Thesis for the Degree of Doctor of Philosophy

A Contribution to the Design and Analysis of Phase III Clinical Trials

Vera Lisovskaja

Division of Mathematical Statistics
Department of Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
Göteborg, Sweden 2013

A Contribution to the Design and Analysis of Phase III Clinical Trials
Vera Lisovskaja

ISBN 978–91–628–8740–7

Department of Mathematical Sciences
Division of Mathematical Statistics
Chalmers University of Technology
and University of Gothenburg
SE-412 96 GÖTEBORG, Sweden
Phone: +46 (0)31-772 10 00

Author e-mail: vera@chalmers.se

Typeset with LaTeX.
Department of Mathematical Sciences
Printed in Göteborg, Sweden 2013

A Contribution to the Design and Analysis of Phase III Clinical Trials
Vera Lisovskaja

Abstract

Clinical trials are an established methodology for evaluation of the effects of a new medical treatment. These trials are usually divided into several phases, namely phase I through IV. The earlier phases (I and II) are relatively small and have a more exploratory nature. The later phase III is confirmatory and aims to demonstrate the efficacy and safety of the new treatment. This phase is the final one before the treatment is marketed, with phase IV consisting of post-marketing studies.

Phase III is initiated only if the conductors of the clinical study judge that the evidence from earlier stages indicates clearly that the new treatment is effective. However, several studies performed in recent years show that this assessment is not always correct. Two papers written on the subject point out average attrition rates of around 45% and 30%. In other words, it is estimated that only around two thirds of the compounds that enter phase III finish it successfully. This thesis examines some of the possible ways of improving efficiency in phase III clinical trials. The thesis consists of four papers on various topics that touch this subject, these topics being adaptive designs (paper I), number of doses (paper II) and multiplicity correction procedures (papers III and IV).

The first paper examines the properties of the so called dual test, which can be applied in adaptive designs with sample size re-estimation. This test serves as a safeguard against unreasonable conclusions that may otherwise arise if an adaptive design is used. However, there is a price of possible power loss as compared to the standard test that is applied in such situations. The dual test is evaluated by considering several scenarios where its use would be natural. In many cases the power loss is minimal or non-existing.

The second paper considers the optimal number and placement of doses used in phase III, with the probability of success of the trial used as optimality criterion. One common way of designing phase III trials is to divide the patients into two groups, one group receiving the new drug and another a control. However, as is demonstrated in paper II, this approach will be inferior to a design with two different doses and a control if there is enough uncertainty in the dose-response model prior to the initiation of the trial.

The last two papers study possible gain that results from optimization of the multiplicity correction procedure that is applied if more than one hypothesis is tested in the same trial. Two families of such procedures are considered. The first one, examined in paper III, consists of a combination of a weighted Bonferroni test statistic with the principle of closed testing. The second one, examined in paper IV, is based on combining the same principle with a "pooled" test statistic. Paper III demonstrates that optimizing a multiplicity testing procedure can lead to a significant power increase as compared to simpler, non-optimized, procedures. The optimization is performed with respect to expected

utility, an approach that originates from decision theory. Paper IV examines the difference between the Bonferroni-based and the pooled-based multiplicity corrections, finding the latter to be superior to the former if the test statistics follow a known multivariate Normal distribution.

Keywords: Adaptive designs, decision theory, dose placement, dual test, closed testing procedures, expected utility, flexible designs, multiplicity, optimization, pooled test, probability of success, recycling procedures, sample size re-estimation, uncertainty, weighted multiplicity correction procedures

Bidrag till Design och Analys av Fas III Kliniska Prövningar
Vera Lisovskaja

Sammanfattning

Kliniska prövningar är en etablerad metodik för utvärdering av effekten av en ny medicinsk behandling. Dessa prövningar brukar delas upp i olika faser, nämligen fas I till IV. Tidigare faser (I och II) är relativt små och mer explorativa till sin natur. Den senare fas III är konfirmatorisk och har som mål att demonstrera att den nya behandlingen är effektiv och ofarlig. Denna fas är den sista fasen innan behandlingen eventuellt kommer ut på marknaden. Under fas IV bevakar man behandlingen efter att den har marknadsförts.

Fas III initieras bara om de som genomför den kliniska studien bedömer att det finns klara tecken från tidigare faser att den nya behandlingen är effektiv. Men flera studier som gjordes på senare tid visar att denna bedömning är inte alltid korrekt. Två artiklar som skrevs om problemet indikerar att runt 45% och 30% av nya läkemedel som påbörjar denna fas misslyckas. Med andra ord, ungefär två tredjedelar av läkemedel som går in i fas III studier är framgångsrika. I avhandlingen undersöks några möjliga sätt att förbättra effektiviteten av fas III kliniska prövningar. Avhandlingen består av fyra uppsatser om olika ämnen som berör kliniska prövningar i fas III, nämligen adaptiva designer (artikel I), val av antal doser (artikel II) och multiplicitetskorrektion (artikel III och IV).

Den första artikeln studerar egenskaper hos det så kallade duala testet, som kan användas i adaptiva designer som tillåter ändring av stickprovsstorleken. Testet skyddar på ett konservativt sätt mot felaktiga slutsatser som kan annars uppkomma då adaptiva designer används. Det finns dock ett pris i form av en möjlig minskning av styrkan jämfört med standardtesten för dessa designer. Dualtestet utvärderas genom examination av vanliga situationer där den skulle kunna användas. I de flesta fall är styrkeförlusten minimal eller icke-existerande.

Den andra artikeln studerar det optimala antalet doser som bör tas med i fas III samt deras placering, med sannolikheten för att denna fas ska lyckas som optimalitetskriterium. Ett vanligt sätt för att utföra fas III studier är att slumpmässigt dela upp patienter i två grupper och ge den ena gruppen den nya behandlingen och den andra kontrollsubstansen. I artikel II demonstreras att detta tillvägagångssätt är underlägset en design med två olika doser om det råder tillräckligt stor osäkerhet i modellen för dos-respons innan försöket initieras.

De sista två artiklarna studerar den möjliga vinsten som kan uppkomma då man optimerar multiplicitetskorrektionen som används när fler än en hypotes testas i samma studie. Två familjer av sådana procedurer studeras. Den första familjen, som undersöks i artikel III, består av en kombination av en viktad Bonferroni test statistika med slutna test principen. Den andra, som undersöks i artikel IV, baseras på en kombination av samma princip och en "poolad" test statistika. Artikel III visar att optimering av en multiplicitetsprocedur kan leda till en markant styrkeökning jämfört med enklare, icke optimerade,

procedurer. Optimeringen görs med avseende på förväntad nytta, ett tillvägagångssätt som härstammar från beslutsteori. Artikel IV studerar skillnaden mellan Bonferronibaserade och pooltestbaserade multiplicitetskorrektioner och visar att den senare är överlägsen den förra om teststatistikorna följer en känd multivariat normalfördelning.

Nyckelord: Adaptiva designer, beslutsteori, dosplacering, dualtest, flexibla designer, förväntad nytta, multiplicitet, optimering, osäkerhet, poolad test, recyclingprocedurer, sannolikhet för ett lyckad försök, slutna test, viktade multiplicitetsprocedurer, ändring av stickprovsstorlek

Acknowledgments

I would like to express my gratitude to...

Carl-Fredrik Burman, for the five years of guidance on the rocky road of my PhD studies...

Ziad Taib, for the quiet and constant support...

Olle Nerman, for the ability to look at a problem from angles that I would never fathom existed...

my mother, for always being there for me.

Vera Lisovskaja
Göteborg, October 22, 2013

List of papers

The PhD thesis is based on the following papers:

I. Burman, C-F. and **Lisovskaja, V.** (2010)

The dual test: Safeguarding p-value combination tests for adaptive designs. *Statistics in Medicine* **29**:797–807.

II. **Lisovskaja, V.** and Burman, C-F. (2013)

On the choice of doses for phase III clinical trials. *Statistics in Medicine* **10**:1661–1676.

III. **Lisovskaja, V.** and Burman, C-F. (*submitted*)

A decision theoretic approach to optimization of multiple testing procedures.

IV. **Lisovskaja, V.** and Burman, C-F. (*work in progress*)

Hierarchical multiplicity correction procedures based on pooled test statistic.

List of papers not included in this thesis

V. Tedeholm, H., Lycke, J., Skoog, B., **Lisovskaja, V.**, Hillert, J., Dahle, C., Fagius, J., Fredrikson, S., Landtblom, A. M., Malmestrom, C., Martin, C., Piehl, F., Runmarker, B., Stawiarz, L., Vrethem, M., Nerman, Olle., Andersen, O. (2013)

Time to secondary progression in patients with multiple sclerosis who were treated with first generation immunomodulating drugs. *Multiple Sclerosis Journal* **19**:765–774.

VI. Ydreborg, M., **Lisovskaja, V.**, Lagging, M., Christensen, P.B., Langeland, N., Buhl M.R., Pedersen, C., Mørch K., Wejstål, R., Norkrans, G., Lindh, M., Färkkilä, M., Westin, J. (*submitted*)

A novel fibrosis index comprising a non-cholesterol sterol accurately predicts HCV-related liver cirrhosis.

VII. Kneider, M., **Lisovskaja, V.**, Lycke, J., Jakobsen, J., Andersen, O. (*submitted*)

Search for peak MRI activity after an upper respiratory infection in relapsing-remitting MS.

Contents

Chapter 1: Introduction

Nowadays, statistical thinking and methodologies are present everywhere in biomedical research, from the most basic experiments performed on animals in a laboratory to the large scale studies of human patients that often span over years or even decades. Numerous journals devoted to the use of statistics in medicine exist, with hundreds of papers written every year. The unofficial guidelines that a paper presenting a medical study should include a statistical analysis has been a source of frustration to many a researcher during the recent years.

Prominent among the medical experiments are the clinical studies performed prior to marketing a new drug. The first clinical study is dated as far back as 1767, when a British physician William Watson explored the efficacy of treatments for smallpox in a group of thirty-odd children (12). Doctor Watson divided the children into three groups, two of those receiving different treatments, and compared the numbers of cured in each. In the following centuries the process of evaluating a treatment in an objective manner evolved and grew, becoming more and more complex and resource-consuming. Now, it is divided into several stages, each consisting of different experiments, or trials. The purpose of the trials ranges from the first tentative evaluation of the new drug in humans (phase I and II) to obtaining sufficient proof of this compounds effectiveness in order for it to be accepted for marketing (phase III). It is these confirmatory trials, conducted during the so called phase III of a clinical study, that are the focus of this thesis.

Despite the staggering amount of resources that goes into clinical studies (the cost of one study is usually measured in millions of USD), only a few of them successfully reach their end. According to (22), around 10% of the compounds that enter a clinical study are marketed. Moreover, many of the new drugs fail in the final third phase, rather than the earlier ones: the same source estimates that of all the drugs that enter phase III only around 55% are successful. A later study, presented in (33), estimated the success rate to be closer to 70%. These numbers seem to point out a deficiency in the selection process imposed by the different phases: ineffective drugs are carried forward, while possibly effective ones may be dropped. The need for increased efficiency, for optimization, is clear.

This thesis examines several possible ways of approaching optimization of designs of the clinical trials that are conducted in phase III of a clinical study. One such approach

is through application of an adaptive design, i.e. a design the different aspects of which (such as sample size) can be changed during the trial. Some of the tests that are used in adaptive designs, although they control the Type I error rate, may pose a controversy from the point of view of common sense. The first of the papers presented in the thesis treats a variation of such tests and that can be an answer to these concerns. Another issue, lifted, among others, by (1), is the number of active doses that are included in phase III clinical trials. Often, these trials concentrate on the comparison of a single dose with a control. However, this approach may fail due to the insufficient information available about the dose-response from the earlier phases. It has thus been suggested to expand phase III to involve several active doses, even if the goal is to market only one dose. It is the effect of this design alteration that is examined in the second paper. The final aspect of the design studied in the last two papers included in the thesis is the choice of multiplicity correction. Phase III trials often include a number of sub-questions, or hypotheses, the rejection of which depends heavily on the applied multiplicity testing procedure (MTP). These procedures can be seen as aspects of the design which, like other ingredients of the trials, may be optimized prior to the initiation of the experiment. Both papers focus on the effect of such optimization using two different families of MTPs.

Chapter 2: Clinical trials

There are several definitions of the concept of a "clinical trial" (see e.g. (12)). Some of these are quite broad, describing a clinical trial as the process of evaluation of the effect that a drug, or other treatment, has on humans. Others are more narrow, specifying that this process should involve comparisons between a group that receives the new medical treatment and another group that receives some controlling substance (e.g. placebo or a previously existing treatment). Yet another definition suggests that clinical trials are a part of a clinical study that is performed, or sponsored, by pharmaceutical companies with the goal of introducing the new treatment into the market. It is this last definition that will be used in this thesis. That is, "clinical trials" will refer to a series of controlled experiments involving human subjects that should, ultimately, lead to an approval of a new medical treatment. This series as a whole will be referred to as a "clinical study".

Within this framework, the specifics of each clinical trial may vary greatly. A trial may involve tens of thousands of subjects, or less than a hundred. It may have a design as simple as dividing the subjects into two groups corresponding to treatment and placebo, or a complex construction involving real-time re-allocation of patients to different dosages of a drug based on Bayesian schemes. The subjects, although usually part of the targeted population (i.e. the group of people with the treated decease), can also be completely healthy (as is usually the case during the first stage of a clinical study). The "treatment" itself may be a drug, a surgical procedure, a diagnostic test, a combination thereof or something else entirely, including no treatment at all. In this thesis we will often assume that the new treatment under study is a drug that can be administered at different dose levels.

Different trials address different issues that may arise when a new medical treatment is introduced. For example, if this treatment is a drug, then the earlier trials concentrate on locating the proper dose at which it should be administered, while the later ones on evaluating its efficacy in an objective manner. It is this need for an objective evaluation that was the original reason for introducing clinical trials in medical research. Nowadays, both the design and the results of a trial are heavily monitored by parties that are independent of pharmaceutical companies, namely the regulatory agencies such as FDA (Food and Drug Administration, USA) and EMEA (European Medicines Agency, EU).

2.1 A brief history of the clinical trials

As described in (36), the field of medical practice historically was heavily individualized, with the physicians trusting their own expertise above all else. The case by case judgment, rather than attempts at generality, was in focus. The numerical methodologies offered by the proponents of clinical trials (or, as this was the early twentieth century and the notion of "clinical trial" was not yet formed, comparison of proportions of cured in two different populations) were criticized heavily on ethical grounds. Each patient was an individual, and, as such, required an individualized approach. The two-population comparison also presupposed that a treatment that was believed to be effective should be withheld from part of the patients, an action the moral implications of which are debated even now, long after the practice of clinical trials was established.

Thus, the push for objectivity and simple, generalizable facts, came not from the medical community, but from the government. It was the government that sponsored the first large clinical trials in Great Britain (which commenced soon after World War II and studied the treatment for tuberculosis) and it was also a governmental agency, in form of FDA, that began to require each new drug to be examined by means of unbiased experiments.

The FDA, which to this day remains the largest regulatory agency, obtained its power through an act of Congress dated 1938. This act, born from a disastrous launching of an untested drug that caused more than 100 deaths, enabled the agency to veto a new substance from being marketed if the pharmaceutical company developing the substance did not provide sufficient evidence with regards to its safety. Another disaster, this time involving thalidomide, led to another act in 1962 that both strengthened the safety requirements and added a new one regarding the efficacy of the new treatment. Since then, the regulatory requirements for drug approval increased in number, with statistical design and analysis of experiments playing a major part. In a century, the field of medical research changed to be among the scientific areas that employ statistical methodology the most.

2.2 Clinical trials today

Today it takes on average 10 to 12 years for a new compound to enter the market. During this time a series of trials is conducted, with the resulting cost coming up to hundreds of millions US dollars (12). The process is heavily regulated by governmental agencies that require that extensive data on both the efficacy and safety of the new compound should be available. The same agencies also have the power to veto a trial if they judge its design to be inferior. The trials are usually divided into four phases. The first one begins after the end of the pre-clinical development, where the compound has been studied *in vitro* and *in vivo* on animals.

Trials in phase I are very small, involving perhaps 20 to 80 subjects, and have as their purpose the first tentative evaluation of the effect that the compound has on humans. No patients with the targeted disease are usually involved at this stage (there are, however, exceptions, such as cancer trials). Rather, the subjects are healthy volunteers and the main goal is evaluation of the safety of the compound. Other goals of phase I are to evaluate the metabolism, the acceptable dosage, the pharmacological and pharmacokinetic profiles of the drug in humans. In this early phase the involvement of the regulatory agencies is minimal and concerns mainly the safety of the subjects.

In the "traditional" designs for phase I, commonly employed in cancer clinical trials, the subjects receive slowly escalated doses of the studied compound until some prespecified safety criteria are no longer met (38). These designs are sometimes called "3+3 designs", the name referring to the size of the groups of patients that are assigned the same dose, which is usually three. The first group receives the lowest dose. If no toxicity is observed, the second group receives the next lowest dose and so on. Depending on the amount of toxicity in the current group, the next group can be given a lower, a higher or the same dose. This process stops when maximum acceptable dose is found, this dose denoted MTD (maximum tolerable dose).

Phase II is usually several times larger than phase I and involves a few hundred subjects. Now the subjects are patients with the targeted decease, and the purpose of this stage is to assess the efficacy of the drug and to estimate the dose-response profile. The safety concerns are still prominent (as they are during the whole continuation of the study) and information of the possible short-term side effects of the compound is collected. Unlike the first phase, where the regulatory agencies mainly require that the subjects are not exposed to unnecessary risks, in the second phase they control the design of the trial, with the requirement that this design will likely produce the data necessary to evaluate the drug.

The simplest, and arguably the most common, design used for dose-response estimation is the parallel group design with several groups assigned to dosages that, from earlier studies, were found to be tolerable with respect to toxicity. In early phase II clinical trials different dose escalation designs may also be used, with the titration design being a classical example. According to this design, all patients start at the same low dose level. If positive response to treatment is observed, the patient continues receive the same dose throughout the trial. If not, the dose is escalated to the next level. Given that no adverse events present themselves, the process continues until either all dose levels are exhausted of a positive effect is observed.

Phase III is confirmatory and has as its main goal ascertaining the efficacy of the drug, as well as capturing more rare or long-term adverse effects. It usually involves several hundred to several thousand patients. At this point of the study, the main objective is often as simple as proving the superiority of the new compound at a certain dosage to a comparator (e.g. placebo or existing treatment) and the statistical design strives to be

uncomplicated. Parallel group designs are common, but more sophisticated approaches, such as cross-over designs, can also be employed. Additional complications may arise if a non-standard randomization scheme is used, e.g. randomization in clusters.

The standard requirement for a drug to be marketed is its demonstrated success in two confirmatory trials. Even then, the monitoring of the drug continues into a phase IV clinical trial. The purpose is to find the very rare adverse events, to evaluate the effect of the drug on morbidity and mortality and to find previously undetected sub-populations for which the effects of the drug may be different.

2.3 Attrition rates

In practice, the process of drug approval outlined above is neither smooth nor straightforward. An often cited paper by Kola and Landis (22) gives the attrition data for ten pharmaceutical companies, years 1991 to 2000. Among other findings, they note that on average only 10% of the drugs that enter the clinical studies (i.e. phase I) are marketed. They also note that the majority of the compounds are stopped during phase II and III rather than phase I. For the supposedly confirmatory phase III, the attrition rates, although different for different therapeutic areas, vary around 45%. That is, on average approximately half of the drugs that entered this phase during the examined period finished it successfully. The authors also give a chart that groups the reasons for attrition. In the year 2000, the two major causes for stopping further development of a drug were lack of efficacy (30% of the stopped compounds) and safety issues (even those around 30%).

A more recent work addressing this issue was written by Paul et al. (33). The authors state that some sources predicted "imminent demise" of the pharmaceutical industry, and that it can not survive without "dramatic increase in R&D productivity". The authors examine key points that would contribute the most to this increase, even they identifying the success rates of phase II and III as being crucial (although they estimate the attrition rate in phase III to be closer to 30%). They also point out that the corresponding attrition rates are increasing. The authors discuss several reasons for this phenomenon, among others naming unprecedented nature of drug targets and greater safety requirements.

2.4 The setting of the thesis

This thesis focuses on the design and analysis of phase III clinical trials, although the results, especially those presented in the multiplicity section, are applicable in other, similar, situations. The setting of phase III clinical trial involves the confirmatory testing of one ore several hypotheses, often of the type $H_k : \theta_k = 0$ with θ_k denoting the parameter of interest and referring to a population average. A typical example is the comparison of a

treatment arm to a control arm. The hypotheses tested are often one-sided, as usually an ordering between θ_k is assumed, as is e.g. the case with treatment vs control, or different dose levels. However, the Type I error rate is commonly taken to be the conservative 0.025 rather than the conventional 0.05. This is done in order to to control the error rate at the 0.05 level as if the test was two-sided (even if we are interested only in the positive alternative in practice). Since phase III usually involves a large number of patients, the assumption of (multivariate) Normally distributed test statistics is not unrealistic.

Chapter 3: Adaptive designs

Adaptive designs have gained increasing popularity in the context of clinical trials during the later years. The term "adaptive design" refers, generally, to division of the process of the data collection in several stages, where the specifics of the data collection or statistical analysis at the later stages may change based on the data accumulated up to the current stage. More formal definitions have been given in e.g. (18), where the authors described an adaptive design as a "clinical study that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial". In recent years several books have been written on the subject, two of them being (11) and (35).

The appeal of such a design in clinical trials is natural. These trials are often sequential, in the sense that the enrolment of patients is a process that is stretched out in time, sometimes over several years. As a consequence, even the data associated with the effects of the treatment becomes available gradually. Another reason are the ethical considerations. An adaptive design would allow to switch patients to a treatment that, according to the accumulated data, is the most effective. In addition, the great cost associated with a phase III clinical trial makes the option of stopping it earlier than planned, if sufficient evidence in favor of the treatment were gathered, very desirable.

A multitude of different kinds of adaptive designs exist. The so called adaptive randomization designs assign new patients to different treatment groups with unequal probabilities, those probabilities being a function of the accumulating data. The treatment-adaptive, covariate-adaptive and response-adaptive designs (discussed in e.g. (12)) fall into this category. Another broad class, sometimes employed in the early phases, are the adaptive dose-finding designs. These designs often are based on the variations of continued reassessment method (32), assuming a model for dose-response (or dose-toxicity) with uncertain parameters and updating the information about these parameters using Bayesian schemes. Examples of their use is determination of the MTD (see e.g. (38)) and location of the optimal dose with respect to a measure that incorporates both efficacy and safety (see e.g. (44)). This thesis, however, touches on the topic of adaptive designs that can be applied in later phases of clinical development and that concentrate on the control of the Type I error rate. These designs are commonly divided into two broad, and not entirely separate, classes: group sequential and flexible designs.

Both of these classes share the same basic construction: the data is collected in blocks of possibly different sizes, with an "interim analysis" performed in between. During this analysis, a decision regarding the remaining blocks is made. In the original group sequential designs, which have been introduced far earlier than flexible ones, this decision was simply stop or go. This option of stopping the data collection earlier than planned is the primary feature of a group sequential design. The reason for stopping can be efficacy (when it is judged that the data collected so far is enough to reject the hypothesis of no treatment effect) or futility (when it is judged that the data collected so far indicates that no clinically significant effect exists). Different stopping rules, or rejection boundaries, exist, the most popular being the ones introduced by Pocock (34) and O'Brien-Fleming (31). In time, the group sequential designs evolved, allowing additional adaptations such as sample size re-estimation (35). However, it is necessary to specify all data-dependent adaptation rules in advance. This necessity is a major point of difference between the group sequential designs and flexible designs with sample size re-estimation that are considered in paper I.

3.1 Flexible designs

Flexible adaptive designs, similarly to group sequential ones, allow for early stopping and change of the sample size. However, there are two important differences. First, the way the design is altered need not be specified in advance. Second, while the group sequential designs focus on the sample size of the trial, the flexible designs allow for a much wider range of possible adaptations, among which are sample size re-estimation, dropping/adding dose levels, focusing on a sub-population and change of statistical hypotheses (e.g. from superiority to non-inferiority).

The flexible designs themselves can broadly be divided into two approaches, which are, essentially, equivalent but were introduced independently of each other and are thus formulated in different ways. The first approach utilizes a so called combination function, while the second is build around a "conditional error probability".

The first approach, introduced in (2), constructs a test statistic at the end of data collection which is a function of the test statistics that were calculated separately for each of the stages. That is, the final test statistic is a combination of the stage-wise test statistics (alternatively, p-values), hence the name of the methodology. Assuming the stage-wise p-values to be independent and $Uniform(0,1)$ if the null hypothesis is true, a combination function can be chosen in such a way that the distribution of the final test statistic under null hypothesis known. The combination function is usually denoted by C. That is, for a two-stage design with p_1 and p_2 the corresponding stage-wise p-values the final test takes the form $C(p_1,p_2) \leq c_\alpha$ or $C(p_1,p_2) \geq c_\alpha$, with α the Type I error rate and c_α the corresponding quantile of the distribution of $C(p_1,p_2)$.

Several combination functions have been suggested in the literature. In the original pa-

per the authors concentrated Fishers combination test. Another popular choice, discussed in detail in (23), is the so called "normal inverse" combination function. As the name suggests, this function consists of the weighted combination of normal inverse functions applied to the stage-wise p-values. The weights are determined in advance and are chosen in such a way that the sum of all the weights across the stages equals unity. Usually, the weights correspond to the proportion of the planned sample size per stage. The attraction of this particular formulation stems from the fact that this combination function will lead to a $N(0,1)$ distributed final test statistics if the stage-wise test statistics are $N(0,1)$. Another advantage is that, if no adaptations are made, then the final test statistic will correspond to that obtained from a group sequential design.

Although the combination function approach was originally suggested for the case with only two stages, the generalization to any number of stages is straightforward. Further generalizations are offered through the observation that, in a two-stage design, the second stage can be viewed as a separate design, possibly itself consisting of several stages. This gives rise to the "recursive combination tests" (4), where a multi-stage design is constructed by recursively applying, possibly different, combination functions to the stage-wise p-values.

The conditional error probability approach (37; 26; 27) considers the same idea from a different perspective. The focus is on the so called conditional error function and the probability of rejecting the null hypothesis given the data from the current stage. In order to understand this approach and its connection to the previously discussed combination functions, consider a two-stage adaptive design. Say that we choose a combination function in such a way that the distribution of the final test statistic is $Uniform(0,1)$. That is, we reject the null hypothesis if $C(p_1,p_2) \leq \alpha$. This rule can be illustrated by plotting the corresponding rejection region in terms of p_1 and p_2. Such an illustration, using the normal inverse combination function, is presented in Figure 3.1, where the rejection region is called "A". Thus, the rejection rule can be formulated as "reject the null if $(p_1,p_2) \in A$". The boundary of the region can be described by a function, call it $f(p_1,\alpha)$. Using this formulation, we have an alternative way of defining the rejection rule, namely "reject the null if $p_2 \leq f(p_1,\alpha)$". It is the function $f(p_1,\alpha)$ that is usually referred to as the conditional error function. The terminology stems from the fact that $f(p_1,\alpha)$ can be viewed as the amount of the Type I error that is allowed in the second stage given the data from the first stage (i.e. p_1). Even the conditional error approach is straightforwardly generalized to the case with arbitrary number of stages.

3.2 Some issues associated with flexible designs

Flexible designs provide the option of adapting virtually any aspect of the trial during the trial. They are entirely correct from the statistical point of view, in the sense that, assuming independent uniformly distributed stage-wise p-values, they do preserve the

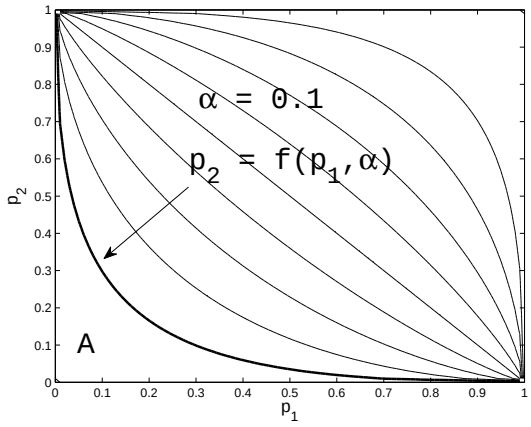

Figure 3.1: Illustration of the rejection rules for a flexible adaptive design. Normal inverse combination function with equal weights and $N(0,1)$ distributed test statistics is used. The contours represent different levels of $C(p_1, p_2)$ with the thick line corresponding to $C(p_1, p_2) = \alpha$, which we choose to be 0.1. The rejection region is in the lower left corner below the thick line, and is denoted by A.

required Type I error rate. However, they have been questioned regarding the soundness of the conclusions that applying such a design can lead to (8) if the adaptation performed is sample size re-estimation. That is, whether they really do provide an adequate test of the null hypothesis in question.

The reason behind these doubts can be discerned by looking at the rejection region of a flexible design, for example a two-stage one employing the normal inverse combination function. This combination function can be explicitly defined as $C(p_1, p_2) = 1 - \Phi\big(\sqrt{w_1}\Phi^{-1}(1-p_1) + \sqrt{w_2}\Phi^{-1}(1-p_2)\big)$ with p_k denoting the stage-wise p-values and w_k the normalized weights associated with the stages. We are going to assume that w_k represent the proportions of the planned sample size per stage. The null hypothesis of no positive effect is rejected if $C(p_1, p_2) \leq \alpha$. Observe that if no sample size re-estimation is made, then this test is equivalent to the common test of mean that would be performed in a non-adaptive setting.

Assume that it was initially planned that the interim analysis will be exactly in the middle of the trial, i.e. $w_1 = w_2 = 0.5$. In this example no stopping rule is incorporated, i.e. the trial continues to the second stage regardless of what data was obtained in the first stage. From the shape of the rejection region (see Figure 3.1) we can see that it is possible to reject the null hypothesis for widely differing p_1 and p_2, in the extreme cases $p_1 \approx 0$ and $p_2 \approx 1$ (or $p_2 \approx 0$ and $p_1 \approx 1$). This, in itself, is acceptable, if no sample size re-estimation is performed. However, the situation becomes more complicated if there is

such an adaptation and the two p-values may be calculated using (normalized) stage-wise sample sizes that are completely different from the w_1 and w_2 planned originally.

Consider the following exaggerated example. Let us say that after obtaining a p_1 close to 1 we would like to stop the trial for futility. However, since no stopping rules are pre-planned, we proceed to the second stage, but take only one observation, on which we base our p_2. It so happens that this particular patient displays a very clear positive response to the treatment and we get a p_2 that is close to 0. According to the flexible design the null hypothesis will be rejected. However, from the practical point of view, the fact that this rejection is based on one observation is highly questionable.

In short, flexible designs may give unequal weights to equally informative observations, breaking against the sufficiency and conditionality principles (13). This makes the analysis potentially unstable and dependent on the outliers. It may be argued that the extreme cases as in the example above can be easily handled by introducing stopping boundaries. This is, to a degree, true. However, the fact remains that the flexible test may reject the null hypotheses in situations, where, should we apply a simple, non-adaptive test to the same data set, the null hypothesis would not be rejected. This discrepancy gave rise to a dispute of whether the rejections based on the flexible test alone can be trusted.

One possible way to address this issue, studied in the first paper included in this thesis, is to use the so called dual test (first introduced in (14)). This test uses the minimum of the flexible and the simple test statistics, i.e. it would reject a null hypothesis only if it is rejected by both the flexible test and the one that would be performed on the data if the fact that adaptations were made was ignored. Since this test is stochastically smaller than the flexible one (the rejection region of the dual test is a subset of that of the flexible test) Type I error rate will be preserved. However, for the same reason, the dual test is less powerful. The question studied in the paper is how much power is lost when using the dual test as compared to the flexible one.

3.3 Summary of paper I

In this paper various properties of the dual test as compared to a one that uses a normal inverse combination function were studied. A two-stage design was assumed, with sample size re-estimation at the interim. The power of the tests (total and conditional) was used as the primary criterion of optimality. With respect to this criterion, an analytical expression for the "no power loss" region was derived. The name refers to possible sample size re-estimation rules that are based on the test statistic from the first stage and for which no power is lost. We proceeded to calculate the power loss resulting from the use of the dual test for several sample size re-estimation rules that may be considered typical. For example, rules aimed at preserving a certain conditional power (probability of rejecting null hypothesis given the interim data) were considered. The difference in the expected total sample sizes was also calculated. It was found that, in many cases, the

power loss associated with the additional requirement the dual test imposes is minimal, and often will not occur at all. The sample size re-estimation rules where this loss reached its maximum were characterized by a decrease of the sample size for relatively small p_1, i.e. situations reminiscent to the one described in example above, if not as extreme.

3.4 The number of doses in phase III clinical trials

Adaptive designs with their option of sample size re-estimation can be used as a safeguard in phase III clinical trials. If the results observed in phase II lead the researcher to choose a dose that is too low, then the total sample size could be increased at the interim and the null hypothesis rejected. As is evidenced by the attrition rates discussed earlier, the uncertainty in the effect of a drug remaining after phase II is large enough to make such an option useful. Another suggestion, examined by e.g. (1), is to allow more than one active dose (namely two) to enter the confirmatory phase. The motivation for this suggestion is clear: intuitively, when taking two active doses, the probability of seeing an effect for at least one of these should be larger than when only one active dose is considered.

It is the impact of this possibility on the probability of success of the trial that is examined in paper II. There, two scenarios are compared. The first one is a classical "single active dose and a control", the second allows for two active doses and a control. The goal was to find optimal designs (e.g. dose placement) for these two cases and compare them with respect to probability of successfully proving efficacy for at least one of the active doses. Unlike paper I, where the focus was on efficacy alone, paper II introduces a new aspect of possibility of unacceptable safety issues, the probability of which increases with dose level. In order to make a fair comparison between the two scenarios, we opted to fix the total sample size to be the same in both of these cases.

3.5 Summary of paper II

The paper begins by introducing models for efficacy and safety profiles that can be used for design of phase III clinical trials. The efficacy model was constructed using a traditional E_{max} curve. The safety model consisted of a certain probability of unacceptable safety being observed at a particular dose level, this probability described by a logistic function. These two were incorporated into a probability of success, which was then used as a measure of optimality of a trial. As mentioned above, two scenarios, one with one active dose, another with two, were considered and optimized with respect to different design features. In the case of a design with one active dose, the only such feature was the dose placement. When two active doses were considered, additional factors such as sample size per dose arm and different multiplicity corrections were examined.

The condition of having the same total sample size for both design types lead to the

design with one active dose being optimal if only a small amount of uncertainty in the models was present (i.e. the data from phase II was sufficient to estimate both the efficacy and safety profiles with good precision). However, if moderate uncertainty in the efficacy and safety profiles was assumed, and if the requirements for the power of the trial were high, the design with two active doses lead to a larger probability of success. It was also found that factors such as multiplicity correction have a large impact on the probability of success of a design with multiple doses, which inspired the work presented in paper III and IV of this thesis.

Chapter 4: Multiplicity

4.1 Possible errors

The concept of Type I error rate, commonly denoted by α, was first introduced in a series of papers by Neyman and Pearson published around 1930 (28; 29; 30). Together with the competing Fisherian approach of p-values, it has, over time, formed the hybrid theory of statistics that is considered to be the foundation of the field. Nowadays, the control of Type I error rate is the primary property that a testing procedure is required to possess.

The Type I error rate is, per construction, a measure applicable only when one single hypothesis is considered. Multiple hypothesis testing, however, involves consideration of several hypotheses simultaneously. A measure similar to the Type I error rate, but designed for this, more general, setting, is desirable. But how, exactly, the concept of Type I error rate should be altered, and whether it should be altered at all, is far from obvious.

The issue was much debated in the second part of the twentieth century, and is still a source of controversy. A wonderful summary describing these debates in allegorical form can be found in (10). One school of thought argued that any alterations of the Type I error rate are unnecessary. Rather, each of the hypotheses should be tested as if no other hypotheses were considered. This point of view was based on the fact that, if these hypotheses have no connection to each other in the sense that each of them leads to a separate conclusion, then it is not logical to change the testing process just because these conclusions happen to be drawn simultaneously. The argument is not unreasonable and has much appeal to practicing researchers, who often strive to answer several questions using a single experiment. This measure of error rate is known as "per-comparison" error rate.

A completely different approach advocated that, when a whole family of hypotheses is considered, then it is natural to control the probability that at least one true hypothesis in this family is rejected. This probability came to be known as family-wise error rate (FWER). The application of the idea required that the significance level for each of the null hypotheses should be re-calculated, often leading to a smaller α used for each of the individual tests than the one that was chosen originally. This, in turn, often has the consequence of greatly reducing the probability of rejecting a particular hypothesis.

The FWER approach, which was much less popular among the performers of experiments due to the loss of statistical power, has over time gained acceptance. It is regarded to be appropriate when the whole experiment, rather than individual hypotheses, are of interest. It should also be used when a hypothesis that serves as a basis for a conclusion has been selected based on the obtained data, as is e.g. the case when several doses of a drug are compared to a control and only the most promising is chosen for marketing (16).

These two approaches, ignoring the multiplicity and controlling the FWER, lie at the opposite ends of a spectrum describing how much error a researcher is willing to tolerate within a family of hypotheses. There is a number of situations where neither of them is appropriate, e.g. when it is desired to control the error but the number of hypotheses is too large for the FWER approach to be feasible. Several altered procedures that compromise between FWER and the comparison-wise error rate have been suggested. One example is the false discovery rate, defined as the expected ratio between the number of false positives and the total number of rejections (7; 15). Today, the usual recommendation for a practicing statistician is to chose the error measure according to their own discretion, based on the particular situation at hand and the different practices that, over time, have been established in different research areas.

The subject of this thesis is the statistical analysis of phase III clinical trials, i.e. confirmatory testing. It is thus plausible to believe that all, or at least most, of the null hypotheses tested are false. At the same time, the error of making even one incorrect rejection has great consequences. The error rate is controlled in the strictest possible sense using the FWER. It is thus this measure that is used in papers III and IV.

The concept of p-values can also be generalized to a multivariate setting. This generalization is quite intuitive, and, unlike the concept of Type I error rate, poses no controversy. In this thesis, let us consider two types of p-values: the unadjusted p-value resulting from the test of this hypothesis ignoring the other hypotheses in the family and the adjusted p-value. The latter is defined as the smallest α for which a hypothesis would have been rejected after applying a particular FWER controlling procedure (40).

4.2 Power and utility

For a test of one single null hypothesis the concept of power refers to the probability of correctly rejecting this hypothesis if it is false. There are several ways to generalize this idea to multiple hypotheses setting. One of the most natural such generalizations (and one of the most common) is the so called disjunctive power (7). In the same spirit as the FWER, the term refers to the probability of correctly rejecting at least one hypothesis in the family. Another generalization is the conjunctive power, which is the probability of correctly rejecting all false hypotheses. A concept uniting the two is the probability of rejecting at least k hypotheses correctly. Yet another measure is the average, or total, power, which is the expected average number of correctly rejected hull hypotheses. It, in

turn, can also be generalized to the weighted average power, which is the expectation of the weighted average of the number of correct rejections.

In the two multiplicity papers in this thesis a tool that encompasses all these concepts, namely the expected utility, is used. Let us define the utility to be a function mapping the space of all possible combinations of acceptances/rejections of the null hypotheses included in a family onto the real line, i.e. $U : \{0, 1\}^K \to \mathbb{R}$, with K denoting the number of hypotheses in the family. That is, let us assign a value to each possible constellation of rejections and acceptances. This value is subjective, different for different situations, and could be based on e.g. the perceived benefit of proving a drug to be effective. For simplicity, let us work with a normalized utility that has the minimum value of 0 and maximum value of 1. Note that this formulation does not include a direct dependence of the utility function on the parameter of interest. Assuming a setting similar to that of a phase III clinical trial, i.e. a one where all the hypotheses are expected to be false, such a formulation is not unreasonable. However, if the purpose of the testing is of a more exploratory nature, some of the hypotheses are expected to be true. In this case it may be desirable to re-define the utility function to encourage not only the rejection of the false hypotheses, but also the failure of rejecting the true ones. That is, to make it directly dependent on the tested parameters. A short exploration of this possibility, giving a possible general approach and some examples, is presented in Chapter 4.8.

The measure of the strength of a testing procedure is then defined as the expectation of the utility function, with the expectation taken with respect to the distribution of the test statistics (alternatively, p-values) that correspond to each of the hypotheses. Observe that, although in the papers presented later, the parameter of interest is usually assumed to be fixed (or, rather, the calculations will be performed given a fixed alternative), this need not be the case. Let us say that we are uncertain about the true value of the model parameters, but that we have some idea about what they are. It may then be possible to describe this uncertainty through a Bayesian prior placed on the parameters. Our goal will change to be maximum expected utility with the expectation taken with respect to the distribution of the parameters as well as the test statistics.

As previously stated, the concept of expected utility encompasses the different definitions of multivariate power mentioned above. This can be realized through understanding that each of these definitions can be expressed in terms of the unions of the disjoint events of accepting a certain number of hypotheses and rejecting the rest, i.e. the domain of the utility function. For example, the disjunctive power, the probability of rejecting at least one hypothesis, corresponds to a utility function that is equal to unity everywhere except for the event of accepting all hypotheses. That is, $U = 1$ for all possible outcomes of the experiment where at least one hypothesis is rejected. For the conjunctive power we have $U = 0$ everywhere except for the single point in its domain that corresponds to the event of rejecting all hypotheses. In this single point $U = 1$. The utility function for the average power assigns the value $1/K$ to the events where one hypothesis is rejected

and the rest accepted, $2/K$ to the events where two hypotheses are rejected and the rest accepted and so on, with the maximum utility, $U = 1$, attained if all hypotheses are rejected. The weighted average power is a generalization thereof, where $1/K$ is replaced by another weight, chosen in such a way that the sum of the K different weights equals to one.

4.3 Multiple testing procedures

The need to control the FWER leads to application of some multiple testing procedure (MTP). Although this term is quite broad, in this thesis we are going to use it to refer to the set of procedures that are based on the unadjusted p-values (alternatively, test statistics) obtained by testing $H_1, ..., H_K$. The most well-known example of these procedures is Bonferroni. According to this MTP, a hypothesis k is rejected if $Kp_k \leq \alpha$, with p_k denoting the unadjusted p-value corresponding to H_k. The product Kp_k then becomes the adjusted p-value (since this is exactly the smallest α for which the hypothesis could be rejected). This procedure preserves the correct FWER due to Bonferroni inequality.

This simple Bonferroni procedure is an example of a single-step MTP, i.e. a MTP where whether a hypothesis k is rejected or not does not depend on the rejection of other hypotheses. One of the most well-known multi-step procedures, which are procedures where such a dependence does exist, is the Holm procedure (21). This MTP orders the unadjusted p-values according to their magnitude, from smallest to largest. The smallest p-value is multiplied with K, and the corresponding hypothesis is rejected if this product is less than α. That is, the first step of Holm MTP is the same as that of Bonferroni. If the hypothesis is rejected, the next smallest p-value is multiplied by $K - 1$ and compared to α and so on. The last p-value is then compared to α directly. This test controls the FWER in the strong sense since it can be viewed as a closed testing procedure (see below).

The Holm procedure is a so called step-down MTP (the name referring to the order in which the p-values are tested, from smallest to largest). An example of a step-up procedure is the Hochberg procedure (20). This procedure also orders the unadjusted p-values according to their magnitude, but now the largest p-value is tested first by comparing it to α. If this p-value is smaller than α, then all hypotheses can be rejected. If not, then the next smallest p-value is multiplied by 2 and compared to α and so on. Even this procedure can be formulated in terms of closed testing and thus preserves the correct FWER.

The three procedures above make no assumption of the exact distribution of the test statistics (although the Hochberg procedure assumes a certain type of dependence (39)). However, there are also MTPs that make such assumptions, two of the most common ones being the Tukey and the Dunnett MTPs (7). The Tukey procedure is used for pair-wise comparisons of means when the test statistics follow a multivariate t distribution. The Dunnett procedure is used when a set of means is compared to a common baseline (e.g.

several treatments against the same control), again assuming a multivariate t distribution. Both procedures make use of the known correlation between the test statistics that arises from the experimental design.

4.4 The principle of closed testing

Many of the common MTPs can be viewed as so called closed testing procedures. The principle of closed testing was first introduced by Marcus et al. (25) and is a general methodology that can be used in the construction of multiple testing procedures that control FWER. This methodology can be described as follows.

Assume that a family of null hypotheses $H_1, ..., H_K$ is to be tested while preserving FWER α. Construct hypotheses intersections of all possible subsets of $H_1, ..., H_K$, i.e. look at the elements of the type $H_C = \cap_{k \in C} H_k$ with $C \subseteq \{1, ..., K\}$. For example, if $H_k : \theta_k = 0$, then $H_{\{12\}} = H_1 \cap H_2 = \{\theta_1 = 0\} \cap \{\theta_2 = 0\}$. Test each of the H_C using some test that preserves Type I error rate. The exact form of this test is not specified. For example, if the θ_k above are the means in a Normal distribution, this test could consist of an ANOVA. A particular null hypothesis H_k can be rejected only if all H_C such that $k \in C$ are rejected. A closed testing procedure usually starts by testing the global intersection, i.e. the intersection of all K hypotheses. If it is rejected, H_C such that $|C| = K - 1$, with $|C|$ denoting the number of elements in C, are tested and so on, down to the simple, or elementary, hypotheses $H_{\{k\}}$. For this reason, the procedure is sometimes referred to as hierarchical testing.

Any multiple testing procedure constructed in this way will preserve the correct FWER. To understand this, consider the simple case when only two hypotheses, H_1 and H_2, are being tested. The following situations are possible: (1) both are false; (2) both are true; (3) one is true and the other is false. If (1) is correct, then no error is committed when any of the hypotheses is rejected. If (3) is correct, e.g. if H_2 is true, then it would be correct to reject $H_{\{12\}}$ and a Type I error is committed only if $H_{\{2\}}$ is rejected during the second step of the procedure, the probability of which is α. Finally, if (2) is correct, then rejecting $H_{\{12\}}$ would be erroneous. The probability of this happening is, again, α. Since in order to make further rejections we have to test $H_{\{1\}}$ and $H_{\{2\}}$ as well, the probability of making at least one false rejection is smaller than the chosen Type I error rate. The same idea can be applied to the general case with K hypotheses. A more thorough description of the procedure can be found in e.g. (7).

As stated above, both Holm and Hochberg procedures can be formulated in terms of closed testing (although they were not constructed using closed testing principle originally and the connection was made later). While based on the same principle, they use different tests for the intersection hypotheses. The Holm procedure results from H_C being tested using the so called Bonferroni test, according to which H_C is rejected if $\min_{k \in C} p_k \leq \alpha / |C|$. To see this, consider again the simple example with two hypotheses. We will reject $H_{\{12\}}$

if $\min(p_1, p_2)$ is smaller than $\alpha/2$, which corresponds exactly to the first step in the Holm procedure. Say that $p_1 < p_2$. If $H_{\{12\}}$ is rejected, we proceed to testing $H_{\{1\}}$, which can be rejected if $p_1 \leq \alpha$. However, since the rejection of $H_{\{12\}}$ is a requirement, we have that $p_1 \leq \alpha/2 < \alpha$. That is, if the intersection hypothesis is rejected, the corresponding elementary hypothesis will be rejected as well, and it is unnecessary to test it explicitly. Finally, $H_{\{2\}}$ is rejected if $p_2 \leq \alpha$, and we have the Holm procedure.

The Hochberg procedure can be obtained if Simes test (41) is used for testing H_C. This test is more powerful than Bonferroni but it assumes a certain type of dependence between the test statistics, and this assumption is carried over to the Hochberg procedure. The closed test MTPs that use Bonferroni and Simes tests possess an important feature that is used in construction of Holm and Hochberg procedures, namely consonance: if H_C is rejected, then at least one of the original hypotheses $H_k : k \in C$ is rejected as well.

4.5 Recycling procedures

For many of the above mentioned multiple testing procedures there exist generalizations that allow incorporation of weights (3). The purpose of these weights is to make a MTP more effective by using additional information that may be available, such as the relative importance of the different hypotheses. For example, a weighted Bonferroni procedure would compare p_k/w_k to α, rather than Kp_k. If the rejection of H_k is of more value to the researcher than the rejection of other hypotheses, it may be natural to choose $w_k > w_i$, with $i \neq k$ since this will make the probability of rejecting H_k greater. Similar extensions exist for Holm and Hochberg procedures (3).

Such weighted generalizations allow for greater adaptability of the MTP to a particular situation. Lately, two even more flexible approaches have been introduced in (5) and (9). Hereafter, the approach described in (5) will be referred to as "graphical procedures" and the approach described in (9) as "recycling procedures". These approaches, although they have been constructed independently of each other and thus have different formulations, rely on the same idea of recycling of the test mass.

The term "test mass" in this context refers to the Type I error rate, i.e. the magnitude of error that is allowed when testing a family of null hypotheses. According to the graphical and recycling procedures, this test mass can be distributed between the hypotheses in different ways. For example, if $w_k\alpha$, $\sum w_k = 1$, is assigned to each of the H_k it will result in the weighted Bonferroni MTP. The test mass can also be re-assigned, or recycled, if a hypothesis has been rejected. That is, if the rejected hypothesis is H_k, then the proportion of α given to H_k can be moved to other hypotheses. Observe that once a test mass α_k has been placed on H_k it can not be re-assigned unless H_k is rejected.

The two procedures are, in their original form, a combination of closed testing principle with the weighted Bonferroni test of the intersection hypotheses (although later they have evolved to include other tests, see (6)). The graphical procedures are formulated in

terms of a vector of weights w reflecting the initial distribution of the test mass among the hypotheses and a $K \times K$ matrix G describing how the test mass should be transferred in case of a rejection. Thus, a graphical procedure is fully specified by $(K-1)+K(K-2)$ parameters. If $K \geq 4$ this number is smaller than $\sum_{k=2}^{K} \binom{K}{k}(k-1)$, which is the number of parameters required for fully specifying a closed testing MTP where each intersection hypothesis is tested with Bonferroni test. Thus, a graphical procedure, although very general, is still a restriction of the full closed testing procedure. The recycling procedures, in their most general form, can be described through sequences of possible rejection orderings, where each ordering is assigned a weight. This can also be understood as follows: according to a recycling procedure, after H_k has been rejected the corresponding test mass $w_k\alpha$ can be re-distributed between the remaining hypotheses arbitrarily. That is, the "recycling rule" is not constrained to follow the one that was initially specified, and this procedure is more general that the graphical one.

Several well-known MTPs can be viewed as special cases of the graphical and the recycling MTPs. For example, in terms of graphical procedures, weighted Holm can be obtained by, first, placing initial weights $w_1, ..., w_K$ on the hypotheses, and, if, say, H_k is rejected, re-distributing the test mass among the remaining hypotheses $H_j : j \neq k$ proportionally to the corresponding weights w_j. Another example is the fixed sequence procedure according to which the hypotheses are tested in some predetermined order at full level α. It can be seen as a graphical procedure that assigns all test mass to a certain hypothesis, and, if this hypothesis is rejected, moves the test mass to the next hypothesis. The fallback procedure and some forms of gatekeeping can also be formulated as recycling or graphical procedures.

4.6 Choosing an MTP

The recycling procedures allow even greater freedom in adapting an MTP to a particular situation than the simpler weighted ones. However, how this freedom should be exploited, i.e. how the weights should be chosen, is far from clear. Intuitively, the weighted procedures should be useful when rejecting some hypotheses is perceived to be more important than others. When presenting the weighted Holm procedure the author stated that the weights may reflect the relative importance of different hypotheses, the greater the perceived value the greater the weight. Later, it has been suggested to optimize the weighted procedures with respect to some definition of multivariate power, the (weighted) average power being the measure chosen most often (see e.g. (42; 3; 45)). Uncertainty in parameters has been considered in (46), where the authors assigned a Bayesian prior to the parameters and optimized the expected number of rejections.

We suggest to combine these ideas by introducing a formal decision theoretic approach into the context. That is, we suggest, for each situation, to explicitly define a utility function as described in Section 4.2. We can then optimize the expected value of this

utility within a family of multiplicity correction procedures, with the expectation taken over the distribution of the p-values/test statistics. The first paper on the subject, paper III the summary of which is presented below, studies this approach when the family of MTPs in question is the recycling procedures. The second paper, paper IV, applies the same idea to the family of closed testing procedures that utilize pooled tests of the intersection hypotheses.

4.7 Summary of paper III

The focus in this paper was on optimization of different Bonferroni based MTPs. Both the possible gain in expected utility and the form that an optimal MTP takes for different utility functions were examined. Several families of MTPs were considered, among others the recycling, weighted Holm and weighted Bonferroni procedures. We formally described the suggested optimization approach and gave algorithms and mathematical formulas that can be used to calculate the expected utility of these procedures. We proceeded to compare the optimal procedures within these classes, both to each other and to other popular MTPs, such as fixed sequence. Several different utility functions were used for this comparison. One of those was a linear utility that corresponds to the weighted average power. Another was a non-linear utility that reflects a requirement that a certain subset of the hypotheses should be rejected in order to be able to reject others (corresponding to a parallel gatekeeping procedure). We examined the case where the test statistics were independent as well as the case where dependence was present. It was found that, for linear utilities and Gaussian test statistics that were weakly correlated, the additional complexity of a recycling procedure gives no discernible advantage over the simpler weighted Holm procedure. However, if the test statistics are strongly dependent, the gain in expected utility increases. For more complex utilities that in a natural manner lead to gatekeeping procedures the recycling MTP has a clear advantage. This advantage increases when the number of hypotheses considered becomes larger. Different constellations of the tested parameters were examined, illustrating the cases in which the optimal MTP approaches a sub-optimal one, such as unweighted Holm or fixed sequence.

4.8 Generalizing utility

The utility function as presented in Section 4.2 was defined as a function of all possible combinations rejections and acceptances of the null hypotheses, but was not directly dependent on the tested parameters. Such a formulation may not be feasible if some of the parameters are, in fact, expected to lie in the space defined by the null hypotheses, as may be the case if some of the hypotheses are of a more exploratory nature. Thus, including such a possibility in the utility formulation is desirable. Let us examine the

effect of such generalizations of utility using the same setting as in paper III.

A natural way to proceed would be to model the parameter vector θ itself as a random variable, with a corresponding probability measure that we, in the spirit of the terminology of Bayesian analysis, refer to as "prior". This prior can be constructed in such a way that the probability of the components of θ belonging to the space defined by the null hypotheses is larger than 0. For example, say that $H_k : \theta_k \leq 0$ is tested against $H'_k : \theta_k > 0$ and it is believed that one of the parameters, θ_1, is 0 with probability $p > 0$. In this case we can introduce a two-point prior on θ defined as $\mathbb{P}((\theta_1, \theta_2, ..., \theta_K) = (0, \theta'_1, ..., \theta'_K)) = p$ and $\mathbb{P}((\theta_1, \theta_2, ..., \theta_K) = (\theta'_1, \theta_2, ..., \theta_K)) = 1 - p$ with θ'_k denoting a positive constant. Observe that such a prior may even be used to include logical relationships between null hypotheses. If, for example, H_1 and H_2 exclude each other, i.e. $\theta_1 \leq 0 \Rightarrow \theta_2 > 0$ and vice versa, then such a restriction can be incorporated by introducing a prior consisting of two points: $(0, \theta'_2, ..., \theta'_K)$ and $(\theta'_1, 0, ..., \theta'_K)$.

The utility function will then take on the form $U(I, \theta) : \{0, 1\}^K \times \theta \rightarrow \mathbb{R}$, with I denoting the vector indicating which hypotheses have been rejected. We will still strive to maximize the expected value of the utility with respect to a family of MTPs, but now the expectation will be taken with respect to the multivariate distribution of (I, θ). That is, we are looking for

$$\text{argmax}_{\text{MTP}} \mathbb{E}_\theta[\mathbb{E}_{I|\theta}[U(I, \theta)]] = \text{argmax}_{\text{MTP}} \int_\theta \int_{I'} U(I', \theta) \mathbb{P}(I = I'|\theta) dF_\theta$$

with I' denoting a particular value of I and F_θ the prior distribution function of θ. Arriving at this generalized formulation is rather straightforward, but defining a multi-dimensional utility and calculating the corresponding expectation in practice may be far more difficult. Below, two simple examples demonstrating the principle are given.

Example 1

As the first example, consider the situation where only two hypotheses, $H_k : \theta_k \leq 0$, are being tested. Assume that the test statistics are independent and Normally distributed, i.e. $T_k \sim N(\theta_k, 1)$. Let us put a two-point prior on θ: $\mathbb{P}(\theta = (2, 2)) = p$ and $\mathbb{P}(\theta = (0, 2)) = 1 - p$. That is, there is a possibility that $\theta_1 = 0$. However, in this example our utility will not be directly dependent of the value of θ. For both possible θ it will be as given in Table 4.1. After maximization over the family of the weighted Holm MTPs we obtain optimal weights and the corresponding maximum expected utility. Both are displayed in Figure 4.1, left-hand side. We can, for example, see that the weight placed on H_1, w_1, increases with p (i.e. probability for non-zero θ_1), reaching 0.5 for $p \approx 1$ and 0 for $p \approx 0$.

U(I,θ)	0	1
0	0	0.5
1	0.5	1.0

Table 4.1: The utility for example 1 as a function of $I = (I_1, I_2)$, the possible values of I_1 row-wise and I_2 column-wise.

$\theta_1 = 2$		
U(I,θ)	0	1
0	0	0.5
1	0.5	1.0

$\theta_1 = 0$		
U(I,θ)	0	1
0	0	1.0
1	-1.0	0

Table 4.2: The utility function for example 2, divided by the two possible values of θ_1 and four possible values of $I = (I_1, I_2)$. The values of I_1 row-wise and I_2 column-wise.

Example 2

In this example, we let the utility be different for the two possible θ_1 values. We define it as given in Table 4.2. That is, if $\theta = (2, 2)$ then we have a linear utility with both of the hypotheses considered to be equally valuable, as in example 1. If, however, $\theta = (0, 2)$ then we attain the maximum utility if we reject H_2 but accept H_1, i.e. make the correct decision for both of the hypotheses. We also have $U((1, 0), \theta) = -1$ since we make wrong decisions about both of the hypotheses and $U = 0$ in the two remaining cases. Maximizing the expected utility over the family of the weighted Holm MTPs, we obtain optimal weights that are displayed in Figure 4.1, right-hand side. In the figure we can see that the weights and utility differ from the ones obtained in the previous example. The curve corresponding to w_1 have a sharper slope and the curve corresponding to the expected utility is no longer monotone. This happens since maximum utility is gained when H_1 is accepted.

4.9 Bonferroni and pooled tests

In paper III the examined multiplicity corrections revolved around the recycling procedures, namely a Bonferroni test combined with closed testing principle. Choosing the minimum of unadjusted p-values as the test statistic for H_C is natural, as the only requirement for this test is that the distribution of the p-values is stochastically smaller than uniform under H_C. That is, no knowledge about the explicit distribution of the test

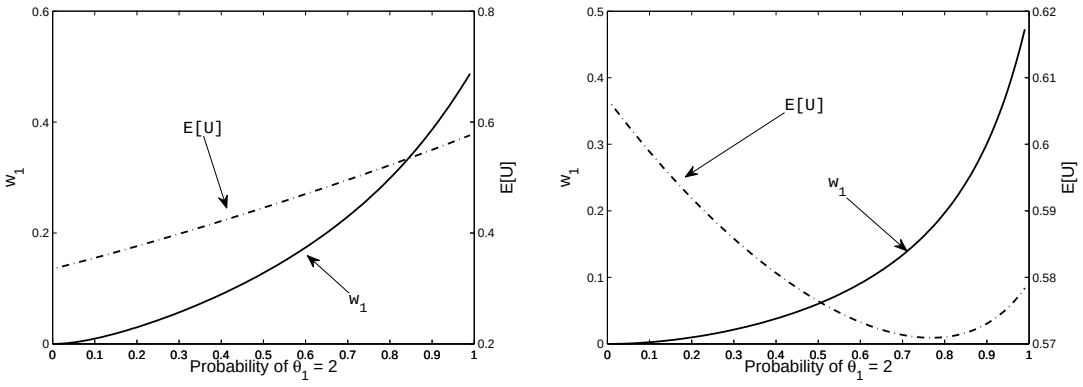

Figure 4.1: Illustrations of example 1 and 2. Probability of non-zero θ_1 (denoted in the text by p) on the x-axis, the weight placed on H_1 on the left y-axis and the corresponding expected utility on the right y-axis.

statistic is required. However, in parametric testing this knowledge is often available, one example, already mentioned earlier, being comparisons of multiple treatments to the same control. If that is indeed the case, a stronger MTP that makes use of this information can be constructed, as is exemplified by the Tukey and Dunnett MTPs.

Originally, both of these tests were single-step procedures similar to the simple Bonferroni MTP. However, later step-up generalizations based on the principle of closed testing have been introduced. For example, the step-up Dunnett procedure involves testing of the intersection hypotheses by comparing the Bonferroni test statistic, i.e. $\min_{k\in C} Kp_k$, not to α but to a slightly larger value derived from the distribution of $\min_{k\in C} p_k$ (which, in turn, is obtained from the distribution of the test statistics). However, there is no result that states that this particular test of the intersection is the best that can be done in the situations where Dunnetts test is applicable.

There exists a multitude of parametric tests for the intersections of the hypotheses of the type $H_k : \theta_k = 0$, many based on the assumption of Normally distributed data, ANOVA being the most well-known example. Another possibility, used in the construction of the step-wise Dunnett test, is to adopt an originally non-parametric test statistic but use the parametric assumptions to derive its explicit distribution. In the next paper we present a new step-up procedure that is more powerful than recycling if the test statistics can be assumed to follow a multivariate normal distribution. This procedure is constructed through combination of the closed testing principle with a so called "pooled" test.

This kind of test has been described in e.g. (17), where it was referred to as "optimal linear combination". As the name suggests, the global test statistic employed by this test consists of a weighted average of the marginal test statistics. If the marginal test statistics follow a multivariate Normal distribution, this test is the most powerful for testing the intersections of the null hypotheses of the type $\theta_1 = ... = \theta_k = 0$. Thus, it

can be expected to give good results even in combination with the principle of closed testing. Note, however, that its optimality for the test of hypothesis intersection does not automatically lead to its optimality when the goal is to maximize the expected value of U, as these two questions are markedly different. In order to evaluate the performance of the suggested test we use the same framework as in paper III, although we restrict ourselves to the family of linear utilities.

4.10 Summary of paper IV

The hierarchical pooled test, which is a combination of the pooled test and the principle of closed testing, was presented. Two variations of this test were also introduced. The first one is a simplification of the test, in the same sense that Holm is a simplification of recycling. The second one alters the rejection regions of the test in order to make it more powerful. These three tests were optimized and compared to each other and the more conventional weighted Holm in terms of expected utility. It was found that, not unexpectedly, all of them perform better than Holm if the test statistics are assumed to follow a fixed multivariate Normal distribution given that the alternative hypotheses are true. However, if it is assumed that the uncertainty in parameters of this distribution is large before the start of the experiment, the Holm procedure may out-perform the pooled test based ones. It was also found that the simplified version of the hierarchical pooled test is, in many situations, comparable to the full version. The second, improved, version of the test often leads to a significant increase in expected utility, but this is accomplished at the cost of heavy and time-consuming numerical computations.

4.11 Optimal tests in closed testing procedures

In paper III we considered the recycling procedures, that can also be viewed as closed testing procedure with a Bonferroni test applied to each of the intersections. In paper IV a hierarchical poled test was examined and found superior if test statistics were assumed to follow a multivariate Normal distribution. So let us now ask ourselves: assuming that the distribution of the test statistics is known both under the null hypotheses and under some alternative, what would be the optimal test statistic for H_C in terms of the marginal p-values?

A special case: two hypotheses

Let us consider the simple case where only two hypotheses, $H_1 : \theta_1 = 0$ and $H_2 : \theta_2 = 0$, are tested against positive alternatives $H_1' : \theta_1 = \theta_1'$ and $H_2' : \theta_2 = \theta_2'$. Say that we have two test statistics that enable us to discriminate between the null and alternatives, with

corresponding p-values p_1 and p_2 with $p = (p_1, p_2)$. Hereafter, we will work directly with p-values rather than test statistics. If both H_1 and H_2 are true, then we will denote the multivariate pdf corresponding to p as $f_0(p)$. If both hypotheses are false, then as $f_{\theta'}(p)$. We will assume, for simplicity, that the distributions of the p-values are continuous. Our goal is to find a rejection region R in the p_1 and p_2 plane where $H_{\{12\}}$ will be rejected and that will have the property of maximizing an expected utility as defined in section 4.2.

The result that we arrive at at the end of this two-hypotheses examination has the form reminiscent of Neyman-Pearson fundamental lemma. It can also be formally proved using the same technique. However, let us start by considering the problem informally and attempt to reason our way to a solution.

As is often the case, it is easier to grasp the problem by approximating it with a similar discrete one. Let us thus begin with finding the region R on a grid, as is illustrated in Figure 4.2. There, we divide the p_1 and p_2 plane into rectangles with side length $\Delta = \alpha/2$. For each such rectangle, denote it by S_{ij}, we have three measures: Q_{ij} which is its probability measure under the null hypothesis, P_{ij} which is its probability measure under alternative and U_{ij} which is the utility that we gain if $p \in S_{ij}$. Since the utility for a particular p is determined by the rejection regions, U_{ij} is constant for all $p \in S_{ij}$.

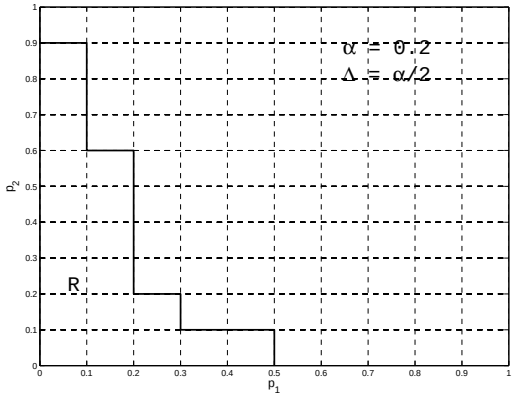

Figure 4.2: Illustration of a discrete rejection region R as a function of p_1 and p_2. The region is delimited by a thick line and is positioned in the lower left corner of the plane.

In our case, we work under the condition that, in order for H_k to be rejected, the corresponding p-value has to be smaller than α. We have thus several potential values of U_{ij} that are functions of the rejection regions for H_1 and H_2. Let $u(p)$ denote the utility as a function of p. For example, say that we gain maximum utility (which we set to 1) if both hypotheses are rejected and we gain nothing if neither of the hypotheses is rejected. Furthermore, the utility value is u_1 if H_1 but not H_2 is rejected and u_2 if the opposite is true. In order for both of the hypotheses to be rejected both p_1 and p_2 have

to be smaller than α, i.e. $u(p) = 1$ if $p \in \{p_1 \leq \alpha\} \cap \{p_2 \leq \alpha\}$. Similarly, $u(p) = 0$ if $p \in \{p_1 > \alpha\} \cap \{p_2 > \alpha\}$, $u(p) = u_1$ if $p \in \{p_1 \leq \alpha\} \cap \{p_2 > \alpha\}$ and $u(p) = u_2$ if $p \in \{p_1 > \alpha\} \cap \{p_2 \leq \alpha\}$.

Our question now becomes similar to a knapsack problem, where we have to fill a sack of certain capacity (in our case α) with objects of different size (in our case Q_{ij}) so as to gain maximum total value (expected utility). The process of "filling the sack" can go something like this. First, we put in the object that has largest value-to-size ratio, i.e. the rectangle with maximum $U_{ij}P_{ij}/Q_{ij}$. Then we continue to the next largest and so on, until we reach maximum capacity, i.e. we reach maximum number of S_{ij} such that $\sum Q_{ij} \leq \alpha$. Letting $\Delta \to 0$ and decreasing the area of S_{ij}, we may realize that the border of the rejection region R should correspond to a contour of $u(p)f_{\theta'}(p)/f_0(p)$. That is, we arrive at the following statement:

Let $u(p)$ be a utility function non-increasing in p, with $\min u(p) = 0$ and $\max u(p) = 1$. In the case of the test of two hypotheses the optimal rejection region R corresponding to $H_{\{12\}}$ can be described through

$$R = \{p : u(p)f_{\theta'}(p)/f_0(p) \leq c\}$$
$$\int_R f_0(p)dp = \alpha$$

with c denoting some constant.

Indication of formal proof. The statement above has the familiar shape of the Neyman-Pearson lemma, with the only difference being the number of dimensions (two rather than one) and the presence of a utility function. The proof can thus follow the same path as the one presented in (24), i.e. by considering R together with some other rejection region and looking at the space where these decision rules differ, but using $u(p)f_{\theta'}(p)$ rather than $f_{\theta'}(p)$ itself.

An example of optimal rejection region obtained in this way is given in Figure 4.3. In the figure, we took $u_1 = 0.3$, $u_2 = 0.7$ and $\theta'_1 = \theta'_2 = 1.5$. We also assumed the test statistics to be independent $N(\theta'_k, 1)$ distributed. In the same figure the region obtained by optimizing a weighted Holm procedure (denote it by R_H) and an improved hierarchical pooled procedure as discussed in paper IV (R_P) are shown for comparison. Since we chose θ_k to be equal, all three procedures will put more mass on the more important hypothesis, H_2. However, the regions belong to different classes, with R_H required to consist of rectangles and R_P following one single contour of $f_{\theta'}(p)$ for $p \in \{p_1 \leq \alpha\} \cup \{p_2 \leq \alpha\}$. The R region is defined by two different contours of $f_{\theta'}(p)$, since $u_1 \neq u_2$. Observe, however, that if we set $u_1 = u_2 = 0.5$ then the regions R and R_P would be the same.

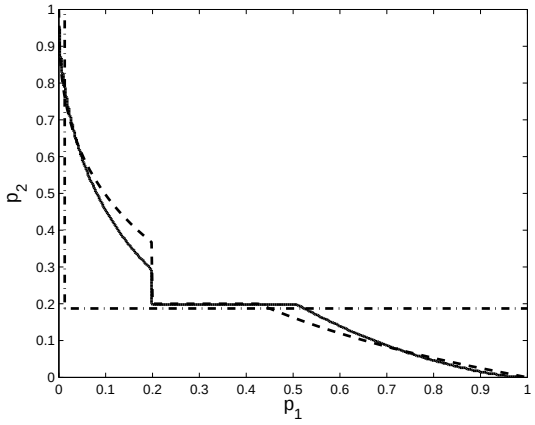

Figure 4.3: The optimal rejection regions for $H_1 \cap H_2$ according to weighted Holm (dash-dotted), improved step-wise pooled pooled procedure (dashed) and "optimal" procedure, R (solid). For clearer illustration, $\alpha = 0.2$ and $\theta_1 = \theta_2 = 1.5$.

More than two hypotheses

Algorithm

Let the utility function be given as $u(p) = \sum_{C \subseteq \Omega} u_C I_C$ with I_C indicating that $H_k : k \in C$ have been rejected and $H_k : k \notin C$ accepted and $0 \leq u_C \leq 1$. Initiate the size of the global intersection to be $L = 2$.

1. Let M be the set of sub-trees in the hierarchy of the closed testing procedure such that the global intersection consists of L hypotheses. Denote each tree in M by m.

2. For each m, define $u(p)$ as above and derive the optimal rejection region for the top-most intersection.

3. Set $L = L + 1$ and repeat. If $L = K$, stop.

The two-hypotheses result above can be generalized to the case with $K > 2$ hypotheses as follows. Again, we will consider only the top-most intersection in the hierarchy, i.e. $H_{\{\Omega\}}$ with $\Omega = \{1, ..., K\}$. We are going to assume that the rejection regions for all other

H_C are already chosen, leading to the potential utility $u(p)$ (with $p = (p_1, ..., p_K)$). The optimal region will again be determined by setting $u(p)f_{\theta'}(p)/f_0(p) = c$, with c a constant determined by α and both $f_{\theta'}(p)$ and $f_0(p)$ this time being K-dimensional. Using this, we can construct a step-wise algorithm for determination of the different R_C, the rejection regions corresponding to H_C, in the same manner as suggested in paper IV.

Chapter 5: Final comments and future work

In this thesis three aspects of a design of a phase III clinical trial have been examined, these aspects being the samples size, the number of active doses and the multiplicity correction that has to be applied when more than one hypothesis is tested. Optimization of these features is part of a larger goal of optimizing the whole design of a phase III trial, or, larger still, optimizing a whole clinical study. A lively area of research that emerged in later years concerns designing phase II and III of a clinical study simultaneously. One popular idea are seamless phase II/III designs, see e.g. (43). Such designs start with several active parallel dose arms, as is typical for phase II, drop most of these dose arms at the interim and seamlessly continue into phase III. The allure of these designs is clear, since they allow the information that accumulated through both of the phases to be used in the final, confirmatory, analysis. Their use, however, has been criticized from a practical point of view, since in reality it is often the case that the response variables measured in these two phases are not the same.

Another, similar idea that arose recently, is to assume that phase II and phase III consist of separate trials, but design phase II in such a way as to maximize the probability of success in phase III. This approach, although seemingly reasonable, is not easily implemented in practice. It consists of adopting a dose-response model with unknown parameters and utilizing a Bayesian scheme to incorporate the possible information that phase II will supply regarding the hypotheses tested in phase III. Given this possible information, it then proceeds to find an optimal phase III design, which, in turn, yields a certain probability of success. All topics discussed in this thesis can be viewed as substeps of such a global optimization. However, although these are different pieces of the same puzzle, more work needs to be done before they can be fitted together seamlessly.

Chapter 6: Bibliography

[1] Antonijevic, Z., Pinheiro, J. et al. Impact of dose selection strategies used in phase II on the probability of success in phase III. *Statistics in Biopharmaceutical Research* 2010; **2**:469–486.

[2] Bauer P, and Köhne, K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1992; **50**:1029–1041.

[3] Benjamini, Y. and Hochberg, Y. Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics* 1997; **24**:407–418.

[4] Brannath, W., Posch. M, and Bauer, P. Recursive combination tests. *Journal of American Statistical Association* 2002; **97**:236–244.

[5] Bretz, F., Maurer, W. et al. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine* 2009; **28**:568–604.

[6] Bretz, F., Posch, M. et al. Graphical approaches for multiple comparison procedures using weighted Bonferroni, Simes, or parametric tests. *Biometrical Journal* 2011; **53**:894–913.

[7] Bretz, F., Hothorn, T. and Westfall, P. *Multiple Comparisons Using R* Chapman & Hall, Boca Raton, 2011.

[8] Burman, C-F. and Sonesson, C. Are flexible designs sound? (with discussion) *Biometrics* 2006; **62**:664–683.

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

[10] Carmer, S.G. and Walker, W.M. Baby Bear's dilemma: a statistical tale. *Agronomy Journal* 1982; **74**:122–124.

[11] Chow, S-C. and Chang, M. *Adaptive Design Methods in Clinical Trials* Chapman & Hall/CRC, 2006.

[12] Chow, S-C. and Liu, J-P. *Design and Analysis of Clinical Trials: Concepts and Methodologies* John Wiley & Sons, Inc., Hoboken, New Jersey, 2004.

[13] Cox, D.R. and Hinkley, F. *Theoretical Statistics* Chapman & Hall, London, 1974.

[14] Denne, J.S. Sample size recalculation using conditional power. *Statistics in Medicine* 2001; **20**:2645–2660.

[15] Edited by Dmitrienko, A., Tamhane, A.C. and Bretz, F. *Multiple Testing Problems in Pharmaceutical Statistics* Chapman & Hall, Boca Raton, 2010.

[16] Dunnett C.W. and Tamhane A.C. A step-up multiple test procedure. *Journal of the American Statistical Association* 1992; **87**:162–170.

[17] Follmann, G. Multivariate tests for multiple endpoints in clinical trials. *Statistics in Medicine* 1995; **14**:1163–1175.

[18] Gallo, P., Chuang-Stein, C. et al. Adaptive designs in clinical drug development - an executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics* 2006; **16**:275–283

[19] Jennison, C. and Turnbull, B.W. *Group Sequential Methods* Chapman & Hall/CRC, 1999.

[20] Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 1988; **75**:800–802.

[21] Holm, S. A simple sequentially rejective multiple procedure. *Scandinavian Journal of Statistics* 1979; **6**:65–70.

[22] Kola, I. and Landis, J. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* 2004; **3**:711–715.

[23] Lehmacher, W. and Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**:1286–1290.

[24] Lehmann, E.L and Romano, J.P *Testing Statistical Hypotheses* Springer Science + Buisness Media, LLC, 2005.

[25] Marcus, R., Eric, P. and Gabriel, K.R. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.

[26] Müller, H.H. and Schäfer, H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and classical group sequential approaches. *Biometrics* 2001; **57**:886–891.

[27] Müller, H.H. and Schäfer, H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.

[28] Neyman, J. and Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: part I. *Biometrika* 1928; **20A**:175–240.

[29] Neyman, J. and Pearson, E.S. On the use and interpretation of certain test criteria for purposes of statistical inference: part 2. *Biometrika* 1928; **20A**:263–294.

[30] Neyman, J. and Pearson, E.S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 1933; **231**:289–337.

[31] O'Brien, P.C. and Fleming, T.R. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.

[32] O'Quigley, J., Pepe, M. and Fisher, L. Continual reassessment method: a practical design for phase 1 clinical trials in cancer. *Biometrics* 1990; **46**:33–48.

[33] Paul S.M., Mytelka, D.S. et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery* 2010; **9**:203–214.

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

[35] Edited by Pong, A. and Chow, S-C. *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development* CRC Press, 2010.

[36] Porter, T.M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* Princeton University press, Princeton NJ, 1995.

[37] Proschan M.A. and Hunsberger S.A. Designed extension of studies based on conditional power. *Biometrics* 1995; **51**:1315–1324.

[38] Rosenberger, W.F. and Haines, L.M. Competing designs for phase I clinical trials: a review. *Statistics in Medicine* 2002; **21**:2757–2770.

[39] Sarkar S.K. Some probability inequalities for ordered MTP_2 random variables: a proof of the Simes conjecture. *The Annals of Statistics* 1998; **26**:494–504.

[40] Shäffer J.P. Multiple hypothesis testing. *Annual Review of Psychology* 1995; **46**:561–584.

[41] Simes R.J. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 1986. **73**: 751–754.

[42] Spjøtvoll, E. On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics* 1972; **43**:398–411.

[43] Stallard, N. and Todd, S. Seamless phase II/III designs. *Statistical Methods in Medical Research* **20**:623–634.

[44] Thall, P.F. and Cook, J.D. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004; **60**:684–693.

[45] Westfall, P.H. and Krishen, A. Optimally weighted, fixed sequence and gatekeeper multiple testing procedures. *Journal of Statistical Planning and Inference* 2001; **99**:25–40.

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