procedures*, IMS Lecture Notes and Monograph Series, **47**, 143–154.

Westfall, P. H. and S. S. Young (1993), *Resampling-based multiple testing*, John Wiley & Sons, New York.

Yang, Y. H. and T. Speed (2003), Design and analysis of comparative microarray experiments, in: T. Speed (ed.), *Statistical analysis of gene expression microarray data*, Chapman & Hall/CRC, Boca Raton.

Received: July 2004. Revised: December 2004.

Unbiased Estimation of Selected Treatment Means in Two-Stage Trials

Jack Bowden[*,1] and **Ekkehard Glimm**[2]

[1] MRC Biostatistics Unit, Cambridge, CB2 0SR, UK
[2] Novartis Pharma AG, Lichtstrasse 35, 4002 Basel, Switzerland

Received 31 January 2008, revised 7 May 2008, accepted 21 May 2008

Summary

Straightforward estimation of a treatment's effect in an adaptive clinical trial can be severely hindered when it has been chosen from a larger group of potential candidates. This is because selection mechanisms that condition on the rank order of treatment statistics introduce bias. Nevertheless, designs of this sort are seen as a practical and efficient way to fast track the most promising compounds in drug development. In this paper we extend the method of Cohen and Sackrowitz (1989) who proposed a two-stage unbiased estimate for the best performing treatment at interim. This enables their estimate to work for unequal stage one and two sample sizes, and also when the quantity of interest is the best, second best, or j-th best treatment out of k. The implications of this new flexibility are explored via simulation.

Key words: Adaptive trial; Selection bias; Point estimation; UMVCUE.

1 Introduction

Conducting biomedical experiments can be exceedingly expensive; therefore it is vital that maximum information be gleaned from existing research, and that new experiments are designed as efficiently as possible. In the pharmaceutical setting, this need has led to the development of a statistical framework to support, where possible, adaptive clinical trials. The advantage of an adaptive trial is that the traditional learning and confirming roles of phase II and III studies can be combined into a single process, thereby making the process of drug development more expedient. Methodological developments have, for example, given trials the flexibility to stop early due to efficacy or futility, Bauer and Kieser (1999), or to re-assess sample sizes, Posch et al. (2003). For a more thorough review see Schmidli et al. (2006). A powerful facet of designs of this sort is that trial modifications need not be pre-defined. Furthermore, with careful utilisation of the closure principle, (Marcus et al., 1976), multiple hypotheses can be simultaneously tested whilst strictly conserving overall type I error.

A commonly pre-defined two-stage strategy is to firstly select the best performing of several candidate treatments after an initial interim analysis, and then secondly to test, in isolation, this treatment against a control. Thall et al. (1988) specifically considered how hypothesis testing could be implemented in this design for binary data, Stallard and Todd (2003) developed an approach to dealing with asymptotically normal responses. If uncorrected, point estimates for selected treatments exhibit a marked bias. However, 'exact' bias correction, by which we mean the proposal of unbiased estimates, has received relatively little attention in the literature. Possible reasons for this are now given. Firstly, the main focus in pharmaceutical statistics is hypothesis testing, and not point estimation. Secondly, maximum likelihood point estimates are often biased. However, rather than

[*] Corresponding author: e-mail: jack.bowden@mrc-bsu.cam.ac.uk, Phone: +44 1223 330385, Fax: +44 1223 330388

attempting to remove this bias completely, conservatively biased corrections have often been preferred, see for example Shen (2001) or Stallard and Todd (2005). Overcorrecting for bias will clearly keep type I error rates low, which naturally fits in with the ethos of industry regulators. Thirdly, biased estimates often have other desirable qualities, such as a small mean squared error, Posch et al. (2005).

Our focus in this paper *is* point estimation, and for that reason we concentrate on the work of Cohen and Sackrowitz (1989), who proposed an unbiased estimate for the mean of the best performing treatment. In Section 2 we describe in more detail the two-stage design under consideration, define three estimators for the selected treatment and decide on a common criteria for rating each estimator. In Section 3 we propose an extension to Cohen and Sackrowitz's UMVCUE and in Section 4 we investigate the properties of this estimator through various simulation studies.

2 The Two-Stage Design

The methods in this paper are motivated by considering the following two-stage clinical trial design. Let $X_i \sim N(\mu_i, \frac{\sigma^2}{n_1})$, $i = 1, \ldots, k$ be the stage 1 outcome measure, expressed as an average over n_1 subjects, for experimental treatments $T_1, \ldots, T_k$. For simplicity we will assume that the μ_i's are unknown but the variance σ^2 is known. All treatments are assessed and assigned a rank order, depending on the magnitude of their associated statistic. The treatment with the largest stage 1 mean, $X_{(1)}$, is then taken forward to a second stage and tested on a new population of size n_2. Let the true mean of this treatment be denoted by $\mu_{(1)}$. The statistic derived in stage 2 will be referred to as Y and, conditional on the stage 1 treatment selection, follows a $N(\mu_{(1)}, \frac{\sigma^2}{n_2})$ distribution.

Since $X_{(1)}$ is the maximum of k random variables it is no longer normally distributed and consequently $E[X_{(1)}] > \mu_{(1)}$. Note that $\mu_{(1)}$ is *not* the maximum of $\mu_1, \ldots, \mu_k$. It is a random variable that can take any of the true treatment mean values $\mu_1, \ldots, \mu_k$. For this reason we believe that, in order to evaluate the performance of a generic estimator for $\mu_{(1)}$, say $\mu^*_{(1)}$, the quantities

$$b_{\text{sel}}(\mu^*_{(1)}) = \sum_{i=1}^{k} E[\mu^*_{(1)} - \mu_i \mid X_{(1)} = X_i]\, P(X_{(1)} = X_i)\,, \tag{1}$$

$$\text{MSE}_{\text{sel}}(\mu^*_{(1)}) = \sum_{i=1}^{k} E[(\mu^*_{(1)} - \mu_i)^2 \mid X_{(1)} = X_i]\, P(X_{(1)} = X_i) \tag{2}$$

are most insightful. (1) and (2) are essentially the weighted bias and mean squared error (MSE) across all values of $\mu_{(1)}$, and were first introduced by Posch et al. (2005). They have also been used to compare the performance of two-stage estimators by Sill and Sampson (2007). No unbiased estimate for $\mu_{(1)}$ exists based on stage 1 data alone, a result commonly attributed to Putter and Rubenstein (1968). A more recent and accessible proof of this fact, albeit only for the case $k = 2$, appears in Stallard et al. (2008).

2.1 Estimation of $\mu_{(1)}$

Obviously, Y is an unbiased estimate of $\mu_{(1)}$. However, the variance of Y $(\frac{\sigma^2}{n_2})$, is large when one considers that $n_1 + n_2$ subjects could in principle contribute to $\mu_{(1)}$'s estimate. For this reason, we will term Y the 'inefficient' estimator. Alternatively, the maximum likelihood estimate (MLE) for $\mu_{(1)}$

$$\hat{\mu}_{(1)} = \frac{n_1 X_{(1)} + n_2 Y}{n_1 + n_2} \tag{3}$$

a weighted average of the first and second stage estimates, could be used instead. However (3) ignores the selection mechanism, hence it is biased. Posch et al. (2005) show that when $k = 2$,

$\mathrm{MSE}_{\mathrm{sel}}(\hat{\mu}_{(1)}) = \frac{\sigma^2}{n_1+n_2}$ for the MLE. They note that this is equal to the MSE for the case when $k = 1$, i.e. where no selection bias is possible. Furthermore, using a symmetry argument, Posch et al. show that this equality holds for $k = 1$ and 2, when *any* selection rule, that depends on the difference between the two treatment estimates X_1 and X_2, is used.

By conditioning on the stage 1 treatment estimates, Cohen and Sackrowitz (1989) proposed an unbiased estimator for $\mu_{(1)}$, which we will call $\tilde{\mu}_{(1)}$. Assuming that $X_1, \ldots, X_k$ and Y have unit variance, that is when $\frac{\sigma^2}{n_1} = \frac{\sigma^2}{n_2} = 1$, Cohen and Sackrowitz's formula for $\tilde{\mu}_{(1)}$ is

$$Z/2 - \frac{1}{\sqrt{2}} \frac{\phi(W)}{\Phi(W)} \tag{4}$$

where $\phi(.)$ and $\Phi(.)$ are the pdf and cdf of the standard normal distribution. $Z = X_{(1)} + Y$ and $W = \sqrt{2}(Z/2 - X_{(2)})$. $X_{(2)}$ is defined as the second best performing treatment out of the k treatments in stage 1. (4) is essentially the MLE $(Z/2)$ minus a correction term. It is an expression for the expected value of the unbiased second stage data Y, conditional on Z and the order of the stage 1 treatment estimates. The sufficiency and completeness of $(Z, X_{(2)}, \ldots, X_{(k)})$ with respect to $(\mu_{(1)}, \ldots, \mu_{(k)})$ mean that by the Lehmann–Schéffe theorem, (4) is the uniformly minimum variance conditional unbiased estimate (UMVCUE) for $\mu_{(1)}$. Note that the UMVCUE fulfills a stronger condition of unbiasedness as that required by formula (1), namely

$$E[\tilde{\mu}_{(1)} - \mu_i \mid X_{(1)} = X_i] = 0\,, \quad \forall i = 1, \ldots, k\,.$$

3 Extending the UMVCUE

Cohen and Sackrowitz only specified the UMVCUE for the case of equal variances in the first and second stage, and also as an estimator for the most extreme out of k statistics. Extending their formula to unequal variances would provide, for example, the flexibility to conduct an interim analysis at any point. Furthermore, interest may not always be solely restricted to treatment $T_{(1)}$ and unbiased estimates for the j-th best performing out of k treatments may also be valued. Finally, one may also wish to select a treatment with a criterion other than its effect size. With this in mind we now propose a corollary to their original proof.

Let X_i, $i = 1 \ldots, k$ be independently $N(\mu_i, \sigma^2_{1,i})$ distributed. Let $Y_i \sim N(\mu_{(i)}, \sigma^2_{2,i})$. The general, two stage UMVCUE for the j-th most extreme statistic is given by

$$\tilde{\mu}_{(j)} = \frac{\sigma^2_{2,j} X_{(j)} + \sigma^2_{1,(j)} Y_j}{\sigma^2_{1,(j)} + \sigma^2_{2,j}} - \frac{\sigma^2_{2,j}}{\sqrt{\sigma^2_{1,(j)} + \sigma^2_{2,j}}} \frac{\{\phi(W_{j,j+1}) - \phi(W_{j,j-1})\}}{\{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})\}} \tag{5}$$

where $W_{s,t} = \frac{1}{\sigma^2_{1,(s)}} \left(\frac{\sigma^2_{2,s} X_{(s)} + \sigma^2_{1,(s)} Y_s}{\sqrt{\sigma^2_{1,(s)} + \sigma^2_{2,s}}} - X_{(t)} \sqrt{\sigma^2_{1,(s)} + \sigma^2_{2,s}} \right)$, with $s = 1, \ldots, k$, $t = 0, \ldots, k+1$ and $X_{(0)} := \infty > X_{(1)} \geq \ldots \geq X_{(k)} > X_{(k+1)} := -\infty$. $\sigma^2_{1,(s)}$ refers to the variance of the (s) largest treatment and not the (s) largest variance.

The proof is very similar to Cohen and Sackrowitz's. The initial factorisation of $X_{(j)}$ and Y_j is different, as are certain limits for integration over these two variables.

Without loss of generality let Q be the event that $X_1 > \ldots > X_k$, so that $X_{(j)} = X_j$. To simplify notation, we will write f instead of f_X for the pdf of a random variable X. Otherwise, we will stick to the convention that capital letters denote random variables and small letters their realizations. The

joint distribution of Y_j and $X = (X_1, \ldots, X_k)$ given Q has the density

$$
\begin{aligned}
& K^{-1}(\mu)\, \frac{1}{\sigma_{2,j}}\, \phi\left(\frac{y_j - \mu_j}{\sigma_{2,j}}\right) \frac{1}{\sigma_{1,j}} \phi\left(\frac{x_j - \mu_j}{\sigma_{1,j}}\right) I_Q(x) \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}} \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right) \\
&= K^{-1}(\mu) \frac{1}{\sigma_{2,j}} \phi\left(\frac{\frac{\sigma_{2,j}}{\sigma_{1,j}} x_j + \frac{\sigma_{1,j}}{\sigma_{2,j}} y_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) \frac{1}{\sigma_{1,j}} \phi\left(\frac{x_j - y_j}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) I_Q(x) \\
&\quad \times \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}} \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right) \\
&= K^{-1}(\mu) \frac{1}{\sigma_{2,j}} \phi\left(\frac{\frac{\sigma_{2,j}}{\sigma_{1,j}} x_j + \frac{\sigma_{1,j}}{\sigma_{2,j}} y_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) \frac{1}{\sigma_{1,j}} \phi\left(\frac{x_j - \frac{\frac{\sigma_{2,j}}{\sigma_{1,j}} x_j + \frac{\sigma_{1,j}}{\sigma_{2,j}} y_j}{\alpha_{1,j}}}{\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} \alpha_{2,j}}\right) I_Q(x) \\
&\quad \times \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}} \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right),
\end{aligned}
$$

where $\alpha_{1,j} = \frac{\sigma_{1,j}}{\sigma_{2,j}} + \frac{\sigma_{2,j}}{\sigma_{1,j}}$ and $\alpha_{2,j} = \frac{\sigma_{2,j}^2}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}$. $I_Q(x)$ is the indicator function for the event, $I_Q(x) = \{\infty > x_1 > \ldots > x_k > -\infty\}$ and $K(\mu) = P_\mu(I_Q(X) = 1)$. Define the set of $k-1$ random variables X_j^c to be $\{X_i, i = 1, \ldots, k, \quad i \neq j\}$. The pair x_j^c and $z_j = \frac{\sigma_{2,j}}{\sigma_{1,j}} x_j + \frac{\sigma_{1,j}}{\sigma_{2,j}} y_j$ are sufficient and complete statistics for $\mu_1, \ldots, \mu_k$. Transforming $f(X, Y_j \mid Q)$ to $f(X, Z_j \mid Q)$ and $f(Y_j, X_j^c, Z_j \mid Q)$ respectively gives

$$
f(X, Z_j \mid Q) = K^{-1}(\mu)\, \phi\left(\frac{z_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) \frac{1}{\sigma_{1,j}^2}\, \phi\left(\frac{x_j - \frac{z_j}{\alpha_{1,j}}}{\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} \alpha_{2,j}}\right) I_Q(x) \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}}\, \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right),
$$

$$
f(Y_j, X_j^c, Z_j \mid Q) = K^{-1}(\mu)\, \phi\left(\frac{z_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) \frac{1}{\sigma_{2,j}^2}\, \phi\left(\frac{(y_j - \frac{z_j}{\alpha_{1,j}})}{\alpha_{2,j}}\right) I_Q(x) \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}}\, \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right).
$$

The density $f(X_j^c, Z_j \mid Q)$ is obtained from $f(X, Z_j \mid Q)$ via the integral

$$
\begin{aligned}
& K^{-1}(\mu) \frac{1}{\sigma_{1,j}^2} \phi\left(\frac{z_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) I_Q(x) \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}} \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right) \int_{x_{j+1}}^{x_{j-1}} \phi\left(\frac{x_j - \frac{z_j}{\alpha_{1,j}}}{\frac{\sigma_{1,j}^2}{\sigma_{2,j}^2} \alpha_{2,j}}\right) \mathrm{d}x_j \\
&= K^{-1}(\mu) \frac{\alpha_{2,j}}{\sigma_{2,j}^2} \phi\left(\frac{z_j - \mu_j \alpha_{1,j}}{\sqrt{\sigma_{1,j}^2 + \sigma_{2,j}^2}}\right) I_Q(x) \prod_{i=1, i\neq j}^{k} \frac{1}{\sigma_{1,i}} \phi\left(\frac{x_i - \mu_i}{\sigma_{1,i}}\right) \{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})\}.
\end{aligned}
$$

The distribution of $f(Y_j \mid X_j^c, Z; Q) = \frac{f(Y_j, X_j^c, Z \mid Q)}{f(X_j^c, Z \mid Q)}$ is

$$
\frac{\frac{1}{\alpha_{2,j}} \phi\left(\frac{\left(y_j - \frac{z_j}{\alpha_{1,j}}\right)}{\alpha_{2,j}}\right)}{\{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})\}} I\left(\frac{\sigma_{2,j}}{\sigma_{1,j}}\left(z_j - \frac{\sigma_{2,j}}{\sigma_{1,j}} x_{j-1}\right) < y_j < \frac{\sigma_{2,j}}{\sigma_{1,j}}\left(z_j - \frac{\sigma_{2,j}}{\sigma_{1,j}} x_{j+1}\right)\right)
$$

and therefore $E[f(Y_j \mid X_j^c, Z_j, Q)]$ is equal to

$$\frac{\frac{1}{\alpha_{2,j}} \int_{\frac{\sigma_{2,j}}{\sigma_{1,j}}\left(z_j - \frac{\sigma_{2,j}}{\sigma_{1,j}} x_{j-1}\right)}^{\frac{\sigma_{2,j}}{\sigma_{1,j}}\left(z_j - \frac{\sigma_{2,j}}{\sigma_{1,j}} x_{j+1}\right)} y_j \phi\left(\frac{\left(y_j - \frac{z_j}{\alpha_{1,j}}\right)}{\alpha_{2,j}}\right) \mathrm{d}y_j}{\{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})\}}.$$

Using a standard result, (see for example Todd et al. (1966)), that

$$\int_{-\infty}^{T} \frac{y}{\sigma} \phi\left(\frac{y-\mu}{\sigma}\right) \mathrm{d}y = -\sigma\phi\left(\frac{T-\mu}{\sigma}\right) + \mu\Phi\left(\frac{T-\mu}{\sigma}\right)$$

and additionally noting that

$$\left(\frac{\sigma_2}{\sigma_1}\left(z - \frac{\sigma_2}{\sigma_1} x\right) - \frac{z}{\alpha_1}\right) \frac{\sigma_1^2}{\sigma_2^2} = \frac{z}{\alpha_1} - x$$

the expectation reduces to

$$\frac{-\alpha_{2,j}\{\phi(W_{j,j+1}) - \phi(W_{j,j-1})\} + \frac{z_j}{\alpha_{1,j}}\{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})\}}{\Phi(W_{j,j+1}) - \Phi(W_{j,j-1})}$$

and we have our result. When treatment $T_{(1)}$ is selected, $\phi(W_{1,0}) = \Phi(W_{1,0}) = 0$, additionally setting $\sigma_{1,(1)} = \sigma_{2,1}$ reduces (5) to Cohen and Sackrowitz's original expression.

3.1 Estimation following a ranking by *p*-value

Suppose that $\sigma_{1,j}^2$, $j = 1, \ldots, k$ are all unique and reasonably dispersed. One may prefer to rank the candidate treatments $T_1, \ldots, T_k$ in order of merit according to the statistical significance, rather than simply the magnitude of their estimated effect. Assuming null hypotheses $\mu_j = 0$, $j = 1, \ldots, k$, with a one sided alternative, $\mu_j > 0$, the consequence of ranking by 'p-value' will be that $x_{(j)} \in \left(\frac{x_{(j-1)}\sigma_{1,(j)}}{\sigma_{1,(j-1)}}, \frac{x_{(j+1)}\sigma_{1,(j)}}{\sigma_{1,(j+1)}}\right)$, and $y_j \in \left(\frac{\sigma_{2,j}}{\sigma_{1,(j)}}\left(z_{(j)} - \frac{\sigma_{2,j}}{\sigma_{1,(j-1)}} x_{(j-1)}\right), \frac{\sigma_{2,j}}{\sigma_{1,(j)}}\left(z_{(j)} - \frac{\sigma_{2,j}}{\sigma_{1,(j+1)}} x_{(j+1)}\right)\right)$. The previous proof follows through almost identically if we condition on the event Q^*: $\frac{X_1}{\sigma_{1,1}} > \ldots > \frac{X_k}{\sigma_{1,k}}$ instead, to leave formula (5), except that the $X_{(t)}$ in $W_{s,(t)}$ is replaced by $\frac{\sigma_{1,(s)} X_{(t)}}{\sigma_{1,(t)}}$.

3.2 'Drop-and-estimate' the loser

If one selected the top $(k-1)$ treatments at interim, then we now have the framework to unbiasedly estimate their means and also the mean of the dropped treatment $\mu_{(k)}$via

$$\breve{\mu}_{(k)} = X_{(k)} + \sum_{i=1}^{k-1} (X_{(i)} - \tilde{\mu}_{(i)}).$$

This follows from the fact that $E(\sum_{i=1}^{k} X_{(i)}) = E(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} \mu_{(i)} = \sum_{i=1}^{k} \mu_i$. Clearly however the the variance of $\breve{\mu}_{(k)}$ will be large.

4 Performance of the UMVCUE: Simulation Studies

4.1 Estimating $\mu_{(1)}$: Varying the interim time

We first consider a two-stage trial involving three candidate treatments T_1, T_2, T_3, with corresponding means μ_1, μ_2, μ_3 = (0, 1/2, 0). Their estimates have the same variance at stage 1 and stage 2 (σ_1^2 and σ_2^2), and only the treatment with the best estimate is selected after stage 1. Letting $\sigma_1^2 = \frac{1}{n_1}$ and $\sigma_2^2 = \frac{1}{n_2}$

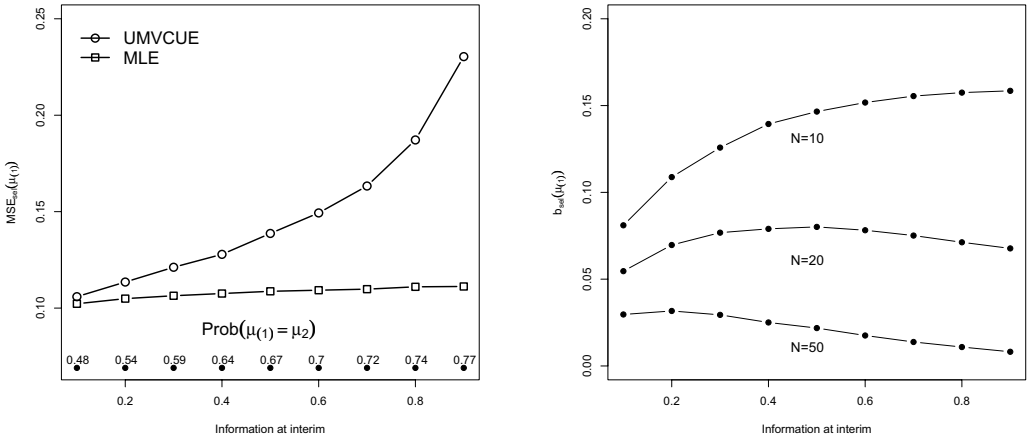


Figure 1 Left: Mean squared error of the MLE and UMVCUE methods when the amount of information at interim, I, is allowed to vary. Right: The bias present in the MLE for varying I and N.

we fix the final sample size, over the two stages, ($n_1 + n_2 = N$) but the relative size of n_1 and n_2 is allowed to vary. This can be viewed as varying the amount 'Information' at interim, which we define as

$$I = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2} = \frac{n_1}{n_1 + n_2}\,.$$

Ideally one would hope that the truly best treatment – T_2, is selected at interim, but this will not always occur. For $\hat{\mu}_{(1)}$ and $\tilde{\mu}_{(1)}$, the size of the MSE is investigated for varying ratios of σ_1^2 and σ_2^2. Figure 1 (Left) shows, for $N = 10$, Monte-Carlo estimates for the MSE as I is varied between 0 and 1. Each estimate is based on 100 000 simulations. The MSE increases dramatically as I increases for the UMVCUE, but it is relatively stable for the MLE, despite the fact that this quantity includes a considerable proportion of bias. Of course, one must not conclude that it is best to go to interim as early as possible, since the probability of choosing the right treatment, T_2, decreases as I increases. The probabilities given at the foot of this plot indicate, for a particular value of I, the proportion of times that $\mu_{(1)} = \mu_2$. For $I = 0.1$, the correct choice is made less than 50% of the time. When $I = 0.9$, T_2 is chosen close to 80% of the time.

Making the wrong choice at interim does not affect the bias of the UMVCUE – it will unbiasedly estimate the mean of whatever treatment is selected, but this is of course not true for the MLE. Figure 1 (Right) shows, for the same trial trial design as before, how the bias of the MLE is modulated by I as well as the total sample size, N. For a relatively small trial ($N = 10$) the bias appears to increase monontonically, albeit more slowly, as I increases. However when the sample size is doubled the bias appears to be maximised at approximately $I = 0.5$. For a trial five times larger ($N = 50$) the bias is maximised for an I closer to 0.2.

4.2 Estimating $\mu_{(1)}$: Varying treatment arm sample size

We now investigate the effect of varying the size of each treatment arm on the ability of the MLE and UMVCUE to estimate $\mu_{(1)}$. We consider a trial with 2 treatments, T_1, T_2 with stage 1 mean estimates drawn from $N(\mu_1, \frac{1}{n_{T_1}})$ and $N(\mu_1, \frac{1}{n_{T_2}})$ distributions respectively. The total sample size for stage 1 is fixed to be $n_{T_1} + n_{T_2} = 10$, but different numbers of subjects are apportioned to each treatment arm.

Table 1 Bias and Mean squared error ($\times 100$) of the MLE and UMVCUE for varying treatment arm sample sizes. All figures are the average of 100 000 simulations.

Treatment arm		Bias $\times 100$ MLE $\hat{\mu}_{(1)}$	Mean squared error $\times 100$	
T_1	T_2		MLE $\hat{\mu}_{(1)}$	UMVCUE $\tilde{\mu}_{(1)}$
$\mu_1 = 0$	$\mu_2 = 0$			
n_{T_1}	n_{T_2}			
9	1	19.90	15.16	15.26
8	2	13.99	9.23	9.75
7	3	11.40	7.46	8.31
6	4	9.70	6.84	7.96
5	5	8.46	6.69	8.03
$\mu_1 = 1$	$\mu_2 = 0$			
n_{T_1}	n_{T_2}			
9	1	12.77	13.60	10.07
8	2	6.26	8.02	6.89
7	3	4.02	6.80	6.57
6	4	2.81	6.55	6.73
5	5	2.28	6.66	7.08
4	6	2.85	7.18	7.55
3	7	4.06	8.23	8.59
2	8	6.27	10.48	10.84
1	9	12.67	16.92	17.29

The treatment with the largest stage 1 mean is selected at interim and tested on 10 further subjects, so that the total sample size of the trial is 20. Rows 1–5 of Table 1 show the results of this simulation when $\mu_1 = \mu_2 = 0$ and the ratio of $n_{T_1} : n_{T_2}$ is varied from 9 : 1 to 5 : 5. When both treatments have the same true mean, a balanced design with 5 subjects in both stage 1 arms appears to be the best option. As the subject ratio increases the bias in the selected treatment mean rises, which is reflected in the increased bias of the MLE. The MSE increases as the arms become less balanced, for both the MLE and the UMVCUE. Rows 6–10 show the results when the true treatment means are different ($\mu_1 = 1$, $\mu_2 = 0$) and the truly best treatment – T_1 is assigned a larger number of subjects. As before, the bias in the MLE increases as the treatment arms become less balanced. However, the MSE of the MLE is minimised when T_1 is assigned to 60% and T_2 is assigned to 40% of the subjects. For the UMVCUE a 70–30% split provides the smallest MSE, and the UMVCUE actually has a smaller MSE than the MLE for higher imbalances. Rows 11–15 show the results of the same simulation where more subjects are assigned to the truly worst treatment – T_2 in stage 1. This appears to be the worst case scenario, as the treatment arm imbalance increases the MSE of the MLE and UMVCUE increase. Although the MLE's MSE is the smallest there is not much to choose between the two estimators.

4.3 Conditional mean squared error of MLE versus UMVCUE

In the case of two treatments with one selected at interim, Posch et al. (2005) show that $\mathrm{MSE}_{\mathrm{sel}}(\hat{\mu}_{(1)})$ equals $\frac{\sigma^2}{n_1+n_2}$ for the MLE. However, it is easy to show that, conditional on stage 1 treatment selection,

$$\mathrm{MSE}_{\mathrm{sel}}(\hat{\mu}_{(1)} \,|\, X_1 > X_2) \geq \mathrm{MSE}_{\mathrm{sel}}(\hat{\mu}_{(1)} \,|\, X_1 < X_2)$$

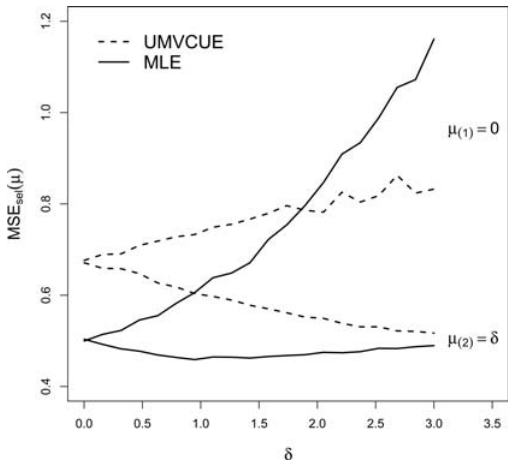


Figure 2 Conditional mean squared error of the MLE and UMVCUE for the case of an adaptive trial with two treatments, with means $\mu_1 = 0$ and $\mu_2 = \delta$.

if $\mu_1 > \mu_2$. For example, if we assume equal sample size $n_1 = n_2$ in both treatments as well as in stage 1 and 2, and $\sigma^2/n_1 = 1$, then

$$\mathrm{MSE}(\hat{\mu}_{(1)} \,|\, X_2 > X_1) = \frac{1}{2} + \frac{1}{8}\frac{\delta}{\sqrt{2}} \cdot \frac{\phi\left(\frac{\delta}{\sqrt{2}}\right)}{\Phi\left(-\frac{\delta}{\sqrt{2}}\right)} \tag{6}$$

and

$$\mathrm{MSE}(\hat{\mu}_{(1)} \mid X_1 > X_2) = \frac{1}{2} - \frac{1}{8}\frac{\delta}{\sqrt{2}} \cdot \frac{\phi\left(\frac{\delta}{\sqrt{2}}\right)}{\Phi\left(\frac{\delta}{\sqrt{2}}\right)}$$

where $\mu_1 - \mu_2 = \delta$. The weighted average of the two MSEs gives

$$\mathrm{MSE}_{\mathrm{sel}}(\hat{\mu}_{(1)}) = \mathrm{MSE}(\mu_{(1)} \,|\, X_1 > X_2) \cdot \Phi\left(\frac{\delta}{\sqrt{2}}\right) + \mathrm{MSE}(\mu_{(1)} \,|\, X_2 > X_1) \cdot \Phi\left(-\frac{\delta}{\sqrt{2}}\right) = \frac{1}{2}$$

in this case. Formula (6) shows that the MSE of the MLE increases dramatically, if the wrong selection is made at interim. This tendency is much reduced with the UMVCUE. Figure 2 shows that the conditional MSE of the UMVCUE also increases if the wrong choice is made at interim, but not nearly as quickly as the conditional MSE of the MLE. The figure is based on 2 million simulations of the described two-treatment scenario (100 000 for each of the twenty δ values evaluated) with $\sigma^2/n_1 = 1$.

4.4 Estimation of $\mu_{(j)}$ and their distribution

For a trial with $k = 5$ treatments, we investigate the MSE of the UMVCUE when estimating the j-th best treatment's mean $\mu_{(j)}$ for $j = 1, \ldots, 5$. To do this we define, for a generic estimator $\mu^*_{(j)}$

$$\mathrm{MSE}_{\mathrm{sel}}(\mu^*_{(j)}) = \sum_{i=1}^{k} E[(\mu^*_{(j)} - \mu_i)^2 \,|\, X_{(j)} = X_i] P(X_{(j)} = X_i)$$

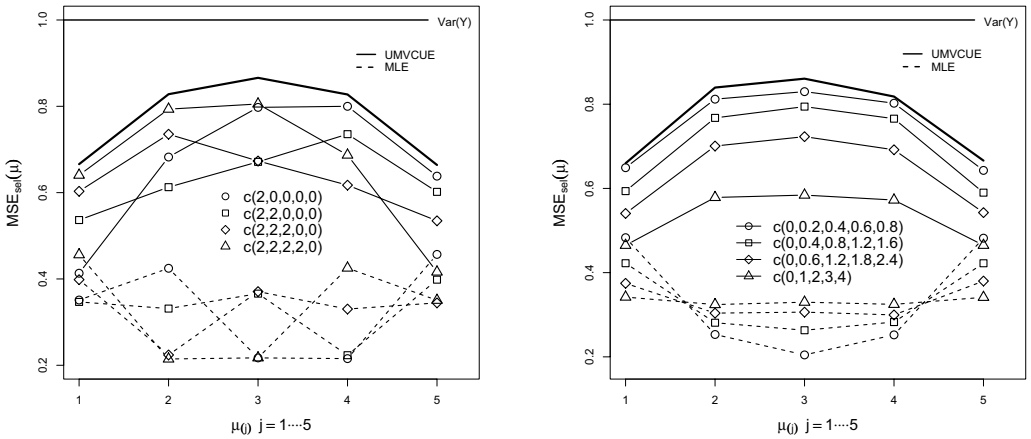


Figure 3 Mean squared error of the modified UMVCUE and MLE under 5 different values of $\mu_1, \ldots, \mu_5$, highlighting; (Left) the effect of unequal distances between true treatment effects, (Right) the effect of increasing the magnitude of the distance between true treatment effects. The case where $\mu_1 = \ldots = \mu_5 = 0$ is shown by a solid black line.

in analogy to formula (2). Each stage 1 treatment mean is given an underlying variance of $\frac{1}{2}$. All treatments are effectively selected and a stage 2 statistic is calculated with a variance of 1, so that $I = \frac{2}{3}$. This enables MLE's and UMVCUE's for $\mu_{(1)}, \ldots, \mu_{(5)}$ to be obtained. Figure 3 (left) illustrates the effect of varying the number of treatments that are *a priori* significantly different from 0 on the MSE of $\hat{\mu}_{(j)}$ and $\tilde{\mu}_{(j)}$. Figure 3 (right) illustrates the effect of varying the magnitude of the difference between each treatment. Values are again Monte-Carlo estimates based on 100 000 simulations.

Estimation of $\tilde{\mu}_{(1)}$ requires $X_{(1)}$ and only one other stage 1 statistic, $X_{(2)}$. Cohen and Sackrowitz regarded this as a negative point, because only a small fraction of the available data is used. Estimat-

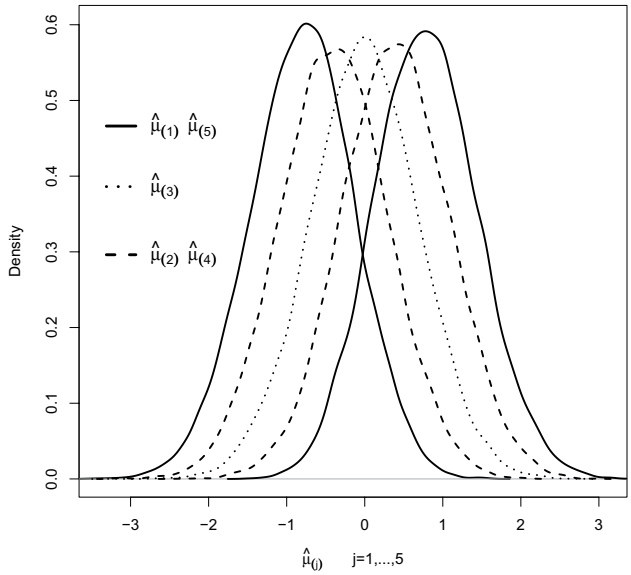


Figure 4 The distribution of the UMVCUE for $\mu_{(1)}, \ldots, \mu_{(5)}$ (shifted on the x axis by $(-0.8, -0.4, 0, 0.4, 0.8)$ to highlight their shape).

ing $\tilde{\mu}_{(j)}$ for $j = 2, \ldots, k-1$ utilises $X_{(j)}$, $X_{(j-1)}$ *and* $X_{(j+1)}$. However, in terms of MSE, the UMVCUE generally performs worse when used to estimate a treatment from the middle of the ordered range, not better. A nice feature of these plots is that $\mathrm{MSE}_{\mathrm{sel}}(\tilde{\mu}_{(j)})$ appears to be maximised when $\mu_1 = \ldots = \mu_k$ (shown by the solid black line in both plots). This makes intuitive sense, and agrees with results in Sill and Sampson (2007), but no formal proof of this is offered in this paper.

Figure 4 shows the distribution of 20 000 UMVCUE estimates for $\mu_{(1)}, \ldots, \mu_{(5)}$, for the case when $\mu_1 = \ldots = \mu_5 = 0$. So that their shape can be seen more clearly, all distributions but for that of $\tilde{\mu}_{(3)}$ have been shifted to the left or right of 0. All distributions are very well approximated by a normal distribution. However, the densities of the estimates for $\tilde{\mu}_{(1)}$ and $\tilde{\mu}_{(5)}$ are more peaked, and close inspection reveals a small amount of asymmetry in their extreme tails. This asymmetry is still present but less marked for $\tilde{\mu}_{(2)}$ to $\tilde{\mu}_{(4)}$.

4.4.1 Confidence intervals for $\mu_{(j)}$

If we are willing to assume the approximate normality of $\tilde{\mu}_{(j)}$ and also accept that an upper bound for the variance of $\tilde{\mu}_{(j)}$ is achieved when $\mu_1 = \mu_2, = \ldots = \mu_k$, then a conservative α-level confidence interval for our modified UMVCUE would naturally take the form $\tilde{\mu}_{(j)} \pm \Phi^{-1}(1-\alpha/2)\, V$, where V^2 is this maximal variance, and is easily approximated to a high degree of accuracy given k, σ_1^2, σ_2^2. Figure 5 shows the results of a simulation study to assess the coverage of this conservative 95% interval, for different values of $\mu_1, .., \mu_5$. When the true difference between each treatment is 0, the coverages for $\tilde{\mu}_{(1)}$ and $\tilde{\mu}_{(5)}$ are just above the nominal level. Conversely, the coverages of $\tilde{\mu}_{(2)}$ to $\tilde{\mu}_{(4)}$ appear slightly below their nominal level. When there is a small difference between the true treatment means, all of the estimate's coverage probabilities appear at, or above, their nominal 95% level.

4.5 Trial data example

We simulate a single two-stage adaptive trial with three active treatment arms and one control arm. For a single subject, treatment and placebo outcomes were generated from a normal distribution with true means μ_1, μ_2, μ_3, $\mu_c = (1, 2, 0.5, 0)$ and a common, known standard deviation of $\sigma = 7$. Fifty

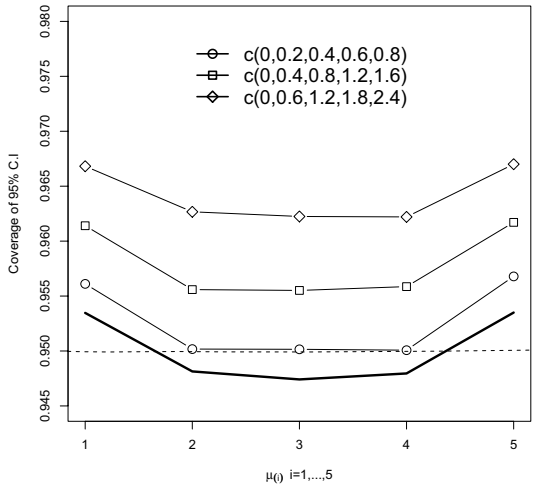


Figure 5 Coverage probabilities of the conservative confidence interval for $\hat{\mu}_{(1)}, \ldots, \hat{\mu}_{(5)}$ under the assumption that the UMVCUE is approximately normally distributed.

Table 2 Simulated two-stage adaptive trial data.

Treatment Arm	**Stage 1**	**Stage 2** (Variance: $\frac{S^2}{100}$)	MLE (Variance: $\frac{S^2}{150}$)	UMVCUE (Variance: V)
T_1: $\mu_1 = 1$	$X_{(2)}$: 1.70	Y_2: 1.16(0.48)	1.34(0.32)	1.37(0.44)
T_2: $\mu_2 = 2$	$X_{(1)}$: 2.82	Y_1: 2.70(0.48)	2.74(0.32)	2.66(0.40)
T_3: $\mu_3 = 0.5$	$X_{(3)}$: -0.57			
Control				0.54(0.32)
95% C.I				
$\mu_{(1)} - \mu_c$		(0.40, 3.90)	(0.61, 3.78)	(0.42, 3.80)
$\mu_{(2)} - \mu_c$		(-1.14, 2.36)	(-0.79, 2.38)	(-0.90, 2.53)

subjects were assigned to each treatment and control arm in stage 1. The top two stage 1 treatments (in rank order T_2 and T_1), were selected and tested along with the control on a further one hundred subjects. Y_1 and Y_2 therefore provide unbiased estimates for $\mu_{(1)} = \mu_2 = 2$ and $\mu_{(2)} = \mu_1 = 1$ respectively. All observations (4×50 from stage 1 and 3×100 from stage 2) were used to obtain a pooled estimate for σ of 6.92. Table 2 shows the inefficient stage 2 estimate, the MLE and the UMVCUE for $\mu_{(1)}$ and $\mu_{(2)}$, as well as their variances. The variances of the UMVCUE were obtained by simulating the two-stage scenario 50 000 times. As explained in Section 4.4.1, all treatments were assigned means of 0 to make the variance estimate conservative. Common stage-1 and stage-2 variances of $6.92^2/50$ and $6.92^2/100$, respectively, were also assumed. Repeated simulation of this trial design with the true parameter values, $\mu_1, \mu_2, \mu_3 = (1, 2, 0.5)$, rather than $\mu_1 = \mu_2 = \mu_3 = 0$ showed that the average bias in $\hat{\mu}_{(1)}$ and $\hat{\mu}_{(2)}$ is approximately 0.2 and 0.02 respectively. The MSE of $\hat{\mu}_{(1)}$ and $\hat{\mu}_{(2)}$ were 0.34 and 0.28 respectively. The MSE for $\tilde{\mu}_{(1)}$ and $\tilde{\mu}_{(2)}$ were 0.39 and 0.42 respectively.

Table 2 also shows 95% confidence intervals for the selected treatment-control differences using the MLE and UMVCUE. As expected, confidence intervals based on the MLE are narrower than those based on the UMVCUE, but their α-level control is clearly suspect. Using the stage 1 data on all three treatments plus the UMVCUE's for the top two treatments as in Section 3.2, an unbiased estimate for $\mu_{(3)}$ is calculated to be $\breve{\mu}_{(3)} = -0.06$.

5 Discussion

Standard two stage designs concentrate solely on the best performing treatment at interim, since this will provide maximal power to prove efficacy at stage 2. When economically feasible, it may prove prudent to take forward more than one treatment if, for example, at a later date the best performing treatment exhibits other undesirable properties. Our modified estimator makes unbiased estimation of multiple selected treatments possible, even if their estimates have varying precision. However, it must be made clear that treatment selection decisions must either be made in advance of the trial commencing, or only be made conditional on the stage 1 estimate's rank order at interim for the UMVCUE to be valid. If other information is allowed to influence which treatments are taken forward, such as their actual stage 1 values, then the estimators' properties could be compromised.

We have attempted to explore the strengths and weaknesses of the UMVCUE and the MLE under various trial designs. In our simulations the MLE generally outperforms the UMVCUE by a clear margin in terms of MSE, although they tend to increase or decrease together when one particular facet of the trial is changed. Whether one prefers the UMVCUE over the MLE is a question of how much the notion of bias matters – in the pharmaceutical setting unbiasedness undoubtedly has a high currency with industry regulators. The UMVCUE performs particularly poorly relative to the MLE when

there is relatively little unbiased information to 'Rao-Blackwellise'. That is, when stage 1 of the trial is larger than stage 2, for a fixed total sample size. This has been previously noted in a simulation study by Sill and Sampson (2007). Although it is not a characteristic of the trial that can be altered, the underlying values of the parameters $\mu_1, \ldots, \mu_k$ also have a large impact on both estimators' performance. Simulations suggest that increasing the precision the of truly best treatment's stage 1 estimate at the expense of the other treatments can reduce the MSE of both the MLE and UMVCUE. However, the UMVCUE appears to benefit even more from this. To make use of this feature it is of course necessary to speculate about the best treatment when allocating the sample size. This can backfire when the wrong treatment is accidentally up-weighted. Perhaps a more realistic situation in which unequal stage 1 variances would be encountered is if treatment arms started off balanced, but were affected by differing amounts of subject drop outs. As long as the drop out data could be assumed to be missing at random, then our modified UMVCUE could be utilised as normal.

Also noted by Sill and Sampson (2007) is the fact that, all other characteristics being equal, the MSE of the UMVCUE for $\mu_{(1)}$ is highest when $\mu_1 = \ldots = \mu_k$. Our simulations show that this is also true for $\tilde{\mu}_{(2)}, \ldots, \tilde{\mu}_{(k)}$ and we exploited this to produce a conservative Monte-Carlo estimate for the UMVCUE's variance. Future research on this topic might concentrate on proving that the variance of the UMVCUE is indeed maximised at $\mu_1 = \ldots = \mu_k$, or better still to obtain an expression for its actual variance. The UMVCUE's confidence interval could then be used directly to prove efficacy over a control treatment, as in Section 4.5. However, it is perhaps asking too much of the UMVCUE to expect that it can be both a point estimate and an effective basis for hypothesis testing. For example, in recent work Sampson and Sill (2005) have shown that an alternative statistic to the UMVCUE forms the basis of a 'uniformly most powerful' test for $\mu_{(1)}$. It would be interesting to see if this could be extended to work for an arbitrary $\mu_{(j)}$ also.

It must be noted that $\tilde{\mu}_{(j)}$ can only claim to be the minimum variance unbiased estimate for $\mu_{(j)}$ marginally, that is when only stage 2 data exists for the j-th largest stage 1 treatment. If for example, as in Section 4.5, two out of three treatments from stage 1 are carried forward, then the stage 2 information on the second best stage 1 treatment, Y_2, could also potentially be incorporated into an estimator for $\mu_{(1)}$. The extra information should lead to a variance reduction. Moreover, rather than conditioning on the full ordering constraint $X_1 > X_2 > X_3$, unbiased estimators employing less stringent interim rules, such as $X_1 > X_3$, $X_2 > X_3$, may also lead to an estimator for $\mu_{(1)}$ with a smaller variance. Future research into these areas would certainly be worthwhile.

Acknowledgements *The authors would like to thank the reviewer for their detailed comments which greatly improved this manuscript.*

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

References

Bauer, P. and Kieser, M. (1999). Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* **18**, 1833–1848.

Cohen, A. and Sackrowitz, H. (1989). Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* **8**, 273–278.

Marcus, R., Peritz, E., and Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63**, 655–660.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine* **22**, 953–959.

Posch, M., Koenig, F., Branson, M., Brannath, W., Dunger-Baldauf, C., and Bauer, P. (2005). Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* **24**, 3697–3714.

Putter, J. and Rubenstein, D. (1968). Technical report tr165: On estimating the mean of a selected population. *University of Wisconsin statistics department, Wisconsin*.

Sampson, A. and Sill, M. (2005). Drop-the-losers design: Normal case. *Biometrical Journal* **47**, 257–268.

Sill, M. and Sampson, A. (2007). Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. *Communications in Statistics-Theory and Methods* **36**, 801–813.

Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: Applications and practical considerations. *Biometrical Journal* **48**, 635–643.

Shen, L. (2001). An improved method of evaluating drug effect in a multiple dose clinical trial. *Statistics in Medicine* **20**, 1913–1929.

Stallard, N. and Todd, S. (2003). Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* **22**, 689–703.

Stallard, N. and Todd, S. (2005). Point estimates and confidence intervals for sequential trials involving selection. *Journal of Planning and Statistical Inference* **135**, 402–419.

Stallard, N., Todd, S., and Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Planning and Statistical Inference* **138**, 1629–1638.

Thall, P., Simon, R., and Ellenberg, S. (1988). Two-stage selection and testing designs for comparitive clinical trials. *Biometrika* **75**, 303–310.

Todd, S., Whitehead, J., and Facey, K. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika* **83**, 453–461.

Research Article

Statistics in Medicine

Received 7 January 2013, Accepted 23 June 2013 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5920

Empirical Bayes estimation of the selected treatment mean for two-stage drop-the-loser trials: a meta-analytic approach

Jack Bowden,[a*†] Werner Brannath[b] and Ekkehard Glimm[c]

Point estimation for the selected treatment in a two-stage drop-the-loser trial is not straightforward because a substantial bias can be induced in the standard maximum likelihood estimate (MLE) through the first stage selection process. Research has generally focused on alternative estimation strategies that apply a bias correction to the MLE; however, such estimators can have a large mean squared error. Carreras and Brannath (*Stat. Med.* 32:1677-90) have recently proposed using a special form of shrinkage estimation in this context. Given certain assumptions, their estimator is shown to dominate the MLE in terms of mean squared error loss, which provides a very powerful argument for its use in practice. In this paper, we suggest the use of a more general form of shrinkage estimation in drop-the-loser trials that has parallels with model fitting in the area of meta-analysis. Several estimators are identified and are shown to perform favourably to Carreras and Brannath's original estimator and the MLE. However, they necessitate either explicit estimation of an additional parameter measuring the heterogeneity between treatment effects or a quite unnatural prior distribution for the treatment effects that can only be specified after the first stage data has been observed. Shrinkage methods are a powerful tool for accurately quantifying treatment effects in multi-arm clinical trials, and further research is needed to understand how to maximise their utility. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: drop-the-loser trials; empirical Bayes estimation; meta-analysis; temporal coherency

1. Introduction

Two-stage drop-the-loser designs provide a framework for picking the most effective treatment out of a larger group of candidates and then testing it against a standard therapy in a confirmatory analysis. Although this design is an efficient way to discover effective treatments, the selection mechanism acts to inflate the type I error of the final test statistic [1, 2] and can also induce a substantial bias into the standard maximum likelihood estimate (MLE). With regard to the former, current regulatory authority guidance (e.g. [3]) is unequivocal that the final analysis must control the type I error rate. With regard to the latter, whilst acknowledging that estimation bias is a serious issue affecting the validity of adaptive trials and that bias should be 'minimised', there is a distinct lack of guidance and consensus on how this should be achieved. Research has generally focused on estimators that apply a bias correction to the MLE. One such class of estimators, referred to as uniform minimum variance conditionally unbiased estimators (UMVCUEs), totally removes the MLE's bias [4–7]. Others have proposed iterative or likelihood-based methods that can substantially reduce the bias of the MLE, without being unbiased [7–9]. Unfortunately, methods that explicitly target bias correction generally lead to an estimator with a mean squared error (MSE) larger than that of the MLE.

[a]*MRC Biostatistics Unit Hub for Trials Methodology Research, Cambridge, U.K.*
[b]*Competence Center for Clinical Trials Bremen, Faculty 3, University of Bremen, Bremen, Germany*
[c]*Novartis Pharma AG, CH-4002 Basel, Switzerland*
**Correspondence to: Jack Bowden, MRC Biostatistics Unit, IPH, Robinson Way, Cambridge CB2 0SR, U.K.*
[†]*E-mail: jack.bowden@mrc-bsu.ac.uk*

Carreras and Brannath [10] have recently proposed the use of shrinkage estimation [11] within the context of a two-stage drop-the-loser trial. In their method, the stage 1 data on all treatments are used to define a shrinkage estimate for the selected treatment at stage 1. This is then combined with the stage 2 estimate for the selected treatment via a weighted average. Under the assumption that the true treatment effects are independent and follow a common normal distribution (with any mean and variance), their estimator is shown to dominate the MLE in terms of MSE, thus providing a very powerful argument for its use in practice. In this paper, we propose an alternative shrinkage estimation strategy for drop-the-loser designs. Our approach is, in some ways, a simpler estimation procedure to that of [10] because it uses all of the available data within a single, standard, shrinkage equation. However, this apparent simplicity does impose some additional complications, which are discussed at length herein.

In Section 2, we introduce our notation for the two-stage drop-the-loser design. In Section 3, we describe the general principle of shrinkage estimation, Carreras and Brannath's original application of shrinkage estimation to the drop-the-loser trial context, and also present our alternative approach. In Section 4, we introduce several shrinkage estimators that naturally flow from our alternative formulation, and in Section 5, we evaluate the performance of the existing and alternative shrinkage estimators for various two-stage drop-the-loser design scenarios. We conclude in Section 6 with a discussion of the issues raised and point to further avenues of research.

2. The two-stage drop-the-loser design

Let $X_i \sim N\left(\mu_i, \sigma_1^2\right)$, $i = 1, \ldots, k$, be the effect estimates (MLEs) of k experimental treatments $T_1, ..., T_k$ at the first stage of a two-stage trial. The common variance term, σ_1^2, is assumed to be known. Assuming that large values indicate the most benefit, the 'best' treatment, T_s, $s \in \{1, ...k\}$, is selected as the one with the top-ranking MLE. That is, $X_s = \text{Max}\{X_1, ..., X_k\}$. Treatment T_s is taken forward in isolation for testing on an independent population in stage 2. Let $Y_s \sim N\left(\mu_s, \sigma_2^2\right)$ be the estimate for μ_s at stage 2.

Let X_0 and Y_0 represent the normally distributed treatment effect estimates for the control group at stages 1 and 2, with mean μ_0 and variances σ_1^2 and σ_2^2, respectively. At the end of the trial, we are interested in estimating the contrast $\mu_s - \mu_0$. Because the control group always proceeds to the final stage, μ_0 is unbiasedly estimated by its MLE, and we therefore focus our attention on estimation of μ_s only. The MLE of μ_s at stage 2 and its (assumed) variance are given by

$$\hat{\mu}_s = \frac{\sigma_2^2 X_s + \sigma_1^2 Y_s}{\sigma_1^2 + \sigma_2^2}, \quad \text{Var}\left(\hat{\mu}_s\right) = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}. \tag{1}$$

Because it ignores the selection of X_s, $\hat{\mu}_s$ is positively biased (potentially seriously so), and $\text{Var}\left(\hat{\mu}_s\right)$ is also incorrect. We can express the most efficient unbiased estimate for μ_s as

$$\tilde{\mu}_s = \hat{\mu}_s - \frac{\sigma_2^2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \frac{\phi\left[\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1^2}\left(\hat{\mu}_s - X_r\right)\right]}{\Phi\left[\frac{\sqrt{\sigma_1^2+\sigma_2^2}}{\sigma_1^2}\left(\hat{\mu}_s - X_r\right)\right]}, \tag{2}$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal density and distribution functions, respectively, and where X_r is defined as the second best performing treatment at stage 2, that is, $X_r = \text{Max}\{X_1, ..., X_k\} / X_s$. This is referred to as the UMVCUE [4, 6].

2.1. Assessing estimators of μ_s

For any estimator of μ_s, μ_s^*, we can express its bias and MSE as

$$\begin{aligned} \text{Bias}\left(\mu_s^*\right) &= \sum_{i=1}^{k} \text{E}\left[\mu_s^* - \mu_i | T_s = T_i\right] P(T_s = T_i), \\ \text{MSE}\left(\mu_s^*\right) &= \sum_{i=1}^{k} \text{E}\left[\left(\mu_s^* - \mu_i\right)^2 | T_s = T_i\right] P(T_s = T_i). \end{aligned} \tag{3}$$

This form (from Posch [12]) makes clear that at the trial outset, μ_s is a random variable. By definition, Bias$(\tilde{\mu}_s) = 0$, and MSE$(\tilde{\mu}_s)$ is smaller than any other μ_s^* that is also unbiased. This of course does not mean that MSE$(\tilde{\mu}_s)$ is smaller than any biased estimator; for example, it is generally true that MSE$(\tilde{\mu}_s) \gg$ MSE$(\hat{\mu}_s)$ (see Section 5 for example). Furthermore, it is commonly agreed that MSE (a measure equivalent to its variance + squared bias) provides a far better summary of an estimator's worth than bias alone.

3. Shrinkage estimation

The standard motivation for using shrinkage methods is to provide simultaneous, accurate estimation for a group of parameters, where accuracy is defined via the combined MSE. For example, if we were interested in jointly estimating the true mean effect of all k treatments $\mu_1, ..., \mu_k$ using only stage 1 data, then it will generally be true that the combined MSE,

$$\sum_{i=1}^{k} \mathrm{E}\left[(\mu_i^* - \mu_i)^2\right], \tag{4}$$

is far smaller when μ_i^* equals $\breve{\mu}_i^L$ as opposed to the MLE X_i, where $\breve{\mu}_i^L$ is Lindley's estimator [13]:

$$\breve{\mu}_i^L = \hat{B}_+ X_i + (1 - \hat{B}_+)\bar{X} \tag{5}$$

and

$$\hat{B}_+ = \max\left\{0, 1 - \hat{C}\right\}, \quad \hat{C} = \frac{(k-3)\sigma_1^2}{\sum_{i=1}^{k}(X_i - \bar{X})^2}, \quad \bar{X} = \sum_{i=1}^{k} X_i/k. \tag{6}$$

Although shrinkage formula (5) was not originally proposed using a Bayesian argument, it can be easily understood and shown to be optimal within a Bayesian framework. Assume that *a priori* $\mu_1, ..., \mu_k$ are themselves independent and identically distributed (i.i.d) $N(\mu, \tau^2)$ random variables and only stage 1 data are available for the k treatments. Given μ_i, the distribution of its MLE X_i is $\hat{\mu}_i|\mu_i \sim N\left(\mu_i, \sigma_1^2\right)$. The *posterior* distribution of μ_i given $\hat{\mu}_i$ is then

$$\mu_i|\hat{\mu}_i \sim N\left(\frac{\tau^2}{\sigma_1^2 + \tau^2}\hat{\mu}_i + \frac{\sigma_1^2}{\sigma_1^2 + \tau^2}\mu, \frac{\sigma_1^2\tau^2}{\sigma_1^2 + \tau^2}\right). \tag{7}$$

$\breve{\mu}_i^L$ can therefore be viewed as an 'Empirical Bayes' estimate for the posterior mean of equation (7), with $\bar{X}$, X_i and $\hat{C}$ substituted for μ, $\hat{\mu}_i$ and $\frac{\sigma_1^2}{\sigma_1^2+\tau^2}$, respectively. $\bar{X}$ is clearly an unbiased estimate of μ, but it is perhaps less obvious that $\hat{C}$ is an unbiased estimate for $\frac{\sigma_1^2}{\sigma_1^2+\tau^2}$, regardless of the true value of τ^2. If $1 - \hat{C}$ gives a negative value, it is replaced by 0 in the definition of $\hat{B}_+$. This 'plus rule' has been shown to further reduce the MSE of the resulting estimate $\breve{\mu}_i^L$ [14].

3.1. Carreras and Brannath's approach

Hwang [15] explicitly considered estimation of a single mean parameter from a k component system, where all k components have normally distributed estimates with a common variance and the single component is identified by having the largest estimate. This is identical to estimating μ_s using only stage 1 data in a two-stage drop-the-losers trial. He proved that when the treatment means follow a $N(\mu, \tau^2)$ prior distribution (for any μ and $\tau^2 \geqslant 0$), so that their posterior distributions obey equation (7) and $\breve{\mu}_s^L$ is defined by equation (5) with $i = s$ (i.e. drop-the-losers selection), then the dominance result MSE$\left(\breve{\mu}_s^L\right) \leqslant$ MSE(X_s) holds. Within the context of a two-stage drop-the-loser trial, Carreras and Brannath use Hwang's result to show that their estimator for μ_s at stage two

$$\breve{\mu}_s^{CB} = t\breve{\mu}_s^L + (1-t)Y_s, \quad \text{for} \quad t = \frac{1/\sigma_1^2}{1/\sigma_1^2 + 1/\sigma_2^2} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{8}$$

analogously dominates $\hat{\mu}_s$ from equation (1). Their result relies on the fact that equation (1) is equivalent to replacing $\breve{\mu}_s^{CB}$ in equation (8) with X_s. This also occurs naturally when the shrinkage factor C is set to 0.

3.2. An alternative formulation

Carreras and Brannath's method for estimating μ_s following a two-stage drop-the-loser trial can itself be viewed as a two-stage approach. That is, only the stage 1 data are used in the standard shrinkage estimator $\breve{\mu}_s^L$, and then, stage 2 data on the selected treatment, Y_s, are added separately in equation (8) afterwards. Although $\breve{\mu}_s^{CB}$ has the nice dominance property over the MLE, it is useful to consider whether it can itself be improved upon. For example, why not use all of the stages 1 and 2 data to define a single shrinkage estimator? Although this sounds straightforward, it does impose some extra complications. Despite being only truly concerned with estimation of the top-ranking treatment's mean, μ_s, Carreras and Brannath's method is defined to find shrinkage estimates given k parameter estimates with a *common* variance. This means that we are assuming the posterior distribution for μ_i implied by (7) (ignoring the stage 2 data), which enables the use of $\breve{\mu}_s^L$ from (5). However, if we use all of the data (including the stage 2 data Y_s), we may assume the following set-up:

$$\begin{aligned}
&\mu_i \sim N\left(\mu, \tau^2\right),\\
&\hat{\mu}_i|\mu_i \sim N(\mu_i, W_i), \quad \text{where}\\
&\hat{\mu}_i = X_i, W_i = \sigma_1^2 \quad \text{if } i \neq s \quad \text{or} \quad \hat{\mu}_i = \frac{\sigma_2^2 X_i + \sigma_1^2 Y_i}{\sigma_1^2 + \sigma_2^2}, W_i = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \quad \text{if } i = s.
\end{aligned}$$

Further assuming that the μ_is are stochastically independent and the $\hat{\mu}_i$s are conditionally independent given μ_i, then

$$\mu_i|\hat{\mu}_i \sim N\left(\frac{\tau^2}{W_i + \tau^2}\hat{\mu}_i + \frac{W_i}{W_i + \tau^2}\mu, \frac{W_i\tau^2}{W_i + \tau^2}\right) \tag{9}$$

becomes the distribution with which to construct shrinkage estimates for the μ_is. The fact that W_i is not constant makes specification of single appropriate shrinkage factor C and estimate for the grand mean μ far less straightforward. In Section 4, we will discuss estimation for this new setting, pointing out its connection with model fitting in the area of meta-analysis.

4. Estimation for the new target

4.1. Direct estimation

Under the framework of equation (9), if the parameters μ and τ^2 were known, the best estimate for μ_s is given by

$$\frac{\tau^2}{W_s + \tau^2}\hat{\mu}_s + \frac{W_s}{W_s + \tau^2}\mu \tag{10}$$

Furthermore, with infinite data, it is clear that directly replacing μ and τ^2 in this expression with consistent estimates, $\hat{\mu}$ and $\hat{\tau}^2$, would give equation (10). If one assumes that τ^2 is known, then the variance of the posterior distribution in (9) also becomes completely known. It is then possible to show that the MLE for μ is

$$\hat{\mu}_{\tau^2} = \frac{\left(\sigma_1^2\sigma_2^2 + \tau^2\sigma_2^2\right) X_s + \left(\sigma_1^4 + \tau^2\sigma_1^2\right) Y_s + (k-1)q\bar{X}_{-s}}{kq + \sigma_1^4}, \tag{11}$$

where $q = \sigma_1^2\sigma_2^2 + \tau^2\left(\sigma_1^2 + \sigma_2^2\right)$ and $\bar{X}_{-s} = 1/(k-1)\sum_{i\neq s} X_i$. τ^2 could then be estimated by maximising the posterior likelihood of (9) with respect to τ^2 at $\mu = \hat{\mu}_{\tau^2}$. See the Appendix for further details. We will refer to the estimate for μ_s obtained by directly plugging in the previous estimates to equation (10) as $\hat{\mu}_s^{MPL}$. Although this approach is fairly crude, it will be interesting to observe the performance of $\hat{\mu}_s^{MPL}$ compared to several 'legitimate' shrinkage estimators that are now introduced.

4.2. Shrinkage estimation: Carter and Rolph's standard prior approach

Carter and Rolph [16] investigate shrinkage methods for jointly estimating the parameters of a k component system with unequal variances, as set up in equation (9). We will use it specifically to yield an estimate for μ_s in the context of a two-stage drop-the-loser trial of the form

$$\hat{B}_{+s}\hat{\mu}_s + \left(1 - \hat{B}_{+s}\right)\hat{\mu}, \quad \text{where} \quad \hat{B}_{+s} = \max\left\{0, 1 - \hat{C}_s\right\}, \tag{12}$$

so as to approximate equation (10). Although not immediately obvious, it can be shown that their suggested approach is equivalent the following procedure. First define a new weight $V_i = (W_i + \tau^2)$, and find the value of τ^2 that solves $Q(\tau^2) = k - 1$, where

$$Q(\tau^2) = \sum_{i=1}^{k} V_i^{-1}\left(\hat{\mu}_i - \hat{\mu}(\tau^2)\right)^2, \quad \text{for} \quad \hat{\mu}(\tau^2) = \frac{\sum_{i=1}^{k} V_i^{-1}\hat{\mu}_i}{\sum_{i=1}^{k} V_i^{-1}}. \tag{13}$$

If $Q(\tau^2) = k - 1$ for a $\tau^2 < 0$, then the estimate is truncated to 0. $Q(\tau^2)$ is known in the area of medical meta-analysis as the 'generalised' Q statistic [17, 18], and this estimation method for μ is known as the Paule–Mandel method of moments algorithm [19, 20]. Let the estimate for τ^2 derived in this manner be equal to $\hat{\tau}^2_{PM}$. The shrinkage factor for μ_s is then approximated by

$$\hat{C}_s = \frac{(k-3)W_s}{\left(\hat{\tau}^2_{PM} + \overline{W}\right)Q\left(\hat{\tau}^2_{PM}\right) + (k-3)(W_s - \overline{W})}, \quad \text{where} \quad \overline{W} = \sum_{i=1}^{k} W_i/k. \tag{14}$$

The form of $\hat{C}_s$ may appear complicated, but it has the nice property that if the W_is are all equal, then it reduces to the original $\hat{C}$ given in equation (6). In our context, the W_is can only approach equality as $\sigma_2^2 \to \infty$ or as the stage 2 sample size tends to zero. $\hat{C}_s$ can be inserted into equation (12) along with the MLE $\hat{\mu}_s$ and grand mean $\hat{\mu}$ given by $\hat{\mu}\left(\hat{\tau}^2_{PM}\right)$, to yield a new shrinkage estimator for μ_s. We will refer to this estimator as $\breve{\mu}_s^{\tau^2}$ – the 'τ^2' denoting that this parameter is additionally and explicitly estimated.

4.3. Shrinkage estimation: Carter and Rolph's proportional prior approach

Carter and Rolph [16] also propose an alternative method for applying shrinkage estimation within the unequal variance context, which usefully avoids an iterative estimation of τ^2. It relies on the assumption of a different prior distribution for the treatment parameters, namely $\mu_i \sim N(\mu, W_i\tau^2)$. This asserts that the prior uncertainty around each treatment's mean is directly proportional to the variance of its estimate, $\hat{\mu}_i$. By replacing τ^2 with $W_i\tau^2$ in (9), it is clear that this implies the posterior distribution:

$$\mu_i|\hat{\mu}_i \sim N\left(\frac{\tau^2}{1+\tau^2}\hat{\mu}_i + \frac{1}{1+\tau^2}\mu, \frac{W_i\tau^2}{(1+\tau^2)}\right), \tag{15}$$

the mean of which becomes an alternative target to estimate. Because this mean does not depend on W_i, it suffices to calculate a single shrinkage factor $\hat{C}$ using all of the data. This can then, in conjunction with an estimate for μ, be used to estimate μ_s via equation (12). Turning first to estimation of μ: Under the proportional prior, the unconditional distribution of the estimates is $\hat{\mu}_i \sim N\left(\mu, \sigma_i^2(1+\tau^2)\right)$. Therefore, given weights $V_i^* = \sigma_i^2(1+\tau^2)$, the inverse variance weighted average

$$\hat{\mu}(\tau^2) = \frac{\sum_{i=1}^{k} V_i^{-1*}\hat{\mu}_i}{\sum_{i=1}^{k} V_i^{-1*}} = \hat{\mu}(0) \quad \forall\tau^2.$$

$\hat{\mu}(0)$ is referred to in meta-analysis as the 'fixed-effects' estimate for μ, as opposed to a 'random-effects' estimate, of which $\hat{\mu}\left(\hat{\tau}^2_{PM}\right)$ is an example. Of course, when τ^2 is estimated to be 0, they are equal.

Turning now to estimation of μ_s via $\hat{C}$: It can easily be shown that Carter and Rolph's approach in this context is equivalent to choosing:

$$\hat{C} = \hat{C}(0) = \frac{k-3}{Q(0)}, \quad \text{where} \quad Q(0) = Q(\tau^2 = 0). \tag{16}$$

$\hat{C}(0)$ can be seen as a simple generalisation of the $\hat{C}$ in equation (6) for the case where the W_i terms are not constant. $Q(0)$ is known as Cochrane's heterogeneity statistic in meta-analysis and is closely related to the DerSimonian and Laird estimator for τ^2, τ^2_{DL} [20]. For example, when $Q(0) > k - 1$,

$$\frac{Q(0) - (k-1)}{Q(0)} = \frac{\tau^2_{DL}}{\tau^2_{DL} + \bar{\sigma}^2} = I^2$$

where $\bar{\sigma}^2$ is called the 'typical' within study variance and I^2 is a popular measure of 'inconsistency' (heterogeneity) among studies in a meta-analysis [21]. We will refer to the resulting estimator (which utilises $\hat{C}(0)$ and $\hat{\mu}(0)$) as $\breve{\mu}_s^0$.

4.4. *Incorporating 'Limiting Translation'*

It is well known that shrinkage methods can perform poorly with respect to specific parameter components of a larger system, when the magnitude of the specific parameters are among the most extreme. For this reason, Efron and Morris [22] suggest the use of a 'Limiting Translation' (LT) strategy that constrains one shrinkage estimator to be within a certain distance of the MLE. Efron [14] and Johnson [23] suggest a practical choice for this constraint of one unit of the MLE's standard error. The effect of applying LT to a single extreme parameter component of a larger system, say a μ_i from $\mu_1, ..., \mu_k$, is to dramatically reduce the i'th contribution to the overall MSE of equation (4), at the expense of increasing the total value of equation (4) by a small margin.

Although from equation (3), we can see that, at the trial outset, μ_s is not a single parameter but rather a weighted mixture of all k fixed parameter values $\mu_1, ..., \mu_k$, the specific values of those parameters may mean that μ_s is consistently an outlier. For example, this would certainly be the case if one treatment was far more effective than any other because it would monopolise the value of μ_s. We can apply LT to the shrinkage estimator $\breve{\mu}_s^0$ by subtracting $\hat{\mu}_s$ from equation (12), constraining the result to be $\leqslant \sqrt{W_s}$ and noting that the definition of $\hat{C}(0)$ in equation (16) becomes

$$\hat{C}(0) = \min\left\{\frac{(k-3)}{Q(0)}, \frac{\sqrt{W_s}}{|\hat{\mu}(0) - \hat{\mu}_s|}\right\}. \tag{17}$$

This estimator will be referred to as $\breve{\mu}_s^0(LT)$. LT versions of all other estimators are clearly possible but are not considered here.

4.5. *Some implications of using* $\breve{\mu}_s^0$

When deriving the form of $\breve{\mu}_s^0$, there is no inherent mathematical difficulty in assuming an $N(\mu, W_i\tau^2)$ prior distribution (with varying W_is) for the μ_is because the resulting posterior distribution for $\mu_i|\hat{\mu}_i$ remains in the normal family. However, this shrinkage approach does raise certain philosophical questions when applied in the context of a two-stage drop-the-loser trial. The primary issue is that we do not know *a priori* which treatment will be selected. So, assigning the $N(\mu, W_s\tau^2)$ prior to μ_s (for general values of μ and τ^2) is only possible after we have observed the first stage data. This violates the principle of 'Temporal Coherency' [24] that states that the prior must be specified in advance and constant in time. Indeed, this principle is overwhelmingly adhered to by practitioners of Bayesian inference in the interest of maintaining scientific objectivity. A consequence of this temporal violation, which becomes most apparent in Section 5, is that there is no general way to simulate data consistent with the assumptions of $\breve{\mu}_s^0$. To understand this, suppose we wanted to generate trial data consistent with the shrinkage estimator $\breve{\mu}_s^{\tau^2}$ instead. We simply start by simulating the μ_is from an $N(\mu, \tau^2)$ density given values for μ and τ^2, which can then be used to generate the trial data for stages one and two $(X_1, ..., X_k, Y_s)$. These data can then be used to specify the distributions $\hat{\mu}_i|\mu_i$ and $\mu_i|\hat{\mu}_i$ from equation (9). Clearly, we can not follow an equivalent data-generating procedure when the μ_is come from an $N(\mu, W_i\tau^2)$ prior density because, as previously stated, the prior can only be specified *after* seeing the data. The single exception is when τ^2=0 (implying a degenerate normal prior) in which case the μ_is all take the value μ with probability 1. This is equivalent to assuming a fixed effects model with only one unknown parameter, μ.

Of course, despite these philosophical concerns, we are still free and able to evaluate $\breve{\mu}_s^0$ in a simulation study without exactly mimicking the data-generating process it relies upon.

5. Simulation study

We simulate trial data under a two-stage drop-the-loser design in order to quantify the bias and MSE of four new estimators $\left(\hat{\mu}_s^{MPL}, \breve{\mu}_s^{\tau^2}, \breve{\mu}_s^0 \text{ and } \breve{\mu}_s^0(LT)\right)$ for μ_s, alongside the existing estimators $\left(\breve{\mu}_s^{CB}, \hat{\mu}_s, \tilde{\mu}_s\right)$. We use the definition of bias and MSE from equation (3), which can be simply and accurately approximated by averaging over all simulations where, in each single case, a treatment T_i out of k is ranked top at the end of stage 1 so that $\mu_s = \mu_i$. We chose four different levels of standard error

associated with the stage 1 and 2 estimates, along with four different scenarios for the six unknown means:

- Scenario *I*. True means $\sim N(0, 1)$;
- Scenario *II*: True means all 0;
- Scenario *III*: One true mean $= 1$, 5 means equal to 0;
- Scenario *IV*: One true mean $= 1.5$, 5 means equal to 0.

The underlying distribution of the treatment parameters is a key factor driving the Bayesian motivation of any shrinkage estimator. Scenario *I* is compatible with the prior assumptions of $\check{\mu}_s^{\tau^2}$ and $\check{\mu}_s^{CB}$. Scenario *II* is compatible with the prior assumptions of all shrinkage estimators, despite being non-stochastic, because it is equivalent to scenario *I* with $\tau^2 = 0$. Carreras and Brannath's dominance result for $\check{\mu}_s^{CB}$ is valid for scenarios *I* and *II* . Scenarios *III* and *IV* are not compatible with any shrinkage estimator. However, all scenarios are compatible with the assumptions of the UMVCUE, in the sense that it maintains its unbiasedness for any constellation of parameter values.

We show the results for the 16 simulation scenarios in Table I. All reported figures are based on 50 000 simulations. For each simulation scenario, we show the bias and $\sqrt{MSE}$ in units of the MLE $\hat{\mu}_s$s naive standard error, $\sqrt{W_s}$, to make comparisons easier. We do not show the bias of $\tilde{\mu}$ because it is always zero, except for sampling error. All of the shrinkage estimators generally outperform the MLE in terms of bias and $\sqrt{MSE}$, the exception being simulation 15, scenario *IV*. Carreras and Brannath [10] show theoretically that the MLE is maximally biased when all treatment means are equal. The results of scenario *II* supports this. $\check{\mu}_s^{CB}$ and $\check{\mu}_s^{\tau^2}$s performances are fairly equal. $\check{\mu}_s^{CB}$ tends to have a smaller bias than $\check{\mu}_s^{\tau^2}$ but a larger $\sqrt{MSE}$. The performance of $\hat{\mu}_s^{MPL}$ varies considerably; it is the best estimator

Table I. Bias and mean squared error (MSE) of the various estimands over the 16 scenarios of a two-stage drop-the-loser trial with $k = 6$ initial treatments.

σ_1, σ_2	Bias (μ_s^*)						$\sqrt{MSE\,(\mu_s^*)}$						
values	$\hat{\mu}_s$	$\check{\mu}_s^{CB}$	$\check{\mu}_s^{\tau^2}$	$\check{\mu}_s^0$	$\check{\mu}_s^0(LT)$	$\hat{\mu}_s^{MPL}$	$\tilde{\mu}_s$	$\hat{\mu}_s$	$\check{\mu}_s^{CB}$	$\check{\mu}_s^{\tau^2}$	$\check{\mu}_s^0$	$\check{\mu}_s^0(LT)$	$\hat{\mu}_s^{MPL}$
Scenario *I*: true means $\sim N(0, 1)$													
1.(1,1)	0.63	0.19	0.22	0.11	0.11	−0.17	1.21	1.12	0.97	0.96	0.95	0.94	0.97
2.(2,1)	0.51	0.18	0.29	0.11	0.11	−0.03	1.08	1.08	0.98	0.97	0.92	0.92	0.91
3.$\left(\frac{1}{2}, 1\right)$	0.51	0.13	0.14	0.11	0.11	−0.22	1.35	1.08	0.99	0.99	0.98	0.98	1.04
4.$\left(1, \frac{1}{2}\right)$	0.40	0.12	0.17	0.00	0.00	−0.14	1.07	1.05	0.99	0.98	0.98	0.98	1.01
Scenario *II*: true means all 0													
5.(1,1)	0.89	0.35	0.45	0.35	0.36	0.16	1.27	1.23	0.92	0.87	0.79	0.79	0.65
6.(2,1)	0.57	0.22	0.39	0.22	0.23	0.12	1.09	1.10	0.97	0.95	0.83	0.83	0.78
7.$\left(\frac{1}{2}, 1\right)$	1.14	0.45	0.48	0.45	0.47	0.17	1.64	1.35	0.86	0.84	0.81	0.82	0.58
8.$\left(1, \frac{1}{2}\right)$	0.57	0.22	0.39	0.23	0.23	0.12	1.08	1.09	0.96	0.94	0.83	0.83	0.77
Scenario *III*: one true mean $= 1$, $5 = 0$													
9.(1,1)	0.78	0.25	0.32	0.21	0.22	−0.03	1.24	1.19	0.94	0.93	0.88	0.88	0.84
10.(2,1)	0.55	0.20	0.36	0.19	0.19	0.08	1.08	1.09	0.97	0.95	0.85	0.85	0.81
11.$\left(\frac{1}{2}, 1\right)$	0.59	−0.07	−0.06	−0.10	−0.09	−0.56	1.40	1.14	1.04	1.05	1.05	1.04	1.20
12.$\left(1, \frac{1}{2}\right)$	0.50	0.16	0.28	0.11	0.11	−0.02	1.08	1.08	0.98	0.97	0.93	0.93	0.93
Scenario *IV*: one true mean $= 1.5$, $5 = 0$													
13.(1,1)	0.64	0.11	0.16	0.05	0.06	−0.24	1.21	1.14	0.98	0.99	0.98	0.98	1.02
14.(2,1)	0.53	0.19	0.33	0.16	0.16	0.04	1.08	1.08	0.97	0.96	0.88	0.88	0.86
15.$\left(\frac{1}{2}, 1\right)$	0.21	−0.40	−0.39	−0.43	−0.43	−0.94	1.17	1.04	1.16	1.17	1.19	1.18	1.49
16.$\left(1, \frac{1}{2}\right)$	0.40	0.07	0.16	−0.01	−0.01	−0.16	1.07	1.05	0.99	0.99	1.01	1.01	1.04

in terms of $\sqrt{MSE}$ in 8 out of 16 simulations but is sometimes the worst estimator by far (e.g. simulations 11 and 15). Unlike the other shrinkage estimators, it is also negatively biased in general. $\check{\mu}_s^0$ and $\check{\mu}_s^0(LT)$ perform similarly and are consistently the most reliable estimators across the 16 scenarios. It is perhaps surprising that their similarity extends to scenarios *III* and *IV*, where one would suspect the LT strategy would come into play. This implies that the difference between $\hat{\mu}_s$ and $\check{\mu}_s^0$ is still almost always less than $\sqrt{W_s}$.

Figures 1–3 show the results of three further simulation studies. In each case, $\sigma_1 = \sigma_2 = 1$. Figure 1 shows the scaled bias and $\sqrt{MSE}$ of the estimators for a trial with $k = 6$ treatments, 5 of which have true mean 0 and one of which has true mean δ, as δ is varied between 0 and 5. In order to highlight the strength of the selection effect as a function of δ, we also plot the average value of μ_s (labelled as '$E[\mu_s]$'). One can see that the bias of the shrinkage estimators changes sign as δ increases whereas the bias of the MLE decreases from 0.8 to 0 as δ increases. Of the shrinkage estimators, $\check{\mu}_s^0$ and $\check{\mu}_s^0(LT)$ have the smallest bias and MSE (and are indistinguishable) for δ up to 1.8. For $\delta \leqslant 2.2$, the shrinkage estimators dominate the MLE in terms of $\sqrt{MSE}$. $\hat{\mu}_s^{MPL}$ has the smallest bias of all for $\delta \leqslant 1.5$ but, by far, has the largest (negative) bias as δ increases. $\hat{\mu}_s^{MPL}$ also has the smallest $\sqrt{MSE}$ of all estimators for small δ, but as δ increases, its MSE increases dramatically.

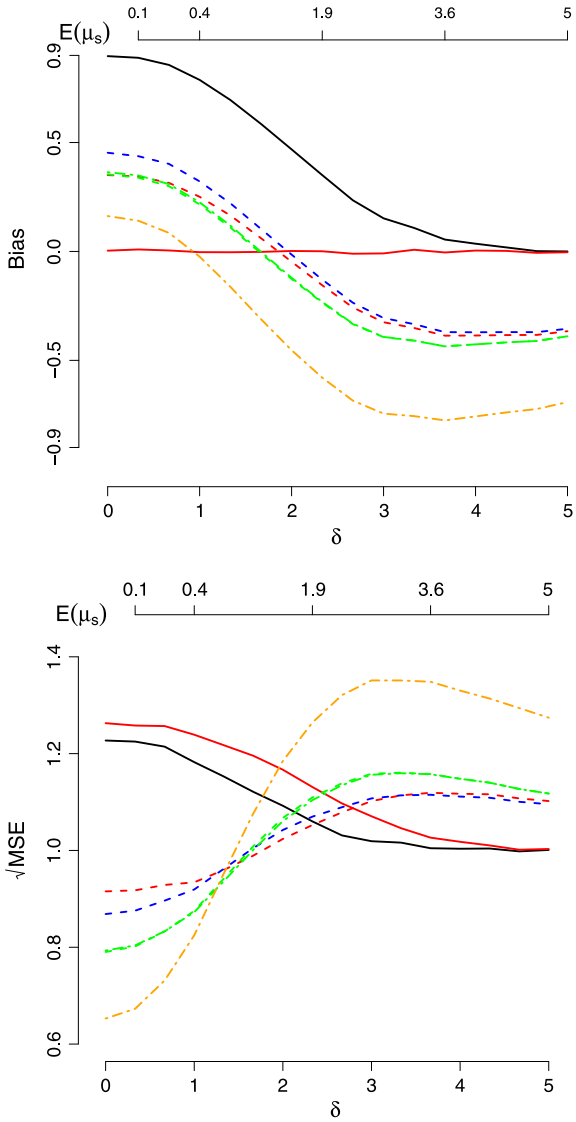


Figure 1. Bias and mean squared error (MSE) of the estimators as a function of δ. Key: maximum likelihood estimate (black), UMVCUE (red), $\check{\mu}_s^{CB}$ (red-dashed), $\check{\mu}_s^{\tau^2}$ (blue-dashed), $\check{\mu}_s^0$ (green dashed), $\check{\mu}_s^0(LT)$ (green dot-dashed) and $\hat{\mu}_s^{MPL}$ (orange dot-dashed).

Statistics in Medicine

Figure 2 shows the results of a simulation for $k = 6$ treatments with mean parameters drawn from a $N(0, \tau^2)$ distribution as τ^2 is varied between 0 and 4 (the choice of $\mu = 0$ is clearly unimportant). The bias of all shrinkage estimators decreases towards 0 as τ^2 increases, although this happens most rapidly for $\check{\mu}_s^0$ and $\check{\mu}_s^0(LT)$ so that they are the least biased. For values of τ^2 greater than 1, $\hat{\mu}_s^{MPL}$ is consistently negatively biased. The $\sqrt{MSE}$ of the shrinkage estimators appears to asymptote upwards towards $\sqrt{W_s}$ (towards 1 after scaling) as τ^2 increases, but remain below that of the MLE in this range.

In an effort to separate $\check{\mu}_s^0$ and $\check{\mu}_s^0(LT)$, Figure 3 shows the results of a simulation assuming that the treatment mean parameters are drawn from a $N(0, 1)$ distribution, but the number of treatments, k, is varied between 5 and 20. As k increases, the positive bias exhibited by $\check{\mu}_s^0$ decreases, quickly becoming large and negative. $\check{\mu}_s^{CB}$ and $\check{\mu}_s^{\tau^2}$ do not to suffer in the same way; their biases asymptote towards 0 as k increases. $\check{\mu}_s^0(LT)$ appears to protect $\check{\mu}_s^0$ well from its tendency for negative bias beyond $k = 10$. From the right-hand panel, one can see that the price $\check{\mu}_s^0(LT)$ pays for this bias protection is an increase in $\sqrt{MSE}$. Interestingly, as k increases beyond 15, even the UMVCUE has a smaller $\sqrt{MSE}$ than the MLE.

5.1. *Summary of findings*

Across all simulations, the performance of $\check{\mu}_s^{\tau^2}$ is most similar to Carreras and Brannath's original estimator $\check{\mu}_s^{CB}$, but the two estimators that performed the best were $\check{\mu}_s^0$ and $\hat{\mu}_s^{MPL}$. However, the reasons for the latter's apparent success are now qualified. Estimation of the heterogeneity parameter τ^2 is

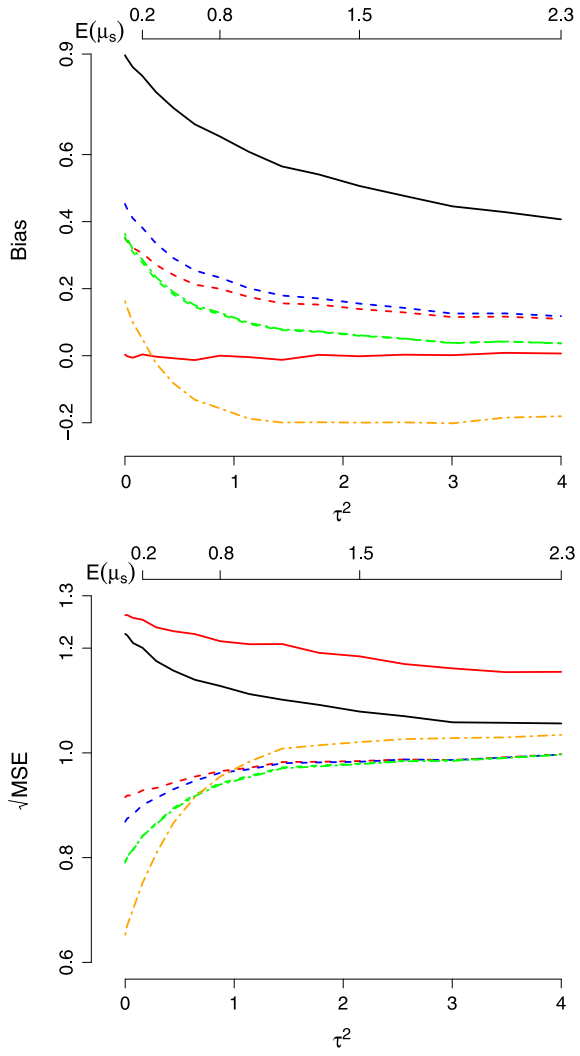


Figure 2. Bias and mean squared error (MSE) of the estimators as a function of τ^2. Key: Same as Figure 1.

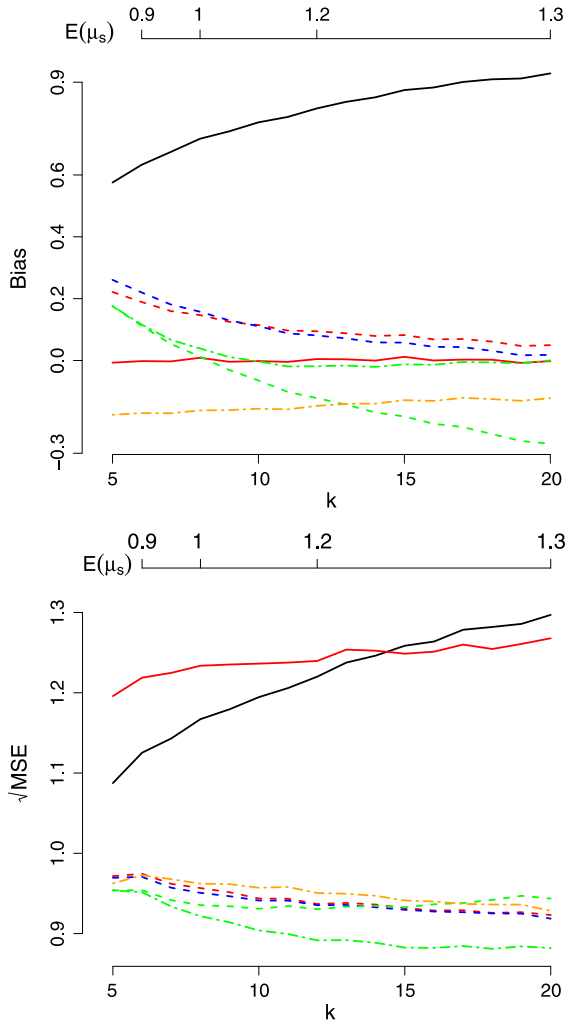


Figure 3. Bias and mean squared error (MSE) of the estimators as a function of k. Key: Same as Figure 1.

challenging when its true value is small or the number of treatments is small. When k is small or there is little apparent variation between treatment effect estimates $(\hat{\mu}_1, ..., \hat{\mu}_k)$, the estimate for τ^2 is often likely to be at or close to zero, even when its true value is larger. This fact is well documented in the meta-analysis literature, where quantification of the (between trial) heterogeneity parameter is often not recommended unless the number of studies is sufficiently large (at least 10). Although poor estimation of τ^2 impacts $\check{\mu}_s^{\tau^2}$ and $\hat{\mu}_s^{MPL}$, the latter is most strongly affected because from equation (10) when $\tau^2 = 0$, it reduces via equation (11) to the fixed effects estimate for μ, $\hat{\mu}(0)$. However, when the Paule–Mandel estimate for τ^2 is zero, $\check{\mu}_s^{\tau^2}$ does not shrink to exactly $\hat{\mu}(0)$ because $\hat{C}_s$ in equation (14) is not equal to 0. This explains why $\hat{\mu}_s^{MPL}$ performs so well in Scenario *II*, Table I, and for small values of δ, τ^2, k in Figures 1–3 because in these situations, the true value of μ_s is always close (or equal) to μ. Conversely, it also explains why it performs badly when μ_s is truly very different from μ (e.g. Scenario *IV*, Table I, and large values of δ, τ^2, k in Figures 1–3).

The LT version of $\check{\mu}_s^0$ only helped to improve its performance in simulations when the number of treatments rose above 10, which is unrealistically large for most clinical trials settings. It therefore appears to be an unnecessary extension in this context. However, it could potentially be implemented in a more sophisticated manner than we have here. For example, the width of the protection region around the MLE can be tuned to control its bias and MSE, rather than being fixed at a specific value as we did. Efron and Morris [22] provide the theoretical framework for doing this in a general shrinkage estimation context, but their method would need to be altered before application to drop-the-loser designs, in order to account for the selection mechanism. This is a topic for further research. One could also argue that by shrinking the prior variance for μ_s after selection, $\check{\mu}_s^0$ already contains and in-built form of LT.

6. Discussion

In this paper, we have reviewed the shrinkage estimation strategy of Carreras and Brannath [10] for two-stage drop-the-loser designs. As an extension, we propose that rather than using only the first stage data, the stage two data should also be used to furnish a single shrinkage equation. Although this strategy replaces two formulae with one, evaluation of this single shrinkage formula is much harder because the variance of its k estimated components are no longer equal. By incorporating the methods of Carter and Rolph [16] and Efron and Morris [22], we identified several alternative procedures. The alternative approaches necessitate either explicit estimation of an additional between treatment arm heterogeneity parameter τ^2 (for $\hat{\mu}_s^{MPL}$ and $\breve{\mu}_s^{\tau^2}$) or a different (and quite unnatural) prior distribution for the mean treatment effects (for $\breve{\mu}_s^0$). The new methods tend to outperform Carreras and Brannath's original estimator, but unfortunately, no equivalent dominance results could be shown.

From looking at the simulations in totality, the estimator that consistently performs well when μ_s is close to and far from the overall mean of all treatments is $\breve{\mu}_s^0$. Some may object philosophically to its use in this context because of concerns over Temporal Coherency. However, as Cox [25] states: There may be certain situations where it is perfectly right to modify one's prior beliefs as more data become available. Furthermore, when one's prior uncertainty about a parameter *is* allowed to change over time in a manner proportional to the (increasing) size of the data sample, then the resulting Bayesian inference starts to approximate a classical significance test [26]. This is exactly what occurs in the drop-the-loser context when, at the point of selection, the variance of the prior for μ_s shrinks from $\sigma_1^2\tau^2$ to $W_s\tau^2$ – for example, by a factor of 2 when $\sigma_1^2 = \sigma_2^2$. It is therefore pertinent to note that the formula for $\breve{\mu}_s^0$ can also be arrived at by applying Lindley's original equal variance shrinkage formula (5) to the standardised MLEs, $\hat{\mu}_i/\sqrt{W_i}$, [16] because they are sufficient test statistics for the null hypothesis $\mu = 0$.

We chose to illustrate the different estimation approaches for trials involving five or more treatments. Apart from $\hat{\mu}_s^{MPL}$, all of the shrinkage estimators discussed in this paper are only defined for $k \geqslant 4$, as indicated by the factors of $(k-3)$ they contain. However, we can crudely apply them for the $k = 3$ case by simply replacing these terms with $(k-2)$ instead – as performed by Carreras and Brannath [10]. We repeated the simulations shown in Figures (1) and (2) for $k = 3$ using this crude fix to see how it affected the performance of the various estimators. The results (not shown) were qualitatively very similar.

Throughout this paper, we have attempted to stress the link between shrinkage estimation in the adaptive trial context with that of meta-analysis. We have shown that $\breve{\mu}_s^0$, which incorporates the fixed effects estimate $\hat{\mu}(0)$, works well as an estimator for μ_s under drop-the-losers selection. It is therefore interesting to note the following: In meta-analyses that exhibit substantial amounts of between study heterogeneity, the random effects estimate for μ is known to be unreliable when the heterogeneity is thought to be driven by dissemination bias (i.e. selective reporting and publication of extreme findings) [18, 27]. In order to address this, it has been advocated that the fixed effects estimate $\hat{\mu}(0)$ be used as the preferred measure of overall effect instead [28, 29]. Thus, despite the fact that in the adaptive trial setting, the parameter of interest is $\hat{\mu}_s$ and in meta-analysis, it is the overall grand mean μ, when the data are affected by some form of selection, Carter and Rolph's proportional prior approach appears to be an effective solution to both problems.

Bowden and Glimm [30] have extended the idea of a two-stage drop-loser-trial to allow the best performing treatment to be identified over multiple stages. The motivation for adding further stages of selection is that one can markedly increase the probability of selecting the truly best treatment (and subsequently declaring it effective in a confirmatory analysis), whilst keeping trial costs to a minimum. Many other multi-arm multi-stage (MAMS) designs incorporating treatment selection rules have also been proposed with a similar motivation in mind, see for example [31, 32]. Because they ignore the selection process altogether and use all of the data to define a target posterior distribution or shrinkage equation, $\hat{\mu}_s^{MPL}$, $\breve{\mu}_s^{\tau^2}$ and $\breve{\mu}_s^0$ should be simple to apply in any of these contexts. It is not so obvious to see how Carreras and Brannath's original estimator, $\breve{\mu}_s^{CB}$, would generalise to the multi-stage context or if the dominance results that make it attractive in the two-stage case would remain in intact.

A simple, straightforward translation of the shrinkage estimators proposed here to other multi-arm trial designs is only immediate if the k treatment effect estimates are independent before selection. If, as in the MAMS design of Royston *et al*. [33], the treatment effect summarised time-to-event data in the form of a log-hazard ratio, then the effect estimates would be intrinsically correlated across treatment arms because of their shared control group data. Extending shrinkage estimation to account for inter-dependence of this sort is another topic for further research.

Appendix A: Details for the calculation of $\hat{\mu}_s^{MPL}$

Assume without loss of generality that $s = 1$, so that μ_1 is the mean of the best performing treatment. Let $\mathbf{Z} = (X_1, Y_1, X_2, ..., X_k,)$ equal the complete vector of data from the two-stage trial. Under the framework of equation (9), the model induces the unconditional distribution $\mathbf{Z} \sim \mathrm{N}(\mu \mathbf{1}_{k+1}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_2 & \mathbf{0} \\ \mathbf{0} & \left(\sigma_1^2 + \tau^2\right) \mathbf{I}_{k-1} \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} \sigma_1^2 + \tau^2 & \tau^2 \\ \tau^2 & \sigma_2^2 + \tau^2 \end{pmatrix}$$

Assuming that μ and τ^2 are known, then the best linear prediction of $\mu_1|\mu$ is given in equation (10), which is equivalent to

$$\frac{\sigma_1^2\sigma_1^2\mu + \sigma_2^2\tau^2 X_1 + \sigma_1^2\tau^2 Y_1}{\sigma_1^2\sigma_2^2 + \sigma_1^2\tau^2 + \sigma_2^2\tau^2}$$

The empirical question is how to estimate μ and τ^2 as accurately as possible. If we were to assume that τ^2 is known, then $\boldsymbol{\Sigma}$ becomes a completely known matrix. Then, it is easily seen from differentiating $\boldsymbol{a}'\boldsymbol{\Sigma}\mathbf{a} + \lambda \boldsymbol{a}'\mathbf{1}_{k+1}$ with respect to $\mathbf{a}$ and λ that the linear combination $\boldsymbol{a}'\boldsymbol{Z}$, which minimises $\mathrm{Var}(\boldsymbol{a}'\boldsymbol{Z}) = \boldsymbol{a}'\boldsymbol{\Sigma}\mathbf{a}$, is given by

$$\boldsymbol{a} = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{1}_{k+1}}{\mathbf{1}'_{\boldsymbol{k+1}}\boldsymbol{\Sigma}^{-1}\mathbf{1}_{k+1}}$$

As this also maximises the multivariate normal likelihood, $\boldsymbol{a}'\boldsymbol{Z}$ is also equivalent to the MLE for μ. Because Σ is diagonal, its inverse is trivially

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_2^{-1} & \mathbf{0} \\ \mathbf{0} & \left(\sigma_1^2 + \tau^2\right)^{-1} \mathbf{I}_{k-1} \end{pmatrix}$$

with

$$\boldsymbol{\Sigma}_2^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 + \sigma_1^2\tau^2 + \sigma_2^2\tau^2} \begin{pmatrix} \sigma_2^2 + \tau^2 & -\tau^2 \\ -\tau^2 & \sigma_1^2 + \tau^2 \end{pmatrix}$$

Hence, we obtain the estimate for $\hat{\mu}_{\tau^2}$ given in equation (11). We then plug $\hat{\mu}_{\tau^2}$ into the log-likelihood for τ^2 given μ:

$$\begin{aligned} ll(\tau^2) &= -\frac{1}{2}\left(\log Det(\boldsymbol{\Sigma}) + (\mathbf{Z} - \hat{\mu}_{\tau^2})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \hat{\mu}_{\tau^2}) + (\mathbf{Z} - \hat{\mu}_{\tau^2})\right) \\ &= \frac{1}{2}\left((k-1)\log\left(\sigma_1^2 + \tau^2\right) + \log\left(\sigma_1^2\sigma_2^2 + \sigma_1^2\tau^2 + \sigma_2^2\tau^2\right)\right. \\ &\quad \left. + (\mathbf{Z} - \hat{\mu}_{\tau^2})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z} - \hat{\mu}_{\tau^2}) + (\mathbf{Z} - \hat{\mu}_{\tau^2})\right) \end{aligned}$$

and maximise to obtain an estimate for τ^2

Acknowledgements

We thank the reviewers for their helpful comments on earlier versions of our manuscript. The Medical Research Council (grant numbers G0800860 and MR/J004979/1) and the FWF (grant number P21763) funded this work.

References

1. Sampson A, Sill M. Drop-the-losers design: normal case. *Biometrical Journal* 2005; **47**:257–268.
2. Wu S, Wang W, Yang M. Interval estimation for drop-the-losers designs. *Biometrika* 2010; **97**:406–418.
3. US Food and Drug Administration (FDA) 2010. Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics.
4. Cohen A, Sackrowitz H. Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters* 1989; **8**:273–278.

5. Sill M, Sampson A. Extension of a two-stage conditionally unbiased estimator of the selected population to the bivariate normal case. *Communications in Statistics-Theory and Methods* 2007; **36**:801–813.
6. Bowden J, Glimm E. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal* 2008; **50**:515–527.
7. Kimani P, Stallard N, Todd S. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. *Statistics in Medicine* 2013. Published online DOI: doi/10.1002/sim.5757/pdf.
8. Stallard N, Todd S. Point estimators and confidence regions for sequential trials involving selection. *Journal of Statistical Planning and Inference* 2005; **135**:402–419.
9. Bebu I, Luta G, Dragalin V. Likelihood inference for a two-stage design with treatment selection. *Biometrical Journal* 2010; **52**:811–822.
10. Carreras M, Brannath W. Shrinkage estimation in two-stage adaptive designs with midtrial treatment selection. *Statistics in Medicine* 2013; **32**:1677–90. Published online DOI: doi/10.1002/sim.5463.
11. James W, Stein C. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Univ. California Press, Berkeley, 1961; 361–379.
12. Posch M, Koenig F, Branson M, Brannath W, Dunger-Baldauf C, Bauer P. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in Medicine* 2005; **24**:3697–3714.
13. Lindley D. Discussion of professor Stein's paper "confidence sets for the mean of a multivariate normal distribution". *Journal of the Royal Statistical Society, series B* 1962; **24**:265–296.
14. Efron B. Biased versus unbiased estimation. *Advances in Mathematics* 1975; **16**:259–277.
15. Hwang J. Empirical Bayes estimation for the means of the selected populations. *The Indian Journal of Statistics* 1993; **55**:285–311.
16. Carter G, Rolph J. Emprical Bayes methods applied to estimating fire alarm probabilities. *Journal of the American Statistical Association* 1974; **69**:880–885.
17. Viechtbauer W. Confidence intervals for the amount of heterogeneity in a meta-analysis. *Statistics in Medicine* 2007; **26**:37–52.
18. Bowden J, Tierney J, Copas A, Burdett S. Quantifying, displaying and accounting for heterogeneity in the meta-analysis of rcts using standard and generalised Q statistics. *Medical Research Methodology* 2011; **11**:41.
19. Paule R, Mandel J. Concensus values and weighting factors. *Journal of Research for the National Bureau of Standards* 1982; **87**:377–85.
20. DerSimonian R, Kacker R. Random effect models for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 2007; **28**:105–114.
21. Higgins J, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses. *British Medical Journal* 2003; **327**:557–560.
22. Efron B, Morris C. Limting the risk of Bayes and empirical Bayes estimators-part II: the empirical Bayes case. *Journal of the American Statistical Association* 1972; **67**:130–139.
23. Johnson D. An empirical Bayes approach to analyzing recurring animal surveys. *Ecology* 1989; **70**:945–952.
24. Hacking I. Slightly more realistic personal probability. *Philosophy of Science* 1967; **34**:311–325.
25. Cox D. Foundations of statistical inference: the case for eclecticism. *Australian Journal of Statistics* 1978; **20**:43–59.
26. Cox D, Hinkley D. *Theoretical Statistics*. Chapman and Hall: London, 1974.
27. Poole C, Greenland S. Random-effects meta-analyses are not always conservative. *American Journal of Epidemiology* 1999; **150**:469–475.
28. Henmi M, Copas J. Confidence intervals for random effects meta-analysis and robustness to publication bias. *Statistics in Medicine* 2010; **29**:2969–2983.
29. Bowater R, Escarela G. Heterogeneity and study size in random-effects meta-analysis. *Journal of Applied Statistics* 2013; **40**:2–16.
30. Bowden J, Glimm E. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multi-stage drop-the-losers trials. *Biometrical Journal* 2013. In press.
31. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
32. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**:494–501.
33. Royston P, Parmar M, Qian W. Novel designs for multi-arm clinical trials with survival outcomes, with an application in ovarian cancer. *Statistics in Medicine* 2003; **22**:2239–2256.

Statistics, Vol. 39, No. 3, June 2005, 207–219

A theorem on the principal components inference

JÜRGEN LÄUTER*† and EKKEHARD GLIMM‡

†Otto von Guericke University Magdeburg, Mittelstr. 2/151, 39114 Magdeburg, Germany
‡AICOS Technologies AG, Basel, Switzerland

(Received 3 September 2002; in final form 23 January 2005)

A method of multivariate data compression and dimension reduction is established, which is based on principal components and avoids all overfitting effects. This method allows the use of 'compressed' data for exact level-alpha tests of hypotheses on the mean vectors. It is a particularity of the method that the coefficients of the constructed linear scores depend solely on the residual sums of products matrix; the empirical means are not necessary to determine the compression. Thus, novel and very simple confidence regions of the unknown multivariate mean vectors are also obtained. The method can be combined with strategies of selecting variables. Furthermore, multiple testing procedures are derived, which serve for finding all sets of variables with deviations from the null hypothesis. The methods are evaluated by computer simulations.

Keywords: Multivariate test; Multivariate confidence region; Exact test; Principal components

AMS 1991 Subject Classification: 62F03; 62F25; 62H15; 62H25

1. Introduction

Principal components serve for data compression and dimension reduction. They have great importance for the representation and interpretation of measurements [1, 2]. However, principal components analysis is usually considered a heuristic method, a method of descriptive statistics. In most cases, it is not clear whether trustworthy inference regarding the structure of the mean values can be performed after the compression into principal components. This paper is devoted to parametric multivariate decisions, which utilize principal components in a mathematically rigorous way.

Some advance has been made since 1995 by the derivation of the so-called spherical tests [3–5]. To perform these tests, the given multivariate data vectors $\mathbf{x}' = (x_1\ x_2\ \cdots\ x_p)$ are transformed into vectors $\mathbf{z}' = \mathbf{x}'\mathbf{D}$ of a smaller dimension q by multiplication from the right with a $p \times q$ weight matrix $\mathbf{D}$. Here, $\mathbf{D}$ can have a fixed value but, more generally, $\mathbf{D}$ may be any function of the 'total sums of products matrix' corresponding to the test problem considered. For example, if $\mathbf{X}$ is an $n \times p$ sample matrix of n independent rows $\mathbf{x}'_{(j)}$ $(j = 1, \ldots, n)$ with the p-dimensional normal distribution $\mathbf{x}'_{(j)} \sim N_p(\boldsymbol{\mu}', \boldsymbol{\Sigma})$ and the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{\mu}'_0$

*Corresponding author. Email: juergen.laeuter@medizin.uni-magdeburg.de

Statistics
ISSN 0233-1888 print/ISSN 1029-4910 online © 2005 Taylor & Francis Group Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/02331880500077228

should be examined, then $\mathbf{W} = \sum_{j=1}^{n}(\mathbf{x}_{(j)} - \boldsymbol{\mu}_0)(\mathbf{x}_{(j)} - \boldsymbol{\mu}_0)'$ is the total sums of products matrix referring to $\boldsymbol{\mu}_0$. One can apply the data compression $\mathbf{Z} = \mathbf{XD}$ by the q first principal components using the weight matrix $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_q)$, where $\mathbf{d}_1, \ldots, \mathbf{d}_q$ are the eigenvectors with the q largest eigenvalues $\lambda_1, \ldots, \lambda_q$ of the eigenvalue problem $\mathbf{Wd} = \mathbf{d}\lambda$. Under the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{\mu}_0'$, the '$q$-dimensional' Hotelling statistic

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{D}(\mathbf{D}'\mathbf{GD})^{-1}\mathbf{D}'(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \tag{1}$$

is exactly distributed according to $(q/(n-q))F(q, n-q)$. Here, $\bar{\mathbf{x}}' = (1/n)\sum_{j=1}^{n}\mathbf{x}_{(j)}'$ is the mean vector and $\mathbf{G} = \sum_{j=1}^{n}(\mathbf{x}_{(j)} - \bar{\mathbf{x}})(\mathbf{x}_{(j)} - \bar{\mathbf{x}})'$ is the within-population sums of products matrix. Thus, exact inference for the p-dimensional mean vector $\boldsymbol{\mu}$ can be carried out, only using the q principal components.

It is remarkable that the spherical tests work with the F distribution or with the other distributions from classical multivariate analysis (like Wilks' Lambda) in spite of the dimension reduction and that no α adjustment is necessary as in the well-known multiple testing procedures. The spherical tests prove to be very effective in many applications, for example, in clinical trials with multiple endpoints or in high-dimensional gene expression analysis based on array technology [6–10]. However, this strategy has the obvious drawback that the weight matrix $\mathbf{D}$ and the derived principal components depend on the special null hypothesis (the value of $\boldsymbol{\mu}_0$) which is being considered. The compression from the data matrix $\mathbf{X}$ into principal components must be performed anew for each given null hypothesis. This is unsuitable, in particular, if confidence regions for multivariate parameters are desired.

To obtain invariance with respect to different null hypotheses, we will now consider principal components, which result from the residual sums of products matrix $\mathbf{G}$. The compression $\mathbf{Z} = \mathbf{XD}_G$ with the weight matrix $\mathbf{D}_G = (\mathbf{d}_{G1}, \ldots, \mathbf{d}_{Gq})$ is used, where $\mathbf{d}_{G1}, \ldots, \mathbf{d}_{Gq}$ are the eigenvectors of the eigenvalue problem $\mathbf{Gd}_G = \mathbf{d}_G\lambda_G$ to the q largest eigenvalues $\lambda_{G1}, \ldots, \lambda_{Gq}$. This means, $\mathbf{GD}_G = \mathbf{D}_G\boldsymbol{\Lambda}_G$ with $\boldsymbol{\Lambda}_G$ being the diagonal matrix of the eigenvalues $\lambda_{G1}, \ldots, \lambda_{Gq}$. We will see that these sample-based principal components are an adequate tool to transform a multivariate testing problem to a smaller dimension. Thus, the principal components acquire a new important role in addition to their 'naive' use for computational data compression [11, 12].

2. Theorem on the principal components inference

Consider a general multivariate linear testing problem with the $n \times p$ data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{(1)}' \\ \vdots \\ \mathbf{x}_{(n)}' \end{pmatrix} \sim N_{n\times p}(\mathbf{M}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}) \tag{2}$$

consisting of n independent p-dimensional normally distributed row vectors $\mathbf{x}_{(j)}'$ $(j = 1, \ldots, n)$. Here $\mathbf{I}_n$ is the $n \times n$ identity matrix and the symbol $\otimes$ denotes the Kronecker product. We would like to test the structure of the $n \times p$ matrix M of the mean. $\boldsymbol{\Sigma}$ is the unknown covariance matrix of the p variables.

$\mathbf{M}$ is supposed to have a linear model structure: we assume that there is an $n \times f_G$ model matrix $\mathbf{E}_G$ with $\mathbf{E}_G'\mathbf{E}_G = \mathbf{I}_{f_G}$ and $\mathbf{E}_G'\mathbf{M} = \mathbf{0}$. Then the residuals of this model are given by

$$\mathbf{E}_G'\mathbf{X} \sim N_{f_G\times p}(\mathbf{0}, \mathbf{I}_{f_G} \otimes \boldsymbol{\Sigma}), \tag{3}$$

and the corresponding residual sums of products matrix is

$$\mathbf{G} = \mathbf{X}'\mathbf{E}_G\mathbf{E}_G'\mathbf{X} \sim W_p(\mathbf{\Sigma}, f_G), \tag{4}$$

where W_p denotes the Wishart distribution.

Furthermore, a null hypothesis is defined: we assume that there exists an $n \times f_H$ matrix $\mathbf{E}_H$ with $\mathbf{E}_H'\mathbf{E}_H = \mathbf{I}_{f_H}$ and $\mathbf{E}_H'\mathbf{E}_G = \mathbf{0}$. The null hypothesis puts an additional condition on the matrix $\mathbf{M}$, namely $\mathbf{E}_H'\mathbf{M} = \mathbf{\Delta}_0$. Here, $\mathbf{\Delta}_0$ has a fixed given value. Under this null hypothesis,

$$\mathbf{E}_H'\mathbf{X} - \mathbf{\Delta}_0 \sim N_{f_H \times p}(\mathbf{0}, \mathbf{I}_{f_H} \otimes \mathbf{\Sigma}), \tag{5}$$

and the hypothesis sums of products matrix

$$\mathbf{H} = (\mathbf{E}_H'\mathbf{X} - \mathbf{\Delta}_0)'(\mathbf{E}_H'\mathbf{X} - \mathbf{\Delta}_0) \sim W_p(\mathbf{\Sigma}, f_H) \tag{6}$$

is obtained. $\mathbf{G}$ and $\mathbf{H}$ are stochastically independent.

THEOREM 1 *Assume a number of variables p and a number of principal components q ($1 \le q \le p$). Let $\mathbf{G}$ and $\mathbf{H}$ be two independent Wishart distributed $p \times p$ matrices*,

$$\mathbf{G} \sim W_p(\mathbf{\Sigma}, f_G), \quad \mathbf{H} \sim W_p(\mathbf{\Sigma}, f_H) \ (\mathbf{\Sigma} \textit{ unknown}, \ \operatorname{rank}(\mathbf{\Sigma}) \ge q, \ f_G \ge q, \ f_H \ge 1). \tag{7}$$

Let $\mathbf{D}_G$ be the $p \times q$ eigenvector matrix and $\mathbf{\Lambda}_G$ be the $q \times q$ diagonal eigenvalue matrix of the eigenvalue problem

$$\mathbf{G}\mathbf{D}_G = \mathbf{D}_G\mathbf{\Lambda}_G \tag{8}$$

pertaining to the q largest eigenvalues $\lambda_{G1}, \ldots, \lambda_{Gq}$ ($\mathbf{D}_G'\mathbf{D}_G = \mathbf{I}_q, \lambda_{G1} \ge \lambda_{G2} \ge \cdots$).

Then, the determinant statistic

$$\text{Lambda}_G = \frac{|\mathbf{D}_G'\mathbf{G}\mathbf{D}_G|}{|\mathbf{D}_G'(\mathbf{G}+\mathbf{H})\mathbf{D}_G|} = \frac{|\mathbf{\Lambda}_G|}{|\mathbf{\Lambda}_G + \mathbf{D}_G'\mathbf{H}\mathbf{D}_G|} \tag{9}$$

is stochastically not smaller than Wilks' Lambda distribution $\Lambda(q, f_H, f_G)$, which corresponds to q variables and f_H, f_G degrees of freedom of the hypothesis and the residuals, respectively, see ref. [13, p. 299]. *This means*

$$Pr(\text{Lambda}_G \le \Lambda_\alpha(q, f_H, f_G)) \le \alpha \quad \textit{for every } \alpha \in (0, 1), \tag{10}$$

where $\Lambda_\alpha(q, f_H, f_G)$ is the α quantile of Wilks' Lambda.

Proof In addition to the eigenvalue problem (8), we consider the corresponding eigenvalue problem with the total sums of products matrix $\mathbf{W} = \mathbf{G} + \mathbf{H}$:

$$\mathbf{W}\mathbf{D} = \mathbf{D}\mathbf{\Lambda}. \tag{11}$$

Here, $\mathbf{D}$ is the $p \times q$ eigenvector matrix and $\mathbf{\Lambda}$ is the $q \times q$ diagonal eigenvalue matrix of $\mathbf{W}$ ($\mathbf{D}'\mathbf{D} = \mathbf{I}_q, \lambda_1 \ge \lambda_2 \ge \cdots$).

An important property of the eigenvectors is that they maximize certain determinants:

$$|\mathbf{\Lambda}_G| = \left|\mathbf{D}_G'\mathbf{G}\mathbf{D}_G\right| = \max_{\mathbf{Y}} \left|\mathbf{Y}'\mathbf{G}\mathbf{Y}\right|, \quad |\mathbf{\Lambda}| = \left|\mathbf{D}'\mathbf{W}\mathbf{D}\right| = \max_{\mathbf{Y}} \left|\mathbf{Y}'\mathbf{W}\mathbf{Y}\right|, \tag{12}$$

where $\mathbf{Y}$ goes through all $p \times q$ matrices with $\mathbf{Y}'\mathbf{Y} = \mathbf{I}_q$; see ref. [14, p. 65]. Therefore,

$$\left|\mathbf{D}'\mathbf{G}\mathbf{D}\right| \le \left|\mathbf{D}_G'\mathbf{G}\mathbf{D}_G\right|, \quad \left|\mathbf{D}_G'(\mathbf{H}+\mathbf{G})\mathbf{D}_G\right| \le \left|\mathbf{D}'(\mathbf{H}+\mathbf{G})\mathbf{D}\right|. \tag{13}$$

Thus,

$$\text{Lambda}_G = \frac{\left|\mathbf{D}_G'\mathbf{G}\mathbf{D}_G\right|}{\left|\mathbf{D}_G'(\mathbf{H}+\mathbf{G})\mathbf{D}_G\right|} \ge \frac{\left|\mathbf{D}'\mathbf{G}\mathbf{D}\right|}{\left|\mathbf{D}'(\mathbf{H}+\mathbf{G})\mathbf{D}\right|} = \text{Lambda}. \tag{14}$$

The Lambda ratio on the right-hand side is distributed according to Wilks' Lambda distribution $\Lambda(q, f_H, f_G)$. This is a consequence of the general theorems in ref. [5] on spherical tests, because $\mathbf{D}$ is a function of the total sums of products matrix $\mathbf{W}$. We see that Lambda_G is pointwisely not smaller than Lambda. Hence, the theorem is proved. ■

Theorem 1 facilitates the application of Wilks' Lambda test to the compressed q-dimensional data $\mathbf{Z} = \mathbf{X}\mathbf{D}_G$. A significant value $\text{Lambda}_G \le \Lambda_\alpha(q, f_H, f_G)$ indicates that the p-dimensional null hypothesis $\mathbf{E}_H'\mathrm{M} = \mathbf{\Delta}_0$ has to be rejected at significance level α. In contrast to the spherical tests derived since 1995, the weight matrix does not depend on the special null hypothesis. The compression of the data matrix $\mathbf{X}$ through the weight matrix $\mathbf{D}_G$ is universally valid in the linear model defined by $\mathbf{E}_G$ for all null hypotheses. The theorem shows that the search of the combinations of variables which have 'maximal residual variances' is compatible with the multivariate test. However, one can also see that the data compression depends on the scales of the variables. Changes in the scale of a variable also change the weight matrix $\mathbf{D}_G$. We briefly comment on this in section 6.

The test based on Lambda_G is conservative, that is, its error of first kind generally is smaller than the given value α. However, in situations with highly correlated variables, the deviation from the given value α is slight.

As far as we could recognize, Theorem 1 cannot be generalized to include other well-known multivariate test statistics, such as Hotelling's T^2.

If the null hypothesis has only one degree of freedom ($f_H = 1$) or if only one principal component is applied ($q = 1$), then Fisher's F distribution can be used instead of Wilks' Lambda distribution. This is detailed in the following corollaries.

COROLLARY 1 *Under the assumption of Theorem* 1 *and if the null hypothesis has only one degree of freedom,* $f_H = 1$*, the distribution of the statistic*

$$F_G = \frac{f_G - q + 1}{q} tr\left((\mathbf{D}_G'\mathbf{H}\mathbf{D}_G)(\mathbf{D}_G'\mathbf{G}\mathbf{D}_G)^{-1}\right) = \frac{f_G - q + 1}{q} tr\left((\mathbf{D}_G'\mathbf{H}\mathbf{D}_G)\mathbf{\Lambda}_G^{-1}\right) \tag{15}$$

is stochastically not larger than Fisher's F distribution $F(q, f_G - q + 1)$*, that is,*

$$Pr(F_G \ge F_{1-\alpha}(q, f_G - q + 1)) \le \alpha \quad \textit{for every } \alpha \in (0, 1), \tag{16}$$

where $F_{1-\alpha}(q, f_G - q + 1)$ *is the* $(1 - \alpha)$ *quantile of Fisher's F distribution.*

COROLLARY 2 *Under the assumption of Theorem* 1 *and if only one principal component is used,* $q = 1$*, the distribution of the statistic*

$$F_G = \frac{f_G}{f_H} \frac{\mathbf{d}_{G1}'\mathbf{H}\mathbf{d}_{G1}}{\mathbf{d}_{G1}'\mathbf{G}\mathbf{d}_{G1}} = \frac{f_G}{f_H} \frac{\mathbf{d}_{G1}'\mathbf{H}\mathbf{d}_{G1}}{\lambda_{G1}} \tag{17}$$

is stochastically not larger than Fisher's F distribution $F(f_H, f_G)$.

These corollaries can be proved from Theorem 1 by application of the well-known relations between Wilks' Lambda and Fisher's F distribution in the cases of $f_H = 1$ or $q = 1$.

Application We consider the one-sample test of dimension p. The given data matrix is $\mathbf{X} \sim N_{n\times p}(\mathbf{1}_n\mu', \mathbf{I}_n \otimes \mathbf{\Sigma})$, where $\mathbf{1}_n$ represents the $n \times 1$ vector consisting of ones only. The corresponding residual sums of products matrix is

$$\mathbf{G} = \sum_{j=1}^{n} (\mathbf{x}_{(j)} - \bar{\mathbf{x}})(\mathbf{x}_{(j)} - \bar{\mathbf{x}})' = (\mathbf{X} - \bar{\mathbf{X}})'(\mathbf{X} - \bar{\mathbf{X}}) \sim W_p(\mathbf{\Sigma}, n-1) \tag{18}$$

with $\bar{\mathbf{x}}' = 1/n \sum_{j=1}^{n} \mathbf{x}'_{(j)}$, $\bar{\mathbf{X}} = \mathbf{1}_n\bar{\mathbf{x}}'$. The null hypothesis $\boldsymbol{\mu}' = \boldsymbol{\mu}'_0$ should be tested. Thus, we obtain the hypothesis sums of products matrix

$$\mathbf{H} = (\bar{\mathbf{X}} - \mathbf{M}_0)'(\bar{\mathbf{X}} - \mathbf{M}_0) = (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)\mathbf{1}'_n\mathbf{1}_n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)', \tag{19}$$

where $\mathbf{M}_0 = \mathbf{1}_n\boldsymbol{\mu}'_0$. Under the null hypothesis, $\mathbf{H} \sim W_p(\mathbf{\Sigma}, 1)$.

In the case that only one principal component is used ($q = 1$), Corollaries 1 and 2 provide the statistic

$$F_G = (n-1)n\frac{\left((\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{d}_{G1}\right)^2}{\lambda_{G1}}, \tag{20}$$

where $\mathbf{d}_{G1}$ is the eigenvector from the eigenvalue problem $\mathbf{G}\mathbf{d}_G = \mathbf{d}_G\lambda_G$ pertaining to the largest eigenvalue λ_{G1} with $\mathbf{d}'_{G1}\mathbf{d}_{G1} = 1$. According to the corollaries, the distribution of F_G under the null hypothesis is not larger than the F distribution $F(1, n-1)$. Consequently, the test outcome $F_G \geq F_{1-\alpha}(1, n-1)$ results in rejection of the p-dimensional null hypothesis. Once more, we emphasize that the weight vector $\mathbf{d}_{G1}$ does not depend on the special value μ_0.

In the following, some simulations of the F test (20) are presented. The simulations refer to a symmetric parameter structure of four variables, that is, four variables with equal means, equal variances and equal correlation coefficients between them. The null and the alternative hypotheses are also assumed to be symmetric:

$$\boldsymbol{\mu}' = \begin{pmatrix}\mu & \mu & \mu & \mu\end{pmatrix}, \quad \boldsymbol{\mu}'_0 = \begin{pmatrix}\mu_0 & \mu_0 & \mu_0 & \mu_0\end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix}1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1\end{pmatrix}. \tag{21}$$

We use samples of size $n = 12$ and a nominal significance level of $\alpha = 0.05$. Table 1 shows the incidence of rejecting the null hypothesis for the correlations $\rho = 0.20, 0.40, 0.60, 0.90$. These values are presented for $\Delta^2 = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\mathbf{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = 1$ and for the 'null condition' $\Delta^2 = 0$. The latter illustrate the extent of conservatism of this test, that is, they show to what extent the test really exhausts the nominal significance level $\alpha = 0.05$. The test has been replicated in the simulations 100,000 times in each case.

The classical multivariate test, Wilks' Lambda test, has power 0.5316 for $\Delta^2 = 1$. Hence, the F_G test (20) is superior to the classical test with higher correlations, even though it does not exhaust the given significance level α.

Table 1. Power values and real level of significance for $\alpha = 0.05$ and for different correlations ρ.

Correlation ρ	0.20	0.40	0.60	0.90
Probability of rejection for $\Delta^2 = 1$	0.5326	0.7851	0.8598	0.8796
Probability of rejection for $\Delta^2 = 0$	0.0091	0.0214	0.0364	0.0465

3. Principal components inference combined with selection of variables

The data compression method of section 2 may be applied in conjunction with a selection of variables. We will use again the maximization of the determinants $\left|\mathbf{Y}'\mathbf{G}\mathbf{Y}\right|$ and $\left|\mathbf{Y}'(\mathbf{G}+\mathbf{H})\mathbf{Y}\right|$, but with other restrictions on the arguments $\mathbf{Y}$.

We will assume an arbitrary set M of subsets m of all p variables ($m = m_1, m_2, m_3, \ldots$). Each subset (or part) m is characterized by the indices $i_1, \ldots, i_{p^{(m)}}$ of the corresponding variables, $m = \text{part}\{i_1, \ldots, i_{p^{(m)}}\}$, where $p^{(m)}$ denotes the number of variables contained in m. For example, we can take the p subsets of single variables (with $p^{(m)} = 1$), all subsets m of two variables each (with $p^{(m)} = 2$), the p ascending subsets $m_1 = \text{part}\{1\}$, $m_2 = \text{part}\{1, 2\}, \ldots, m_p = \text{part}\{1, \ldots, p\}$ (with $p^{(m)} = 1$ to $p^{(m)} = p$), or all $2^p - 1$ subsets m consisting of at least one variable (with $1 \le p^{(m)} \le p$). In our strategy, principal components are calculated for each of the given subsets $m_1, m_2, m_3, \ldots$ of variables. Subsequently, the subset with principal components revealing the 'highest residual variance' is chosen for the multivariate test. In many applications, this method leads to a valuable subset of variables and to a high power of the test.

Let m be a fixed subset of variables and q a fixed number of principal components with $1 \le q \le p^{(m)}$. Then corresponding $p \times q$ weight matrices $\mathbf{Y}^{(m)}$ are defined. These matrices consist of $p^{(m)}$ 'essential' rows, assigned to the variables of m, and of $p - p^{(m)}$ 'unessential' rows of zeros only. From all matrices $\mathbf{Y}^{(m)}$ of this form, $\mathbf{D}_G^{(m)} = (\mathbf{d}_{G1}^{(m)}, \ldots, \mathbf{d}_{Gq}^{(m)})$ is the eigenvector matrix whose columns $\mathbf{d}_{G1}^{(m)}, \ldots, \mathbf{d}_{Gq}^{(m)}$ are just the solutions of the eigenvalue problem $\mathbf{G}^{(m)}\mathbf{d}_G^{(m)} = \mathbf{d}_G^{(m)}\lambda_G^{(m)}$ pertaining to the q largest eigenvalues $\lambda_{G1}^{(m)}, \ldots, \lambda_{Gq}^{(m)}$ in decreasing order. Here, $\mathbf{G}^{(m)}$ is the $p \times p$ sums of products matrix obtained from $\mathbf{G}$ by setting to zero all rows and columns with an index outside m. It is easy to show that $\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)} = \mathbf{D}_G^{(m)\prime}\mathbf{D}_G^{(m)}\mathbf{\Lambda}_G^{(m)}$ holds, where $\mathbf{\Lambda}_G^{(m)}$ is the $q \times q$ diagonal matrix of the eigenvalues $\lambda_{G1}^{(m)}, \ldots, \lambda_{Gq}^{(m)}$.

THEOREM 2 *Assume several subsets $m = m_1, m_2, m_3, \ldots$ of variables with the sizes $p^{(m)}$ and a number of principal components q with $1 \le q \le p^{(m)} \le p$ for all m. Let $\mathbf{G}$ and $\mathbf{H}$ be two independent Wishart distributed $p \times p$ matrices,*

$$\mathbf{G} \sim W_p(\mathbf{\Sigma}, f_G), \quad \mathbf{H} \sim W_p(\mathbf{\Sigma}, f_H) \ (f_G \ge q, f_H \ge 1), \tag{22}$$

and let all $p^{(m)} \times p^{(m)}$ submatrices of $\mathbf{\Sigma}$ corresponding to the subsets m have a rank larger than or equal to q. Let, for any m, $\mathbf{D}_G^{(m)}$ be the $p \times q$ eigenvector matrix and $\mathbf{\Lambda}_G^{(m)}$ the $q \times q$ diagonal eigenvalue matrix of the eigenvalue problem

$$\mathbf{G}^{(m)}\mathbf{D}_G^{(m)} = \mathbf{D}_G^{(m)}\mathbf{\Lambda}_G^{(m)} \tag{23}$$

pertaining to the q largest eigenvalues $\lambda_{G1}^{(m)}, \ldots, \lambda_{Gq}^{(m)}$, where the special normalization condition

$$\mathbf{D}_G^{(m)\prime}\mathbf{D}_G^{(m)} = \frac{1}{\sqrt{p^{(m)}}}\mathbf{I}_q \tag{24}$$

is applied. Select from all given subsets of variables the subset $m = m_{\text{opt}}$ with the maximum value of the assessment criterion

$$\text{O}^{(m)} = \left|\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)}\right| = \frac{1}{\left(p^{(m)}\right)^{q/2}}\left|\mathbf{\Lambda}_G^{(m)}\right| \tag{25}$$

as the 'optimum subset', and let $\mathbf{D}_G^{\text{opt}} = \mathbf{D}_G^{(m_{\text{opt}})}$ be the corresponding 'optimum weight matrix'.

Then, the distribution of the statistic

$$\text{Lambda}_G = \frac{\left|\mathbf{D}_G^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}_G^{\text{opt}}\right|}{\left|\mathbf{D}_G^{\text{opt}\,\prime}(\mathbf{G}+\mathbf{H})\mathbf{D}_G^{\text{opt}}\right|} \tag{26}$$

is stochastically not smaller than Wilks' Lambda distribution $\Lambda(q, f_H, f_G)$.

Proof According to the preceding construction, $\mathbf{D}_G^{\text{opt}}$ is the $p \times q$ weight matrix with

$$\left|\mathbf{D}_G^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}_G^{\text{opt}}\right| = \max_{\mathbf{Y}} \left|\mathbf{Y}'\mathbf{G}\mathbf{Y}\right|, \tag{27}$$

where $\mathbf{Y}$ goes through all $p \times q$ matrices having the restrictions, first, to correspond to one of the subsets m and, secondly, to satisfy the normalization $\mathbf{Y}'\mathbf{Y} = (1/\sqrt{p^{(m)}})\mathbf{I}_q$. Just as in Theorem 1, we also consider the analogous maximization problem for the total sums of products matrix $\mathbf{W} = \mathbf{G} + \mathbf{H}$. The optimum solution $\mathbf{D}^{\text{opt}}$ is obtained from the eigenvectors of the $p \times p$ matrices $\mathbf{W}^{(m)}$, with zeros in the rows and columns not pertaining to m. $\mathbf{D}^{\text{opt}}$ has the property

$$\left|\mathbf{D}^{\text{opt}\,\prime}(\mathbf{G}+\mathbf{H})\mathbf{D}^{\text{opt}}\right| = \max_{\mathbf{Y}} \left|\mathbf{Y}'(\mathbf{G}+\mathbf{H})\mathbf{Y}\right|, \tag{28}$$

where $\mathbf{Y}$ still satisfies the same restrictions. Then, of course,

$$\left|\mathbf{D}^{\text{opt}\prime}\mathbf{G}\mathbf{D}^{\text{opt}}\right| \le \left|\mathbf{D}_G^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}_G^{\text{opt}}\right| \quad \text{and} \quad \left|\mathbf{D}_G^{\text{opt}\,\prime}(\mathbf{H}+\mathbf{G})\mathbf{D}_G^{\text{opt}}\right| \le \left|\mathbf{D}^{\text{opt}\,\prime}(\mathbf{H}+\mathbf{G})\mathbf{D}^{\text{opt}}\right|, \tag{29}$$

and therefore

$$\text{Lambda}_G = \frac{\left|\mathbf{D}_G^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}_G^{\text{opt}}\right|}{\left|\mathbf{D}_G^{\text{opt}\,\prime}(\mathbf{H}+\mathbf{G})\mathbf{D}_G^{\text{opt}}\right|} \ge \frac{\left|\mathbf{D}^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}^{\text{opt}}\right|}{\left|\mathbf{D}^{\text{opt}\,\prime}(\mathbf{H}+\mathbf{G})\mathbf{D}^{\text{opt}}\right|} = \text{Lambda}. \tag{30}$$

For an arbitrary fixed set of subsets $m_1, m_2, m_3, \ldots$, the optimum weight matrix $\mathbf{D}^{\text{opt}}$ is a function of the total sums of products matrix $\mathbf{G} + \mathbf{H}$. Therefore, the general theorems on spherical tests [5] yield that the Lambda ratio on the right-hand side of equation (30) has exactly Wilks' Lambda distribution $\Lambda(q, f_H, f_G)$. Thus, the inequality $\text{Lambda}_G \ge \text{Lambda}$ verifies Theorem 2. ■

It should be noted that the normalization condition (24) is not imperative for the correctness of Theorem 2. Any condition $\mathbf{D}_G^{(m)\,\prime}\mathbf{D}_G^{(m)} = c^{(m)}\mathbf{I}_q$ with any value $c^{(m)} > 0$ depending on the subsets m could be used. The normalization coefficients $c^{(m)} = 1/\sqrt{p^{(m)}}$ are intended to avoid a trivial preference for extremely large or extremely small subsets of variables. For example, the coefficients $c^{(m)} = 1$ would particularly emphasize subsets with many variables and $c^{(m)} = (1/p^{(m)})$ would prefer subsets of small sizes $p^{(m)}$.

Theorem 2 does not only provide a result about the p-dimensional test of the means, but also a characterization of the identified subset of variables. The following Addendum shows that the randomly selected subset m_{opt} can be regarded as relevant with respect to the deviations from the null hypothesis if significance is attained. The Addendum does not require that the null hypothesis is fulfilled for the parameters of all p variables.

ADDENDUM TO THEOREM 2 *Consider the assumptions from Theorem* 2 *with the exception of the distribution of the hypothesis sums of products matrix* $\mathbf{H}$. *We now assume a fixed subset*

of variables m_0 with p_0 variables, which meets the true null hypothesis. In practice, we do not need to know m_0. $\mathbf{H}$ *is supposed to be a random positive semidefinite $p \times p$ matrix being independent of* $\mathbf{G}$ *and having the Wishart distribution*

$$\mathbf{H}^{(m_0)} \sim W_p(\mathbf{\Sigma}^{(m_0)}, f_H), \tag{31}$$

where $\mathbf{H}^{(m_0)}$ *and* $\mathbf{\Sigma}^{(m_0)}$ *are the $p \times p$ matrices obtained from* $\mathbf{H}$ *and* $\mathbf{\Sigma}$*, respectively, by setting to zero all rows and columns not pertaining to m_0.*

If the 'optimum subset' m_{opt} and the 'optimum weight matrix' $\mathbf{D}_G^{\text{opt}}$ *are derived as in Theorem* 2 *under these modified conditions, then the probability of $m_{\text{opt}} \subseteq m_0$ in conjunction with significance* $\text{Lambda}_G \leq \Lambda_\alpha(q, f_H, f_G)$ *is at most α.*

Proof It is important to note that the values of $\mathbf{D}_G^{(m)}$, $\left|\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)}\right|$ and

$$\text{Lambda}_G^{(m)} = \frac{\left|\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)}\right|}{\left|\mathbf{D}_G^{(m)\prime}(\mathbf{G}+\mathbf{H})\mathbf{D}_G^{(m)}\right|}$$

are determined exclusively by the variables pertaining to m.

At first, consider only those subsets m of variables that are contained in m_0. Then, applying Theorem 2 yields an optimum subset $m_{\text{opt}} \subseteq m_0$ and an optimum weight matrix $\mathbf{D}_G^{\text{opt}}$ such that the statistic Lambda_G is significant with probability α, at most. Now, if all given subsets m are taken into account, the maximum of $\left|\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)}\right|$ might of course be reached for some subset $m = m_{\text{opt}}$ that is not contained in m_0. As this possibility reduces the probability of selecting $m_{\text{opt}} \subseteq m_0$ and then obtaining significance, the Addendum is also proved in the general case. ■

As the consequence of the Addendum, a resulting significance at the identified subset of variables m_{opt} is interpreted as indication of deviations from the null hypothesis in this set and, therefore, in all supersets of m_{opt}. To attain significance of any set of variables, all subsets m contained in it can be utilized. Nevertheless, α adjustment like Bonferroni or Bonferroni–Holm is not necessary. The method exactly keeps the 'multiple level' α. Theorem 2 establishes a new method for the selection of variables, which is based solely on the residual sums of products matrix $\mathbf{G}$. Knowledge of the mean values is not required.

A particularly interesting case considers only the p single-element sets $m_1, \ldots, m_p$ corresponding to the variables 1 to p. We have $q = 1$, and according to Theorem 2, the index $i = i_{\text{opt}}$ with the maximum value of $\text{O}^{(m_i)} = \mathbf{d}_{G1}^{(m_i)\prime}\mathbf{G}\mathbf{d}_{G1}^{(m_i)} = g_{ii}$ is determined. Then the test statistic

$$\text{Lambda}_G = \frac{g_{i_{\text{opt}}i_{\text{opt}}}}{g_{i_{\text{opt}}i_{\text{opt}}} + h_{i_{\text{opt}}i_{\text{opt}}}} \quad \text{or} \quad F_G = \frac{f_G}{f_H}\frac{h_{i_{\text{opt}}i_{\text{opt}}}}{g_{i_{\text{opt}}i_{\text{opt}}}}, \tag{32}$$

respectively, is calculated. Thus, the p-dimensional test is performed with the formula from a univariate test. The transition from dimension p to dimension 1 is managed simply by selecting the variable i with the largest residual sum of squares g_{ii}. This alone suffices to render any α adjustment superfluous. If significance is attained, that is $\text{Lambda}_G \leq \Lambda_\alpha(1, f_H, f_G)$ or $F_G \geq F_{1-\alpha}(f_H, f_G)$, variable i_{opt} can be regarded as relevant. This principle is also used in section 5 for the construction of a multiple test procedure that identifies all relevant variables.

Application. To illustrate the application of Theorem 2, we consider again the one-sample problem with observations $\mathbf{X} \sim N_{n\times p}(\mathbf{1}_n\boldsymbol{\mu}', \mathbf{I}_n \otimes \boldsymbol{\Sigma})$ and the null hypothesis $\boldsymbol{\mu}' = \boldsymbol{\mu}_0'$. Then, the matrices $\mathbf{G}$ and $\mathbf{H}$ are determined by equations (18) and (19). In the case of one principal component ($q = 1$), the optimal subset m_{opt} can be selected from the given sets $m_1, m_2, m_3, \ldots$ by Theorem 2 and the statistic

$$F_G = (n-1)n\frac{\left((\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\mathbf{d}_{G1}^{\text{opt}}\right)^2}{\mathbf{d}_{G1}^{\text{opt}\,\prime}\mathbf{G}\mathbf{d}_{G1}^{\text{opt}}} \tag{33}$$

is determined. If $F_G \geq F_{1-\alpha}(1, n-1)$, we conclude that $\mu_i \neq \mu_{0i}$ somewhere within m_{opt}.

We will present some corresponding simulation results. The simulations are done for four variables and a special parameter structure, where the first two variables contribute genuine information and the other two represent 'independent noise'. In reality, it is of course unknown whether such irrelevant variables exist and at which positions they are. The structures considered here are

$$\boldsymbol{\mu}' = \mu\begin{pmatrix}1 & 1 & 0 & 0\end{pmatrix}, \quad \boldsymbol{\mu}_0' = \mu_0\begin{pmatrix}1 & 1 & 0 & 0\end{pmatrix}, \tag{34}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix}1 & \rho & 0 & 0\\ \rho & 1 & 0 & 0\\ 0 & 0 & 1-\rho & 0\\ 0 & 0 & 0 & 1-\rho\end{pmatrix} = (1-\rho)\mathbf{I}_4 + \rho\begin{pmatrix}1\\1\\0\\0\end{pmatrix}\begin{pmatrix}1 & 1 & 0 & 0\end{pmatrix}, \tag{35}$$

where ρ is the correlation coefficient of the first two variables and $1-\rho$ is the 'specific variance' of each variable in the four-dimensional 'one-factor structure'. It is desired that the relevant block part{1, 2} is correctly identified.

Just as in section 2, we consider $n = 12$ and $\alpha = 0.05$. Table 2 contains the probabilities of identification of the blocks part{1, 2} and part{3, 4} if all 15 possible subsets of variables $m_1 =$ part{1}, $m_2 =$ part{2}, $\ldots$, $m_{15} =$ part{1, 2, 3, 4} are used. The block part{1, 2} is regarded as 'correctly identified' if $m_{\text{opt}} =$ part{1}, $m_{\text{opt}} =$ part{2} or $m_{\text{opt}} =$ part{1, 2} are selected and, additionally, significance is attained. The selection is 'wrong' if $m_{\text{opt}} =$ part{3}, $m_{\text{opt}} =$ part{4} or $m_{\text{opt}} =$ part{3, 4} are chosen in connection with significance. For the correlation, we consider $\rho = 0.20, 0.40, 0.60, 0.90$. The Mahalanobis distance is fixed as

$$\Delta^2 = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) = \frac{2(\mu - \mu_0)^2}{1+\rho} = 1. \tag{36}$$

Table 3 gives the probabilities that arise from restricting attention to the four single-element sets $m_1 =$ part{1}, $\ldots$, $m_4 =$ part{4}. In this case, only univariate criteria are used in the selection procedure (see (32)). Table 3 shows that for large correlations ρ, the relevant block part{1, 2} is not as well recognized as in case of table 2.

In addition, a comparison to the well-known Bonferroni method based on an α adjustment is performed in table 3. The Bonferroni method checks the 'best' single variable with a significance level of $\alpha/4$. It turns out to be superior to the method based on Theorem 2 in identifying

Table 2. Probability of recognition of the relevant and irrelevant subsets of variables ($\Delta^2\{1, 2\} = 1$, $\Delta^2\{3, 4\} = 0$, with all 15 subsets, $\alpha = 0.05$).

Correlation ρ	0.20	0.40	0.60	0.90
Probability of significance with $m_{\text{opt}} \subseteq$ part{1, 2}	0.3546	0.5578	0.7488	0.8802
Probability of significance with $m_{\text{opt}} \subseteq$ part{3, 4}	0.0030	0.0011	0.0001	0.0000

Table 3. Probability of recognition of the relevant and irrelevant subsets of variables ($\Delta^2\{1, 2\} = 1$, $\Delta^2\{3, 4\} = 0$, with four subsets of the single variables, $\alpha = 0.05$).

Correlation ρ	0.20	0.40	0.60	0.90
Probability of significance with $m_{\text{opt}} \subseteq \text{part}\{1, 2\}$	0.3701	0.5545	0.7143	0.8446
Probability of significance with $m_{\text{opt}} \subseteq \text{part}\{3, 4\}$	0.0035	0.0013	0.0002	0.0000
Probability of Bonferroni's significance in part{1, 2}	0.6405	0.6898	0.7178	0.7237
Probability of Bonferroni's significance in part{3, 4}	0.0132	0.0121	0.0108	0.0109

part{1, 2} if the correlation is small. However, it is inferior if the correlation is large. Of course, the selection method of Bonferroni's method is not only based on the within-population sums of products. Hence, this comparison is of minor importance for the data compression treated in this paper. The number of simulation runs for tables 2 and 3 was once again 100,000.

4. Confidence regions of the unknown mean vectors

It is obvious that the tests from sections 2 and 3 can be used to derive confidence regions. Under the assumptions of Theorem 1, we have

$$\text{Lambda}_G > \Lambda_\alpha(q, f_H, f_G) \tag{37}$$

with a probability of at least $1 - \alpha$. Hence, the values $\boldsymbol{\Delta}_0 = \mathbf{E}_H'\mathbf{M}_0$ that fulfill

$$\frac{|\boldsymbol{\Lambda}_G|}{|\boldsymbol{\Lambda}_G + \mathbf{D}_G'\mathbf{H}\mathbf{D}_G|} > \Lambda_\alpha(q, f_H, f_G) \quad \text{with } \mathbf{H} = (\mathbf{E}_H'\mathbf{X} - \boldsymbol{\Delta}_0)'(\mathbf{E}_H'\mathbf{X} - \boldsymbol{\Delta}_0) \tag{38}$$

define a confidence region for the unknown parameter $\boldsymbol{\Delta} = \mathbf{E}_H'\mathbf{M}$ at confidence level $1 - \alpha$. Note that the calculation of the confidence region in a concrete application is simplified by the fact that $\boldsymbol{\Lambda}_G$ and $\mathbf{D}_G$ are constants for a given fixed sample $\mathbf{X}$. If we vary $\boldsymbol{\Delta}_0$, then $\mathbf{H}$ is the only term in the inequality (38) that changes.

The confidence region has a particularly simple form in the case of $f_H = 1$ (Corollary 1). In this case, we have $\mathbf{H} = \mathbf{h}\mathbf{h}'$, where $\mathbf{h}' = \mathbf{x}_H' - \boldsymbol{\delta}_0'$ with $\mathbf{x}_H' = \mathbf{E}_H'\mathbf{X}$, $\boldsymbol{\delta}_0' = \boldsymbol{\Delta}_0 = \mathbf{E}_H'\mathbf{M}_0$. Then, the confidence region (38) is characterized by

$$\frac{f_G - q + 1}{q}(\mathbf{x}_H - \boldsymbol{\delta}_0)'\mathbf{D}_G\boldsymbol{\Lambda}_G^{-1}\mathbf{D}_G'(\mathbf{x}_H - \boldsymbol{\delta}_0) < F_{1-\alpha}(q, f_G - q + 1). \tag{39}$$

In the p-dimensional space, $\boldsymbol{\delta}_0'$ is in the interior of an elliptic cylinder with center $\mathbf{x}_H'$. If, in addition, $q = 1$ is assumed, the condition simplifies to

$$\frac{f_G}{\lambda_{G1}}\left((\mathbf{x}_H - \boldsymbol{\delta}_0)'\mathbf{d}_{G1}\right)^2 < F_{1-\alpha}(1, f_G) \tag{40}$$

or

$$\mathbf{x}_H'\mathbf{d}_{G1} - \sqrt{\frac{\lambda_{G1}F_{1-\alpha}(1, f_G)}{f_G}} < \boldsymbol{\delta}_0'\mathbf{d}_{G1} < \mathbf{x}_H'\mathbf{d}_{G1} + \sqrt{\frac{\lambda_{G1}F_{1-\alpha}(1, f_G)}{f_G}}. \tag{41}$$

This confidence region for the parameter $\boldsymbol{\delta}' = \mathbf{E}_H'\mathbf{M}$ is a 'hyperdisk', that is, a region limited by two parallel hyperplanes. Of course, this is an unbounded region in the p-dimensional space, but with respect to the most informative directions in this space, the region is very flat.

This special case illustrates the fundamental difference between our approach and Hotelling's classical elliptical confidence region, which is determined by

$$\frac{f_G - p + 1}{p}(\mathbf{x}_H - \boldsymbol{\delta}_0)'\mathbf{G}^{-1}(\mathbf{x}_H - \boldsymbol{\delta}_0) < F_{1-\alpha}(p, f_G - p + 1). \tag{42}$$

Let us consider a sample of size n from the population $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. According to equation (41), the limits of the confidence region for the unknown mean parameter $\boldsymbol{\delta}' = \sqrt{n}\boldsymbol{\mu}'$ are given by

$$\sqrt{n}\bar{\mathbf{x}}'\mathbf{d}_{G1} - \sqrt{\frac{\lambda_{G1}F_{1-\alpha}(1, n-1)}{n-1}} < \sqrt{n}\boldsymbol{\mu}_0'\mathbf{d}_{G1} < \sqrt{n}\bar{\mathbf{x}}'\mathbf{d}_{G1} + \sqrt{\frac{\lambda_{G1}F_{1-\alpha}(1, n-1)}{n-1}}. \tag{43}$$

If, for example, the 8×3 sample matrix

$$\mathbf{X} = \begin{pmatrix} 0.37 & 1.28 & 0.69 \\ 1.03 & 1.29 & 0.95 \\ 0.24 & 1.92 & 1.21 \\ 1.29 & 2.83 & 2.58 \\ -1.30 & -0.03 & -0.36 \\ 0.49 & 0.50 & 0.84 \\ 0.93 & 0.19 & 0.64 \\ 0.73 & 0.82 & 1.93 \end{pmatrix}, \quad \bar{\mathbf{x}}' = \begin{pmatrix} 0.47 & 1.10 & 1.06 \end{pmatrix} \tag{44}$$

has been observed, a confidence region at level $1 - \alpha = 0.95$ is obtained by

$$\sqrt{8}\bar{\mathbf{x}}'\mathbf{d}_{G1} - \sqrt{\frac{\lambda_{G1} \cdot 5.59}{7}} < \sqrt{8}\boldsymbol{\mu}_0'\mathbf{d}_{G1} < \sqrt{8}\bar{\mathbf{x}}'\mathbf{d}_{G1} + \sqrt{\frac{\lambda_{G1} \cdot 5.59}{7}}, \tag{45}$$

that is,

$$1.21 = 4.43 - 3.22 < 1.35\mu_{01} + 1.76\mu_{02} + 1.75\mu_{03} < 4.43 + 3.22 = 7.65. \tag{46}$$

Obviously, there is no individual limit on each component of $\boldsymbol{\mu}_0$. Rather, constraints apply to a certain linear combination, which is derived from the within-population sums of products matrix $\mathbf{G}$.

In a corresponding way, Theorem 2 can be used to derive confidence regions. In this case, the calculation of the confidence region involves a selection of variables. According to equation (26), all values $\boldsymbol{\Delta}_0 = \mathbf{E}_H'\mathbf{M}_0$ yielding

$$\frac{\left|\mathbf{D}_G^{\text{opt}\,\prime}\mathbf{G}\mathbf{D}_G^{\text{opt}}\right|}{\left|\mathbf{D}_G^{\text{opt}\,\prime}(\mathbf{G} + \mathbf{H})\mathbf{D}_G^{\text{opt}}\right|} > \Lambda_\alpha(q, f_H, f_G) \tag{47}$$

belong to a confidence region for $\boldsymbol{\Delta} = \mathbf{E}_H'\mathbf{M}$.

Suppose that this method is applied to the sets of individual variables $m_1 = \text{part}\{1\}, \ldots, m_p = \text{part}\{p\}$ with $q = 1$. Then, the inequality characterizing the confidence region can be

written according to equation (32) as

$$\frac{f_G}{f_H}\frac{h_{i_{\text{opt}}i_{\text{opt}}}}{g_{i_{\text{opt}}i_{\text{opt}}}} < F_{1-\alpha}(f_H, f_G), \tag{48}$$

where i_{opt} is the index of the variable with the largest g_{ii} $(i = 1, \ldots, p)$. If we consider the estimation of the mean vector $\boldsymbol{\mu}$ from $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, this means that

$$(n-1)n\frac{(\bar{x}_{i_{\text{opt}}} - \mu_{0i_{\text{opt}}})^2}{g_{i_{\text{opt}}i_{\text{opt}}}} < F_{1-\alpha}(1, n-1) \tag{49}$$

or

$$\bar{x}_{i_{\text{opt}}} - \sqrt{\frac{g_{i_{\text{opt}}i_{\text{opt}}}F_{1-\alpha}(1,n-1)}{(n-1)n}} < \mu_{0i_{\text{opt}}} < \bar{x}_{i_{\text{opt}}} + \sqrt{\frac{g_{i_{\text{opt}}i_{\text{opt}}}F_{1-\alpha}(1,n-1)}{(n-1)n}}. \tag{50}$$

For the numerical example (44), this yields $i_{\text{opt}} = 2$ and the 0.95 confidence region

$$0.31 = 1.10 - 0.79 < \mu_{02} < 1.10 + 0.79 = 1.89. \tag{51}$$

In this special case, one of the variables is selected, and a usual confidence interval is determined for it.

5. Multiple procedure for finding several relevant sets of variables

Theorem 2 can also be applied to derive a multiple procedure for recognizing all subsets of variables that deviate from a fixed null hypothesis $\mathbf{E}_H'\mathbf{M} = \boldsymbol{\Delta}_0$. Once again, let us consider several subsets of variables $m_1, m_2, \ldots$. Following Theorem 2, these subsets are first sorted in decreasing sequence according to the corresponding values of $|\mathbf{D}_G^{(m)\prime}\mathbf{G}\mathbf{D}_G^{(m)}|$, then they are tested in the resulting order (that is, $\text{Lambda}_G \leq \Lambda_\alpha(q, f_H, f_G)$ is checked successively). As soon as the first non-significant result occurs, the procedure is terminated. All sets obtained up to that point are regarded as relevant.

This procedure indeed keeps the multiple level α. To show this, assume that the null hypothesis is true for a set m_0. According to Theorem 2, the first subset m in the succession satisfying $m \subseteq m_0$ yields significance with a probability of α, at most. This is sufficient for the proof.

We are now applying this procedure to example (44) with the null hypothesis $\boldsymbol{\mu} = \mathbf{0}$ and the three subsets part{1}, part{2}, part{3}. First of all, the subsets are sorted according to the sums of squares g_{ii}. Subsequently, the steps of the multiple procedure for $\alpha = 0.05$ are performed:

$$\begin{aligned}
&\text{part\{2\}:}\ g_{22} = 6.28, \quad F_G = 10.82 \geq 5.59\ \text{(significant)},\\
&\text{part\{3\}:}\ g_{33} = 5.49, \quad F_G = 11.45 \geq 5.59\ \text{(significant)},\\
&\text{part\{1\}:}\ g_{11} = 4.47, \quad F_G = 2.78 < 5.59\ \text{(not significant, procedure stops)}.
\end{aligned} \tag{52}$$

However, this procedure is probably not the best one to find the non-null subsets of variables in a situation, where the non-centrality $\boldsymbol{\Delta}_0$ has a fixed value. Kropf [8] has proposed a similar procedure in which the p variables are sorted by the total sums of squares $w_{ii} = g_{ii} + h_{ii}$ $(i = 1, \ldots, p)$. Then, the sample means influence the order of the variables and thus the power can be increased [9]. The method of principal components inference presented in this paper is more adequate to applications with uncertain values of $\boldsymbol{\Delta}_0$.

6. Conclusion

In this paper, some theorems on principal components inference are presented which open new possibilities of multivariate data compression, of multivariate tests without overfitting effects, of confidence regions for the multivariate means, of model choice and of multiple procedures for finding all relevant sets of variables.

Unfortunately, our method of principle components inference crucially depends on the measurement scales of the given variables. The power of the dimension-reduced tests treated here and the corresponding confidence regions are influenced by a change in the scales. The Theorems 1 and 2 are primarily appropriate if all variables have been measured on nearly the same scale. In a factorial model, it might be desirable to have identical specific variances (for example, see the one-factor structure of equations (34) and (35)). If these standard conditions are not met, they may sometimes be reached by a transformation of the original data provided the corresponding coefficients are known from previous independent experiments. If two or more samples are to be compared regarding their means, we might be able to exploit the stochastic independence of the overall means from the sums of products matrices **G** and **H**. This independence implies that we can use the overall means to standardize the data, if these means can be considered as representations of the individual scales of the single variables.

The procedures presented in this paper tend to give attention to variables or linear combinations of variables with a large residual variation. Thus, relations between the means and the covariance matrix are necessary for these methods to obtain a high power. Factorial parameter structures are a class of models with great practical relevance, where this is the case. However, the main target of this paper remains to find a data compression, which is based solely on the residual covariances. The level of significance is always strictly kept.

References

[1] Seber, G.A.F., 1984, *Multivariate Observations* (New York: Wiley).
[2] Reyment, R.A. and Jöreskog, K.G., 1993, *Applied Factor Analysis in the Natural Sciences* (Cambridge: Cambridge University Press).
[3] Läuter, J., 1996 Exact *t* and *F* tests for analyzing studies with multiple endpoints. *Biometrics*, **52**, 964–970.
[4] Läuter, J., Glimm, E. and Kropf, S., 1996, New multivariate tests for data with an inherent structure. *Biometrical Journal*, **38**, 5–23. [Erratum in *Biometrical Journal*, **40**, 1015.]
[5] Läuter, J., Glimm, E. and Kropf, S., 1998, Multivariate tests based on left-spherically distributed linear scores. *The Annals of Statistics*, **26**, 1972–1988 (Correction: *The Annals Statistics*, **27**, 1441).
[6] Glimm, E., Heuer, H., Engelen, B., Smalla, K. and Backhaus, H., 1997, Statistical comparisons of community catabolic profiles. *Journal of Microbiological Methods*, **30**, 71–80.
[7] Reitmeir, P., 1999, *Statistische Methoden zur Analyse multipler Endpunkte in klinischen Studien.* Dissertation, Mediz. Fakultät d. Universität Köln.
[8] Kropf, S., 2000, *Hochdimensionale multivariate Verfahren in der medizinischen Statistik* (Aachen: Shaker).
[9] Kropf, S. and Läuter, J., 2002, Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal*, **44**, 789–800.
[10] Schuster, E. and Kropf, S., 2002, A new proposal for pairwise multiple comparisons with repeated measurements. *Pakistan Journal of Statistics*, **18**, 197–211.
[11] Läuter, J. and Kropf, S., 2002, Multivariate inference in clinical trials – tests with model choice, confidence regions. *Biocybernetics and Biomedical Engineering*, **22**, 5–16.
[12] Läuter, J. and Kropf, S., 2002, Data compression and selection of variables, with respect to exact inference. In: W. Härdle and B. Rönz (Eds) *Proceedings in Computational Statistics, Compstat 2002* (Heidelberg: Physika-Verlag), pp. 273–278.
[13] Anderson, T.W., 1984, *An Introduction to Multivariate Statistical Analysis* (New York: Wiley).
[14] Rao, C.R., 1973, *Linear Statistical Inference and its Applications*, 2nd edn (New York: Wiley).