The Limb-Leaf Design:
A New Way to Explore the Dose Response Curve in Adaptive Seamless Phase II/III Trials

John Spivack

Submitted in partial fulfillment of the Requirements for
the degree of Doctor of Philosophy
under the Executive Committee of the Graduate School of
Arts and Sciences

COLUMBIA UNIVERSITY
2012

Abstract

The Limb-Leaf Design:

A New Way to Explore the Dose Response Curve in Adaptive Seamless Phase II/III Trials

John Spivack

This dissertation proposes a method to explore a dose response curve adaptively, allowing new doses to be inserted into the trial after initial results have been observed. The context of our work is adaptive seamless Phase II/III trials and a systematic Limb-Leaf Design is developed. In a case of a nonmonotonic dose response curve where the desired level of effect exists in only a narrow dose range, a simulated comparison between a Limb-Leaf Design and a standard (Thall, Simon, and Ellenberg or TSE-type) adaptive seamless design shows a savings in risk adjusted expected sample size of up to 25%.

Chapter 1 is a review of concepts and particular adaptive seamless designs of interest. Chapter 2 proposes dose addition in adaptive seamless designs and identifies ALS research as an area of application. Chapter 3 develops dose addition as an application of existing methodology. Chapter 4 identifies shortcomings of this approach and proposes a new Horizontal Test as the basis for the Limb-Leaf Design. Chapter 5 supports the development of the Limb-Leaf Design with several theoretical observations. The Limb-Leaf Design is developed in Chapters 6 and 7. Chapter 8 shows a comparison of the Limb-Leaf Design with a TSE-type adaptive seamless design by simulation. Future work is suggested in Chapter 9.

Key words: Horizontal Test, adaptive design, dose response curve, nonmonotonicity, closed testing principle, dose addition.

Contents

Acknowledgement

Here I must thank my advisor, Professor Bin Cheng, for the enormous amount of time, attention, and care he has invested in my work and my progress over many years at Columbia. He has always been there to help me and to challenge me to explore a little further.

I must thank Professor Bruce Levin, who was exceedingly generous with his time and his encouragement. As Chairman of the Department of Biostatistics, he was known and loved for his dedication to the students.

I am also most grateful to the other members of the Dissertation Committee for their work in examining me. I had the good fortune to take my first course in Statistical Inference under Professor Kenneth Cheung and my first course in Clinical Trials under Professor Emilia Bagiella. These fine classes helped to set my path. I am also indebted to Professor Antai Wang of the Herbert Irving Comprehensive Cancer Center for devoting substantial time and effort to my examination process.

Finally, Professor Roger Vaughan, Interim Chairman of the Department of Biostatistics, was exceedingly kind and went out of his way to help coordinate this examination.

Chapter 1

Adaptive Seamless Designs: Introduction and Examples

1.1 Introduction

The traditional process of drug development consists of four (or more) phases: Phase 1, to find which doses can be tolerated, particularly the maximum tolerated dose (MTD); Phase 2, to determine the biological activity and adverse event rates of the tolerated doses; Phase 3, to determine efficacy of a selected dose; and Phase 4, after regulatory approval of the drug, as a review of safety and other long term results. Traditionally, Phase 3 is run and analyzed independently of Phase 2 and the information on Phase 2 results is not used in the final determination of efficacy.

One method of reducing the large costs in time, money, and patient exposure of this process is to combine these phases together and to eliminate the gaps and delays between them. It is often possible to meet the objectives of Phases 2 and 3 within one less costly, combined study and to replace the effort and delay of organizing two studies with that of organizing one. In general, a *seamless design* is a design that combines the objectives of multiple phases of the development process into a single trial.

New issues emerge in the design, conduct, and final inference of such a study. The Phase

2 data on efficacy and adverse event rates will not be available at the start of the study. Since this information is necessary to complete the design of a Phase 3 study with adequate power and control of the type I error probability, the latter part of the trial will have to be allowed to change based on data from the earlier part. An *adaptive design* is one that allows the modification of some aspect of the trial based on data that emerges during the trial. Since there are differences in terminology among authors, we accept the definition used by Jennison [14] under which designs that allow treatment selection; sample size reestimation; changes of endpoints, test statistics, or subpopulations; or other modifications, are considered adaptive. Adaptive designs may be rigid, with selection rules prespecified; partially flexible, allowing adaptations within a given framework; or totally flexible. The term "rigid adaptive design" does not involve any contradiction. In all cases, however, the changes must be "adaptive by design" rather than post hoc in order to preserve statistical validity.

An *adaptive seamless* design is one that: (1) combines the objectives of different stages, (2) allows modification of the trial based on emerging data, and (3) is inferentially seamless in the sense that the final analysis combines data from before and after any adaptation. A basic example of an *adaptive seamless design* combining Phases 2 and 3 would have two stages, the first for learning, the second for confirming. Several patient cohorts, defined by some factor such as dose level or subpopulation would be studied during the learning stage. At an interim analysis a selection decision would be made about which cohorts would be worth further study with additional recruitment in the second stage. After the second, confirmatory stage, the results of both stages would be combined in a way that allows a pre-specified type I error rate and power to be achieved. Effect estimates and confidence intervals should be derived although the process is more involved in this case than in the case of a single independent Phase 3 trial.

The benefits and limitations of adaptive designs have been well discussed. The benefits are attractive: to speed up the process of drug development, to reduce organizational effort, and to allow the data from earlier stages to contribute to the final analysis. However, there

are two major kinds of criticisms. From a practical point of view, adaptive seamless designs are well suited to cases where the time needed to reach the endpoint which is the basis for adaptation is not too long relative to the recruitment rate of the study. During the time prior to the adaptation, there will be a period during which some of the patients have not yet been followed long enough to have been evaluated for the endpoint being used for the modification. If the time to reach this needed endpoint is long relative to the recruitment rate of the study then many patients will wind up being randomized to treatment arms that are not desired and will not contribute to the final analysis. Attempting to pause recruitment would disrupt the study, waste time, and lose many of the benefits of the adaptive design. This objection can be overcome in many circumstances by basing the adaptation decision on well chosen early outcomes or surrogate markers. Maca et al. [28] explain that only a well understood early outcome or, ideally, a surrogate should be used.

The theoretical criticism of adaptive designs is that the tests involved may not depend on sufficient statistics and so do not achieve the maximum theoretical efficiency. Such a departure from the theoretical ideal could counteract the design's intended benefit of being able to draw extra strength from the use Phase II data. This criticism is most relevant to totally flexible designs and to specific inefficient adaptations such as the (unscheduled) rescue of underpowered studies or sample size reestimation based upon an estimate of treatment effect. Jennison [14] and Liu and Anderson [1] among others defend the usefulness and efficiency of adaptive designs, particularly in the rigid or partially flexible forms.

A further issue is that the treatment selection and possibility of early stopping in an adaptive seamless trial leads to statistical bias in the maximum likelihood estimate of the selected treatment's effect, and also to inaccurate coverage of the associated confidence interval. However, appropriate modifications have been developed. A confidence set can be derived through the duality of hypothesis tests and confidence intervals, leading to the "repeated" confidence interval that is strictly conservative. Alternatively, a confidence interval can be constructed using a stagewise ordering of the sample space as described by Brannath,

Posch, and Bauer [7]. The bias of maximum likelihood estimates has been evaluated, and this allows the use of bias-adjusted estimates. The statistical properties of point estimates and simultaneous confidence intervals have also been investigated by Stallard and Todd [40].

The origins of adaptive seamless designs are in early work on fully sequential identification and ranking from the 1960's. The aim of these procedures was to select the treatment with the highest mean or rate under a requirement on the probability of correct selection under a specified alternative. Early elimination of weak treatments and adaptive allocation of subjects to treatment arms were considered by Bechhofer, Kiefer, and Sobel among others [5].

To overcome the need for continuous monitoring, the first two-stage design was proposed by Thall, Simon, and Ellenberg [42]. Details are given in Section 1.2. In this design, the first stage is used to select the best of several candidate treatments and the second stage focuses only on the selected treatment. Both stages include a control arm and the data from both stages are pooled for the final inference. This design applies only to binary outcomes. For survival comparisons, a two stage screening design was proposed by Schaid, Wieand, and Therneau [33].

A generalization that includes multiple stages and the use of a test based on the score statistic was proposed by Stallard and Todd [38]. As explained in Section 1.2, the score based method accommodates a general endpoint, which, for instance, could be normal, binary, ordinal or a time to an event. Other designs have considered all pairwise comparisons with the control, such as the group sequential design of Follman, Proschan, and Geller [9]. An early endpoint for treatment selection and a primary endpoint for confirmation were used by Stallard and Todd in [39] Another route for the development of adaptive seamless designs has been through the adaptive P-value combination tests used by Bauer and Köhne [3]. This approach allows information from earlier stages to be combined with that of later stages, and treatment selection at adaptive interim analyses to be based on all previous information from inside and outside the trial. Midtrial modifications are possible without inflating the familywise type 1 error rate. The key ideas are: the construction of P-values

with conditionally (sub)uniform distributions given the previous stages of the experiment, the pooling of evidence across stages using prespecified combination rules, and the use of a closed testing procedure to control the overall familywise error rate for the multiple hypotheses under study. Section 1.2 discusses this first use of adaptive combination tests. The method of adaptive combination tests is very general and Müler and Schäfer [29] show how it includes group sequential tests, the two stage design of Thall, Simon, and Ellenberg [42], and further generalizations as special cases.

Other formulations are possible. A Bayesian model-based design was proposed by Inoue, Thall, and Berry [13] in which the decisions to stop early, continue Phase II, or proceed to Phase III are based on predictive probabilities of concluding superiority of the new treatment. The design uses survival time as a primary event as well as a short term event that is related to the primary event by means of a mixture model. This model asks for the impact of the short term event on the primary outcome to be demonstrated by the accruing information in the trial. In this and other Bayesian formulations, parameters must be tuned and frequentist properties must be shown by simulation to enforce the needed type I and type II error rates.

We now discuss three of the most important adaptive seamless designs. They are the starting point for our proposal in Section 2, however, only the main ideas of the next section are necessary for continuity.

1.2 Particular Designs of Interest

Bauer and Köhne Design

The original Bauer and Köhne Design [3] was largely in terms of a two stage experiment for testing individual stagewise null hypotheses. However, it introduced a general structure in which there can be multiple stages and test statistics on several hypotheses (perhaps adaptively chosen) are available at each stage. This original version also introduced early stopping boundaries and conditional error functions for this class of designs.

Consider a global null hypothesis that is the intersection, $H_{\{1,2\}} = H_1 \bigcap H_2$, of hypotheses from the first and second stages, respectively. If, for instance, changes in the endpoints or the type of design are allowed, then the individual component null hypotheses may refer to different potential endpoints or to different comparisons performed in the experiment, with H_2 selected or modified based on the first stage data. Based on the interim analysis of the first stage data, it is decided whether to continue to the second stage or to stop the study early for evidence of futility or efficacy. Let p_1 and p_2 be the observed P-values from each stage from tests against their respective null hypotheses. Bauer and Köhne's procedure is as follows:

1. Set a test for stage 1, determine the stopping rules for the interim decision, and pre-specify the combination function, C, of p_1 and p_2 for the final analysis. A combination function $C(p_1, p_2)$ is a function of two (sub)uniform P-values that pools both pieces of evidence into a (sub)uniform combined P-value. Bauer and Köhne recommend the use of Fisher's inverse Chi square method; in this case $p_{\text{F}} = 1 - F(-2\log(p_1 p_2))$, where F is the cdf of the χ^2_4 distribution.

2. Conduct stage 1 of the study and record p_1.

3. Based on p_1, decide whether to stop at interim (either rejecting or retaining $H_{\{1,2\}}$ and H_1) or to continue the study.

4. If the study is continued, use all information, internal and external to the study, to design the second stage and its sample size.

5. Conduct stage 2 of the study, resulting in p_2. The random variable P_2 is constructed to be independent of P_1, given the first stage data, under $H_{\{1,2\}}$.

6. Combine p_1 and p_2 using $C(p_1, p_2)$ and decide for or against $H_{\{1,2\}}$ by comparing C with an appropriate critical value. In case of rejection, decide for or against H_2 using p_2.

The design controls the familywise error rate strongly by applying the closed testing principle. (The closed testing principle is a method of achieving strong control of the familywise error rate with respect to a given set of hypotheses by considering and testing all possible intersection hypotheses. The definitions of strong familywise error rate control and closed testing procedures are given in Chapter 3.) The critical regions for tests at level α of the individual null hypotheses, H_1 and H_2, are $\{p_1 \leq \alpha\}$ and $\{p_2 \leq \alpha\}$.

The possibility of early stopping can be built into the experiment. Fisher's criterion leads to rejection of $H_{\{1,2\}}$ if $p_1 p_2 \leq c_\alpha$, where $c_\alpha = \exp\{-\frac{1}{2}\chi^2_{4,\alpha}\}$, and $\chi^2_{4,\alpha}$ is the $(1-\alpha)$-th percentile of the Chi-squared distribution with 4 degrees of freedom. Since $p_2 \leq 1$, one could stop the experiment at the first interim analysis with rejection of $H_{\{1,2\}}$, if $p_1 \leq c_\alpha$. Early stopping for either futility or efficacy can also be allowed. One can choose α_1, $c_\alpha \leq \alpha_1 \leq \alpha$, and modify the experiment to stop with rejection of $H_{\{1,2\}}$ when $p_1 \leq \alpha_1$. One can choose α_0, $\alpha \leq \alpha_0 \leq 1$, with the requirement that the experiment stops without rejection of $H_{\{1,2\}}$ when $p_1 \geq \alpha_0$. The overall type 1 error rate of the procedure can be directly calculated as $\alpha_1 + c_\alpha\{\log \alpha_0 - \log \alpha_1\}$. This allows α_0 and α_1 to be set for a desired performance.

We give two brief examples of this method. The basic pieces are the (combination) tests of individual hypotheses and the closed testing procedure which ensures strong control of the familywise error rate.

Assume there are two treatments to be tested for superiority to a control, so that their associated null hypotheses are H_{T_1} and H_{T_2}. Denote the intersection hypothesis as $H_{\{T_1,T_2\}}$. At interim, we decide which one to carry forward into the second stage. The final analysis of the selected treatment includes the patients of both stages according to the prespecified combination rule. Assume that Treatment 1 is selected at interim to continue to the second stage so that there is no second stage data available from Treatment 2. Consequently, the intersection hypothesis $H_{\{T_1T_2\}}$ within this second stage is equal to H_{T_1} and its test is performed using only the test of H_{T_1}. Since we have to reject both H_{T_1} and $H_{\{T_1,T_2\}}$ to declare treatment 1 significantly different from the control, we require that

$C(p_{1,T_1}, p_{2,T_1}) < c$ and that $C(p_{1,\{T_1,T_2\}}, p_{2,T_1}) < c$. This may expressed equivalently as $p_{2,T_1} < \min(A(p_{1,T_1}), A(p_{1,\{T_1,T_2\}}))$, for suitably chosen $A(p_{1,T_1})$ and $A(p_{1,\{T_1,T_2\}})$. These values may be thought of as *conditional error functions* (CEF's), expressing the probability of a type 1 error, given the first stage results.

As a second example, consider a study which was originally planned to investigate a formulation of a new medicine, Treatment 1. Let H_{T_1} be the associated one sided null hypothesis. Assume that, at interim, safety problems are detected and it is decided to discontinue the present treatment arm and continue with a different formulation. Let this treatment, Treatment 2, be a prespecified modification of the original treatment. Let H_{T_2} be the associated new null hypotheses. At stage 1 no data are available for Treatment 2 and, vice versa, at stage 2 no data is available for Treatment 1. The related stagewise P-values for the intersection hypothesis are just the corresponding P-values from the elementary hypotheses H_{T_1} and H_{T_2}. Applying the closed test procedure, H_{T_2} is rejected if $C(p_{1,T_1}, p_{2,T_2}) < c$ and $C(p_{2,T_2}) < c$. The first statement corresponds to rejection of $H_{\{T_1,T_2\}}$; the second corresponds to rejection of H_2 and may be written more simply as $p_{2,T_2} < \alpha$. In some situations, the first condition may not be severe and rejection of H_{T_2} often just depends on the comparison of p_{2,T_2} with α. According to [8], the common practice of stopping a study for futility at interim and starting a seemingly independent second study of Treatment 2 at level α can often be justified from this point of view. However, this approach strains the credibility of any final conclusion both because of the unplanned nature of the adaptation and because the overall familywise error rate control is not explicitly demonstrated.

The theory of this design was not completely worked out in the original paper. Counterexamples, where the design fails due to bad behavior of the P-values under certain classes of adaptations are given by Liu, Proschan, and Pledger [26]. For a similar reason, more justification of the use of closed testing procedures may be needed when the second stage hypothesis is chosen data dependently from uncountably many candidates. Still, this is some of the most important early work on adaptive seamless methods.

Thall, Simon, and Ellenberg (TSE) Design

The Thall, Simon and Ellenberg (TSE) Design [42] is a two stage selection and testing design that specializes the general structure of adaptive tests introduced by Bauer and Köhne. The type of adaptation allowed here is limited to selection of the best treatment in the first stage so that it can be studied further in the second stage. This allows clarification of the role of the alternative hypothesis, a definition of the power of the design, and the use of an optimality condition for sample size determination. The relationship of this design to the more general structure of adaptive designs will be explained toward the end of this subsection.

The specific goal is to identify the best of J experimental treatments, $T_1, \ldots, T_J$, and to compare it to a control, T_0. The observed data are binomial counts of success/failure outcomes from patients on each treatment. The success probabilities for patients on the different treatments are denoted by $\mu_{T_j}, j = 0, 1, \ldots, J$. The null hypothesis under test is $H_0 : \mu_{T_0} = \mu_{T_1} = \ldots = \mu_{T_J}$ and it is assumed that $\mu_{T_1} \leq ... \leq \mu_{T_J}$. The specific parameters of the design are n_1 and n_2, the sample sizes of each treatment arm in the first and second stages, respectively, and y_1 and y_2, the first and second stage cutoff values. They are optimized to minimize a version of the total expected sample size, as explained below.

Let $a(p) = \sin^{-1}\sqrt{p}$, the variance stabilizing transformation appropriate for binomial data. At stage i, for $i = 1$ or 2, let n_i be the number of patients in each treatment arm, let $\bar{X}_{i,T_j}$ be the average number of successes in the jth arm, and let $\bar{X}^*_{i,T_j} = (4n_i)^{\frac{1}{2}}\, a(\bar{X}_{i,T_j})$. Let $n = n_1 + n_2$, $\pi = n_1/n$, and y_i be the cutoff value to reject H_0 at stage i. Given specific values of these parameters, the procedure has two steps:

1. Randomize $(J+1)n_1$ patients equally to $T_0, T_1, \ldots, T_J$. Based on their results, compute the first stage test statistic $Y_1 = \frac{1}{\sqrt{2}} \max_{1\leq j\leq J}(\bar{X}^*_{1,T_j} - \bar{X}^*_{1,T_0})$. If $Y_1 > y_1$ then select treatment T_s having the highest observed success rate. Otherwise stop and accept H_0.

2. Randomize $2n_2$ additional patients equally to T_s and T_0. Let $Y_2 = \frac{1}{\sqrt{2}}\{\sqrt{\pi}(\bar{X}^*_{1,T_s} - \bar{X}^*_{1,T_0}) + \sqrt{1-\pi}(\bar{X}^*_{2,T_s} - \bar{X}^*_{2,T_0})\}$. If $Y_2 > y_2$ then reject H_0 and conclude $\mu_{T_s} > \mu_{T_0}$. Otherwise accept H_0.

This design introduces concepts of size and power that are appropriate for its setting. Call experimental treatment T_s "chosen" if it is selected in the first stage and then H_0 is rejected with the conclusion $\mu_{T_s} > \mu_{T_0}$ after the second stage. The size of the procedure is defined as the probability that any T_s is chosen when H_0 holds. The definition of power is based on the recognition that no procedure can adequately discern between two group means that are arbitrarily close and that only certain differences are medically important. Let $0 < \delta_1 < \delta_2$ be two constants such that success rate $\mu_{T_0} + \delta_1$ is not considered a medically important improvement while $\mu_{T_0} + \delta_2$ is worthwhile. Let any T_j for which $\mu_{T_j} \geq \mu_{T_0} + \delta_2$ be considered an acceptable selection. For the purpose of defining power assume that at least one T_j is acceptable and that no μ_{T_j} lies in the interval $(\mu_{T_0} + \delta_1, \mu_{T_0} + \delta_2)$. (This is necessary because no test can discriminate between sufficiently close parameter vectors with prespecified power.) The power function is then defined as $1 - \beta(\vec{\mu})$ equal to the probability that an acceptable choice is made given $\vec{\mu} = (\mu_{T_0}, \ldots, \mu_{T_J})$.

Design parameters must be set based on the desired size and power of the test. Since power is a function of the alternative parameter configuration, it is necessary to decide which alternative configuration is of interest. This is taken to be the *least favorable* configuration, which minimizes power for given μ_{T_0}, δ_1 and δ_2. To this end the authors prove as their "Theorem 1" the intuitive result that $1 - \beta(\vec{\mu})$ is minimized for given μ_{T_0}, δ_1, and δ_2 when $\mu_{T_1} = \cdots = \mu_{T_{J-1}} = \mu_0 + \delta_1$ and $\mu_{T_J} = \mu_{T_0} + \delta_2$. Formulas for the size and power in terms of design parameters n_1, n_2, y_1, and y_2 are derived by direct computation of the probability of rejection under null hypothesis H_0 and under the alternative least favorable configuration. The expression for the size shows that it is independent of μ_{T_0} in the asymptotic limit. This is as one would expect considering that each $\bar{X}^*_{i,T_j}$ would tend to a normal distribution

as sample size went to infinity and that the variances have been stabilized with respect to shifts in μ_{T_0}.

Using these formulas, the design parameters may be chosen to minimize the *risk adjusted expected (total) sample size*, $\frac{1}{2}E(N|H_0)+\frac{1}{2}E(N|\vec{\mu}_{\text{ least favorable}})$, so that the design will perform reasonably well under both the null and the least favorable configurations. This criterion is a form of Bayes' risk and is also used by Liu and Pledger [25]. The specific minimization procedure is not very interesting in itself, but it is a starting point for the routines that are used to set design parameters in more complex situations.

A new feature that arises here is the possibility of a wrong selection. Although this is not a type 2 error, since H_0 is correctly rejected, it is still highly undesirable. Such an incorrect selection might have been considered a familywise type 1 error because it corresponds to rejection of some true null hypothesis of no medically significant treatment effect at a particular dose; it could be dealt with by imposing strong FWER control on the set of tests of each $H_0 : \mu_{T_j} = \delta_1$, for $j = 1, \ldots, J$. Here, the effect is only identified as a third form of error and its probability is denoted by γ. The authors show in a numerical study that this error probability is small over reasonable parameter configurations.

This design may be shown to follow the general structure using combination tests organized into a closed testing procedure. The adaptation rule is rigid (entirely prespecified). In this interpretation, Y_2 combines independent first and second stage $\bar{X}^*_{i,T_S}$ values, according to prespecified weights, to form a test of the selected treatment. Tests of the other required (intersection) hypotheses may be written similarly, and the possibility of early stopping can be included in the conditional error functions of these tests. As is sometimes the case with rigid designs, the final test statistic here looks similar to the (sufficient) statistic we would report in a nonadaptive treatment comparison experiment.

The contribution of the TSE Design is to specialize the adaptive structure to the two stage hypothesis selection and testing situation with binomial data. This specialization permits a meaningful discussion of the power of the design and allows concepts such as the least

favorable configuration and the risk adjusted expected sample size to be introduced. For its use, it is more convenient and slightly more efficient than the original Bauer and Köhne experiment. The restrictions to binomial data and two analysis points, as well as the desire to incorporate covariate information in the analysis are all motivations for further work. These needs are met in the next design.

Stallard and Todd Design

The Stallard and Todd Design, presented in [38], is a generalization of the TSE Design. The goal is again to select the best of several experimental treatments and to confirm its efficacy against a control. It extends the previous work in two ways: It uses the score based method that can handle general forms of data, incorporate covariate information, and allow for nuisance parameters. The design also lets the trial have a prespecified larger number of stages, $I \geq 2$, although selection must still happen at the first interim analysis. The definitions of size and power developed in the TSE Design are applied and the least favorable configuration is again used to set the sample size. However, there is more work in calculating the distributions of the test statistics and in deciding on the stopping boundaries of each stage than there was before.

The plan is a group sequential trial with up to I analysis points. It begins with J experimental treatments, T_j, $j = 1, \ldots, J$, and a control, T_0. At the first interim analysis, the best treatment is identified for further study based on the primary outcome. If this treatment is sufficiently promising, then, in up to $I - 1$ additional stages, new patients are recruited to this arm and to the control. Cumulative score statistics for the superiority of the chosen treatment over the control are compared with upper and lower stopping boundaries. A prespecified (familywise) type 1 error rate, α, and power to identify the correct dose under a specified alternative are achieved. We now go over the details of this plan.

Let θ_{T_j} be a measure of the superiority of T_j over T_0, and let the null hypothesis be $H_0 : \theta_{T_1} = \ldots = \theta_{T_J} = 0$. For $j = 1, \ldots, J$, let z_{1,T_j} and ν_{1,T_j} be the efficient score and

observed Fisher's information for θ_{T_j} at the first interim analysis. The selected treatment is denoted by T_s, so that S is a discrete random variable with values in $\{1, \ldots, J\}$. It is assumed that all $\mathcal{v}_{1,T_j}$ may be taken to be equal and that T_s is selected because it has the largest observed $\mathcal{z}_{1,T_j}$. At later analyses, specified by $i = 2, \ldots, I$, let $\mathcal{z}_{i,T_s}$ and $\mathcal{v}_{i,T_s}$ denote the efficient score and observed information for θ_s. The design enforces weak control of the FWER, however it is possible to show that strong control is achieved as well. The requirement enforced is that, under H_0, the probability of concluding that any θ_{T_j} exceeds 0 be at most α. The power of the study is defined as it was in TSE Design. Quantities $0 < \delta_1 < \delta_2$ are identified such that δ_1 is a marginal improvement of T_i over the control and δ_2 is a clinically meaningful improvement. Assuming that $\theta_{T_1} \geq \delta_2$ and that no θ_{T_j}, $j = 2, \ldots, J$, is in (δ_1, δ_2), the authors follow Thall, Simon, and Ellenberg [42], and argue that the power is at least $1 - \beta$, for all $\theta_{T_2}, \ldots, \theta_{T_J}$ not in (δ_1, δ_2), if it equals $1 - \beta$ in the least favorable configuration. This is given by $\theta_{T_1} = \delta_2$ and $\theta_{T_2} = \ldots = \theta_{T_J} = \delta_1$. Power is evaluated under this alternative.

The stopping decision at the ith analysis point is based on the value of the efficient score statistic, $\mathcal{z}_{i,T_s}$. For $i = 1, \ldots, I-1$, the test continues to the $(i+1)$th interim analysis if the ith interim analysis takes place and $\mathcal{z}_{i,T_s} \in (l_i, u_i)$. Two different *spending functions* are used to give the upper and lower boundaries. The upper boundary points, $\{u_i : i = 1, \ldots, I\}$, are defined using $\alpha_u^* : [0,1] \to [0, \alpha]$, a non-decreasing function with $\alpha_u^*(0) = 0$ and $\alpha_u^*(1) = \alpha$. They must satisfy $Pr(\mathcal{Z}_{i,T_S} \geq u_i, \mathcal{Z}_{1,T_S} \in (l_1, u_1), \ldots, \mathcal{Z}_{i-1,T_S} \in (l_{i-1}, u_{i-1}) | H_0) = \alpha_u^*(t_i) - \alpha_u^*(t_{i-1})$, where t_i is the proportion of the maximum information, $V_{\max}$, that will be available at the ith interim analysis. Similarly, the lower boundary points $\{l_i, : i = 1, \ldots, I\}$ are defined using a nondecreasing function, $\alpha_l^* : [0,1] \to [0, 1-\alpha]$, with $\alpha_l^*(0) = 0$ and $\alpha_l^*(1) = 1 - \alpha$. The corresponding condition is that $Pr(\mathcal{Z}_{i,T_S} \leq l_i, \mathcal{Z}_{1,T_S} \in (l_1, u_1), \ldots, \mathcal{Z}_{i-1,T_S} \in (l_{i-1}, u_{i-1}) | H_0) = \alpha_l^*(t_i) - \alpha_l^*(t_{i-1})$. This method ensures that the upper and lower boundaries meet at the last analysis, however, it is necessary to search over $V_{\max}$ to ensure the desired power at the least favorable configuration.

To implement the design using given error spending functions, forms for the test statistics and specific distributional results are needed. Forms for θ and the corresponding efficient score and observed information in many settings are given by Whitehead [45]. Under a local alternative and appropriate regularity conditions, it is possible to use the following approximation in a study comparing a single experimental treatment to a control: $\mathcal{Z}_1 \overset{\text{approx}}{\sim} N(\theta\mathcal{V}_1, \mathcal{V}_1)$ and $\mathcal{Z}_i - \mathcal{Z}_{i-1} \overset{\text{approx}}{\sim} N(\theta(\mathcal{V}_i - \mathcal{V}_{i-1}), \mathcal{V}_i - \mathcal{V}_{i-1})$, with $\mathcal{Z}_i - \mathcal{Z}_{i-1}$ independent of $\mathcal{Z}_{i-1}$, for $i = 1, ..., I$. The generalization to a multivariate normal (approximate) distribution of a vector of score statistics is discussed in Section 4.2.

These results enable the first stage cutoff values to be approximated from the distribution of the maximum of correlated (multivariate) normal random variables. Subsequent cutoff values are then derived by considering the distribution of the score statistic at a later analysis point as the sum of two pieces: (1) the first stage statistic; and (2) an independent increment, normally distributed with known mean and variance. Numerical calculation of integrals is then used to set the boundaries.

Overall, this design is successful in accommodating other forms of data, allowing covariates and/or nuisance parameters, and making decisions based on the amount of information gathered rather than the treatment of a preset number of patients. The extension to $I \geq 2$ analyses is useful too. A weakness of the original paper is that the asymptotics were not correctly worked out. However, the needed properties can be found directly from the unconditional distribution of the score statistic under a local alternative, and the design is still valid. It is clearly useful in applications.

This design motivates work in several directions. One is the use of multiple endpoints as Stallard and Todd did in [39]. Another is to consider the data dependent promotion of more than one dose into the second stage, and/or the the addition of new treatments. The next section makes such a proposal and identifies an important application.

Chapter 2

A New Two Stage Limb-Leaf Procedure

2.1 Potential Benefits of Dose Addition

Given any unknown dose response curve, there is an important distinction between d^*, a dose with the desired effect (or the maximum possible effect), and $\hat{d}^*$, its estimated value. The corresponding effects of these doses, θ_{d^*} and $\theta_{\hat{d}^*}$, could be different in a meaningful way, with $\theta_{\hat{d}^*} < \theta_{d^*}$. Heuristically, the more closely $\hat{d}^*$ approximates d^*, the greater the chance that the study will reject the global null hypothesis, $H_0 : \theta_d = 0$ for all doses d, and the closer the final dose recommended to patients will be to that which gives them the desired (or maximum) benefit. This is a strong motivation for a better exploration of the dose response curve in adaptive seamless designs.

One possibility for better exploration in an existing design such as the TSE Design [42] (described in Section 1.2) is to start a study with a large number of closely spaced first stage doses. However, there are reasons to expect the performance to suffer. Under the alternative, the true d^* may be hard to identify because it will have many competitors, some with nonzero effects. Also, a large number of first stage patients will have to be randomized to areas of the dose response curve that are not relevant to the final recommendation. Under

the global null hypothesis of no treatment effect at any dose, many patients will have been treated before an early stopping decision can be made. We note that under either hypothesis, treating an excessive number of patients with an ineffective treatment or with ineffective doses of an otherwise worthwhile treatment is not ethically desirable. Some of these issues are acknowledged by Thall, Simon and Ellenberg in the final sections of [42], however they play less of a role in the case where there are only a few first stage doses with broad spacing between them.

We propose to use a two-stage selection procedure in which second stage doses are not only promoted from a modest number of first stage candidates but in which new doses may also be added in response to first stage results. We aim to improve the estimation of d^* and to use patients' data more efficiently, this being particularly so under the global null, where the probability of early stopping should be large. Such promotion and addition decisions could be based on all the available information, including efficacy and toxicity, whether this information comes from within the study or from an outside source.

Two approaches will be developed. One is a completely free addition of second stage doses out of (a countable collection of) all possible dose levels. This first approach will be developed as an extension of known methods in Chapter 3. It is then criticized in Chapter 4 because of its risk of low efficiency and because it does not recognize the inherent structure of the two stage selection process. A more systematic "limb-leaf" approach is developed in Chapters 6-8 as the favored solution. This solution is built from a "horizontal" test that is fundamentally different from the P-value combination rules that have been used in adaptive designs, and whose performance is tailored to the needs of a limb-leaf exploration process. It would also be valid to implement a limb-leaf strategy using the more generic method of Chapter 3; we do not oppose this given the result in Chapter 5 that there does not exist a uniformly best multiple testing procedure in this setting.

The Limb-Leaf Design organizes a prespecified set of candidate doses into two nested partitions. The first uses a coarse increment, and we call it the full set of "limbs", $\mathcal{L}_{\text{full}} =$

$\{L_j : j = 1, \ldots, |\mathcal{L}_{\text{full}}|\}$. We intend for there to be only a few limbs, and reasonable values for $|\mathcal{L}_{\text{full}}|$ are 1 to 4. We can also consider the control dose to be L_0.

The second partition is finer. To any limb L_j we may associate the "leaf" doses $\ell_j = \{l_{jk} : k = 1, \ldots, M_j\}$. The full set of leaves, ℓ_{full}, is then $\ell_1 \cup \cdots \cup \ell_{|\mathcal{L}_{\text{full}}|}$ and the second partition is $\mathcal{S} = \mathcal{L}_{\text{full}} \cup \ell_{\text{full}}$. We call $\mathcal{S}$ a limb-leaf system because it includes the limbs as well as their leaves.

In referring to a set of limbs, we may write them as $\{L_j : j = 1, \ldots, J\}$. In referring to a set of leaves (not necessarily all associated with the same limb) we may sometimes relabel them as $\{l_k : k = 1, \ldots, K\}$ for convenience.

In the two stage selection process, the first stage investigates only the limb doses. At the end of this stage, there is an interim analysis in which it is decided which limbs deserve further study in the second stage as well as which of their leaves. These leaves are intended for further exploration around an area of promising activity, in order to better approximate d^*. The promotion/addition decision can be based on all available information, such as efficacy, safety, and cost, from sources internal or external to the study. In the most general case the number and locations of the promoted limbs and the number and location of the added leaves can all be adaptively chosen. Later, we will specialize to only select one limb and a variable number of leaves to promote.

At the second analysis, a dose is selected for final confirmation. The remainder of the study may use a group sequential design to confirm the efficacy of this $\hat{d}^*$ (the best available approximation to d^*). The study is adaptively seamless in the sense that all of the stages will be allowed to contribute to the final determination of efficacy.

In order to set the parameters of dose addition designs and to justify these schemes in comparison with existing adaptive seamless designs, we use the risk adjusted expected sample size, $w_1\, E(N|H_{\text{global null}}) + w_2\, E(N|H_{\text{chosen alternative}})$, with $w_1 + w_2 = 1$. In the comparison between a Limb-Leaf Design and a standard seamless design in Chapter 8, we identify and choose a "least favorable locatable" alternative. Here w_1 and w_2 are weights

expressing the prior belief in the efficacy of the drug. In applications such as ALS research in Section 2.2, it may be realistic to choose w_1 to be as high as .9 (with w_2 as .1). Generally, if it is too difficult to find a least favorable configuration in closed form, one might set parameters to guarantee desired performance under a chosen alternative, which could be a least favorable configuration among a more restricted class of dose response curves. One would then use simulation studies to assess the performance of the dose addition design against its competitors under plausible alternatives.

We note that limb-leaf designs imply a different, stronger role for the clinicians in the design and interim analyses of the study. They will provide expert judgement on the range of doses to consider, the classification of these doses into limbs and leaves, and the parameters of the search pattern (Section 7.1). This will be done in consultation with the statistician. Potential sources for prior knowledge include preclinical toxicity studies, human or animal trials, pharmacodynamic models, the behavior of related therapies in the same illness, and the behavior of the given treatment in related illnesses. We anticipate that clinicians will appreciate this greater chance to apply their knowledge to influence the design of the study.

2.2 Identification of ALS Research as a Potential Application

Amyotrophic Lateral Sclerosis is a devastating, incurable, neurodegenerative disease with an annual incidence rate of 1 to 2 per 100,000 person years. It usually leads to death within 2-4 years of onset. The disease mechanism is not fully understood, but it is thought that oxidative stress and mitochondrial impairment contribute to neuronal loss.

The Food and Drug Administration (FDA) has approved only one drug for the disease: Riluzole (Rilutek). Clinical trials with ALS patients showed that Riluzole lengthens survival by several months, and may have a greater survival benefit for those with a bulbar onset of the disease. The safety and tolerability of this medication is an issue and those taking

the drug must be monitored for liver damage (occurring in 10% of patients) and other side effects. There is hope that the progression of ALS may one day be slowed or stopped by new medications or a combination of drugs.

In the design of studies to investigate new ALS treatments, it is important to note that the treatment effect of any new medication is likely to be small and patient benefit may exist over only a limited dose range. There may be no justification for assuming a monotone dose-response relationship; the mechanism by which a new drug or combination therapy works may be complex, and benefits may decline before an actual toxicity occurs. There may be several simultaneous mechanisms of action, with negative interactions between them. For instance, Riluzole is believed to reduce damage to motor neurons by decreasing the release of glutamate via activation of glutamate transporters. In addition, the drug may offer other neuroprotective effects by means of sodium and calcium channel blockades, as reported by Hubert et al. [12]; inhibition of protein kinase C, as reported by Noh et al. [30]; and the promotion of NMDA receptor antagonism, as reported by Beal, Lang, and Ludolph [4]. An incorrect assumption of a monotone dose response relationship is a favored explanation for the failure of a recent major ALS study, as argued by Ludolph and Jesse [27]. We conclude that to extract small treatment benefits in ALS, it may be necessary to explore a dose response curve carefully and even to approach a toxicity boundary (without actually crossing it).

Existing Phase II and Phase II/III designs for ALS research are reviewed by Schoenfeld and Cudkowicz [36]. These include: futility designs, lead in designs, multidrug ranking designs, and sequential designs such as the "Christmas Tree". A classic example of the futility design referenced there is a Phase II trial of Coenzyme Q10 by Levy et al. [24], which found insufficient evidence to justify a further confirmatory study. (See also Levin [22].) According to the review, there is a large pipeline of potentially beneficial treatments and an extended debate about methods to test new drugs. The recommendation is that many different designs be developed and that a symposium on clinical trial design become

a regular part of meetings of the ALS research community.

The proposed Limb-Leaf Design combines and enhances many of the known approaches. The possibility for early stopping after initial (first stage) investigation is similar to the concept of a futility design, the advantages of multidrug (or multidose) ranking are also gained, and group sequential continuation after the selection stages could also be allowed. The addition of doses at adaptively determined locations and a systematic process of dose-response exploration are potentially valuable enhancements.

A controversy exists regarding the appropriate choice of endpoint(s) for Phase II or Phase II/III trials. Survival is the outcome of primary interest and should be chosen for Phase III studies. However, for Phase II studies, there is a legitimate case that a measure of muscular function such as the ALSFRS-r could be used. Such a measure is more quickly observed and may be closer along the causal chain to the actual disease process. The ALSFRS-r was used as the primary outcome by Levy et al. in [24], and further evidence for its applicability was given by the same authors in [19]. However, as reported by Bensimon, Lacomblez, and Meininger [6], a major trial of Riluzole showed a survival benefit without a corresponding benefit on the rating scale. (See also Levin et al. [23].) Clearly, this controversy needs to be resolved by experts within the field. In order to recognize the potential need to accommodate multiple endpoints, we will suggest how this could be done for Limb-Leaf Designs in Chapter 9.

We believe that Limb-Leaf Designs could offer worthwhile advantages for the study of new ALS treatments. An adaptive seamless design is desirable because of the scarcity of patients available to be enrolled in such a study, the organizational efficiency of an integrated program, and the urgent need to bring out any successful therapy to those currently suffering from the disease at the earliest time. Given previous experience, it might be reasonable to expect such a trial to have to reveal a small dose range over which there would be a worthwhile treatment effect. Also, within such a dose range, one would need to locate a dose with the desired level of effect (and/or the maximum effect) as well as possible, both to improve the trial's chance

of success and to enhance the treatment's benefit to future patients. The opportunity to end a study early after a crude investigation of the dose response curve with negative results would also be highly desirable. These are features of the Limb-Leaf Design to be developed in Chapters 6-8. Further advantages include the benefit to participating patients of having a greater chance of being assigned to doses that are effective (as opposed to not), and the opportunity for clinicians to contribute more of their knowledge and intuition to the process of study design.

Before approaching our preferred formulation of the Limb-Leaf Design, we will discuss an extension of the original work by Bauer and Köhne.

Chapter 3

Basic Dose Addition: Formulation and Justification

The method of Bauer and Köhne (see Section 1.2) can be generalized to allow data dependent promotion and addition of doses as, for instance, in a Limb-Leaf Design. We begin with the basic background on closed testing procedures and combination tests.

3.1 Closed Testing Procedures and Combination Tests

Suppose there are J null hypotheses, $H_j : \theta_j \leq 0$ for $j = 1, \ldots, J$, available to be tested in a given experiment. Inflation of the number of type 1 errors committed could result from testing each hypothesis at its nominal α value. Instead, these tests must be organized into a multiple testing procedure, and an appropriate generalization of the type 1 error rate condition must be imposed. In an adaptive design, final test statistics may not be available on all J hypotheses, but the approach given here still controls for the presence of these other hypotheses and the biases introduced by selections.

A multiple testing procedure is a rule to decide which of the available hypotheses to reject. Its *familywise error rate* (FWER), under a particular set of parameter values, $(\theta_1, \ldots, \theta_J)$, is $Pr($ Reject any true $H_j)$. Since the event {The selected true null hypothesis is rejected} $\subseteq$

{Any true hypothesis is rejected}, if the familywise error rate is controlled at level α for a particular set $(\theta_1, .., \theta_J)$, then, even if we select which hypotheses to test in a data dependent way, the probability of committing a false rejection is less than or equal to α.

So far, this only concerns control of the familywise error rate under a particular parameter vector $(\theta_1, \ldots, \theta_J)$. There is a consensus in the literature that a multiple testing procedure needs to control the familywise error rate *strongly*, which means that this error rate needs to be at most α for all possible vectors $(\theta_1, \ldots, \theta_J)$. That is, $Pr($ Reject any true $H_j) \leq \alpha, \forall\, (\theta_1, \ldots, \theta_J)$. Using such a procedure, the probability of choosing to focus on any hypothesis and then falsely rejecting that hypothesis is at most α regardless of the parameter configuration.

The argument for strong control of the familywise error rate as opposed to weak control (under only the null hypothesis) is made by Tamhane, Hochberg, and Dunnett in [41], where procedures that do not control the FWER strongly are criticized. The authors' example considers the problem of finding the minimum effective dose, MED $= \min\{j : \theta_j =$ effect of treatment $j > 0\}$, in two separate experiments. The first uses $J = 4$ treatments and the second uses $J = 5$. Assume that the true MED $= 5$ and that the familywise error rate is controlled in each experiment (weakly) only under $H_{0J} : \theta_1 = \ldots = \theta_J = 0$. Then, estimating dose 4 to be the MED would be a correctly controlled type 1 error in the first experiment (where the global null hypothesis is true) and not in the second experiment (where the global null hypothesis is false). Because the true MED can be any one of the dose levels, control of the FWER is needed and only strong control is acceptable.

A general family of procedures that control the FWER strongly is the class of *closed testing procedures*. A closed testing procedure for a given set of hypotheses $\{H_j : j = 1, \ldots, J\}$ is as follows. For each subset S of $\{1, \ldots, J\}$, define the intersection hypothesis $H_S = \bigcap_{j \in S} H_j$. Construct a level α test of each H_S. Such a test, for instance, might combine tests of the individual H_j's using the Bonferroni correction or some other more efficient adjustment for the multiplicity of hypotheses. Finally, the hypothesis $H_j : \theta_j \leq 0$

is rejected overall iff H_S is rejected for every set S such that $j \in S$. The proof that this procedure controls the FWER strongly is as follows. Let S^* be the set of the indices of all true hypotheses. For a familywise error to be committed, H_{S^*} must be rejected. Since H_{S^*} is true, $Pr(\text{Reject } H_{S^*}) \leq \alpha$. Therefore, the probability of a familywise error can be no greater than α. A further discussion on closed testing procedures, an argument for why there is no need to look beyond this class, and their applications to Limb-Leaf Designs are given in Chapter 5

In order to use a closed testing procedure in a multistage trial, we must consider how to form the tests that constitute it. The established approach is to use combination tests that pool stagewise summary measures of evidence.

Consider a null hypothesis of no treatment effect, H_0, that is studied in $I \geq 2$ stages of an experiment. Often, the evidence that emerges from the ith stage is expressed as a P-value, P_i, interpreted as the evidence against H_0 in stage i. Equivalently, a Z-value is reported according to $Z_i = \Phi^{-1}(1-P_i)$. It is commonly assumed that under H_0, $P_i \sim U(0,1)$ (or $Z_i \sim N(0,1)$) and that stagewise P-values are independent. Rigorously, P_i must have a conditional sub-uniform null distribution. It must satisfy $sup_{\theta \in \Theta_0} Pr(P_i \leq u) \leq u, \forall u : 0 \leq u \leq 1$ conditional on the data from previous stages of the exeriment.

Methods of testing H_0 by combining P_i's across stages come from meta-analysis. Two preferred choices are Fisher's inverse Chi-square, and the weighted inverse normal method.

The Inverse Chi-square method follows from $-2\log(P_1 P_2 \cdots P_I) \overset{H_0}{\sim} \chi^2_{2I}$. The associated test rejects H_0 at level α if the left hand side exceeds $\chi^2_{2I,\alpha}$. The associated P-value is the value of α for which equality is achieved.

Another means of combining evidence is the weighted inverse normal method (Weighted Rule). Here, a set of weights $\{w_1, \ldots, w_I\}$ is prespecified such that $\Sigma_{i=1}^I w_i^2 = 1$. The level α test rejects if $\Sigma_{i=1}^I w_i Z_i > z_\alpha$, where z_α is the $(1-\alpha)$th percentile of the standard normal distribution.

These combination rules can be used in sequence, and the type 1 error rate will be

protected when stages of the study are designed adaptively as long as the proper structure is observed. Consider the Weighted Rule, with weights w_1 and w_2, for a two-stage study. After stage 1, it could be decided to design stage 2 as a combination of $M(Z_1)$ sub-stages, also to be evaluated by the weighted inverse normal method. Before conducting these $M(Z_1)$ sub-studies, weights $w_{21}...w_{2M(Z_1)}$ that satisfy $\Sigma_{i=1}^{M(Z_1)} w_{2i}^2 = w_2^2$ must be specified. Let the results of these substages be expressed by the Z-values $Z_{21}...Z_{2M(Z_1)}$. The combined test statistic from the second stage would be Z_2, where $w_2 Z_2 = w_{21}Z_{21} + ... + w_{2M(Z_1)}Z_{2M(Z_1)}$ and the overall test would still be based on $Z = w_1 Z_1 + w_2 Z_2$.

This technique can be used recursively in such a way that the w_i's can depend on previous data and the number of stages can also depend on the emerging data. This leads to the method of *variance spending*. We may consider that each chosen weight, w_i, spends an amount w_i^2 of the variance, and the study ends when the total variance spent equals 1, the variance of the final Z-statistic. However, if such a variance spending rule is not followed the resulting procedure will not control the type 1 error rate correctly.

There are two other combination methods worth noting. Tippett's rule assigns a P-value to H_0 based on $\min(P_1, \ldots, P_I)$ and accepts or rejects accordingly: specifically, $P_{\text{T}} = 1 - (1 - \min(P_1, \ldots, P_I))^I$. This combination rule is recognized for good performance by Goutis, Casella, and Wells [10]. It is used implicitly in several adaptive seamless designs, such as those of TSE and Stallard and Todd, to combine evidence for a combination hypothesis within a stage rather than across stages. When applied to Z-values, we will call it the Maximum Rule because the smallest P-value corresponds to the largest Z-value. This Maximum Rule is different from the "Maximum Rule for P-values", according to which a level α test of H_0 rejects if $\max(P_1, \ldots, P_I)^I < \alpha$.

The review of the combination of evidence through the combination of P-values by Goutis, Casella, and Wells [10] does not produce a clear winner and states that the choice of combination rule for a particular setting is a matter of judgement. In flexible designs where the adaptation rule is impossible to write explicitly, expressing the results by a combination

of P-values may be the only feasible method. This is the dominant approach in adaptive designs and it has been used very successfully.

3.2 A Useful Formulation

The following representation of a two stage study makes it easy to generalize the method of Bauer and Köhne. We use d to refer to any given dose, and θ_d to refer to its effect relative to control. For any collection of doses, D, let H_D be the associated null hypothesis, $\theta_d = 0 : \forall d \in D$. Let $Z_{i,D}$ denote the Z-value against hypothesis H_D in stage i. When D is a singleton, we will suppress its identity as a set and write expressions such as H_{L_1} and Z_{L_1}. Assuming that results on all doses are available in the first stage, a two stage study can be represented as:

Hypothesis	Stage 1 Statistic	Stage 2 Statistic	Combined Statistic
H_{L_1}	Z_{1,L_1}	Z_{2,L_1}	$Z_{L_1} = C_{L_1}(Z_{1,L_1}, Z_{2,L_1})$
H_{L_2}	Z_{1,L_2}	Z_{2,L_2}	$Z_{L_2} = C_{L_2}(Z_{1,L_2}, Z_{2,L_2})$
$\vdots$			
H_{L_J}	Z_{1,L_J}	Z_{2,L_J}	$Z_{L_J} = C_{L_J}(Z_{1,L_J}, Z_{2,L_J})$
$H_{\{L_1,L_2\}}$	$Z_{1,\{L_1,L_2\}}$	$Z_{2,\{L_1,L_2\}}$	$Z_{\{L_1,L_2\}} = C_{\{L_1,L_2\}}(Z_{1,\{L_1,L_2\}}, Z_{2,\{L_1,L_2\}})$
$H_{\{L_1,L_3\}}$	$Z_{1,\{L_1,L_3\}}$	$Z_{2,\{L_1,L_3\}}$	$Z_{\{L_1,L_3\}} = C_{\{L_1,L_3\}}(Z_{1,\{L_1,L_3\}}, Z_{2,\{L_1,L_3\}})$
$\vdots$			
$H_{\{1,\ldots,J\}}$	$Z_{1,\{L_1,\ldots,L_J\}}$	$Z_{2,\{L_1,\ldots,L_J\}}$	$Z_{\{L_1,\ldots,L_J\}} = C_{\{L_1,\ldots,L_J\}}\{Z_{1,\{L_1,\ldots,L_J\}}, Z_{2,\{L_1,\ldots,L_K\}}\}$

Here the $Z_{i,D}$ values are pooled across stages by prespecified combination rules. The combined Z_D values are then used to test individual effects through a closed testing procedure: Reject any H_d iff H_D is rejected according to $Z_D > z_\alpha,\ \forall D \ni d$.

The set of first stage $Z_{1,D}$ values which are available is prespecified. The set of second stage $Z_{2,D}$ values which are available is a function of the first stage data because, in general,

treatments could be dropped from or added to the study.

One way to form $Z_{i,D}$ values for intersection hypotheses within each stage that have the desired conditional sub-uniform property is by maximization and rescaling. This is a special case of Tippett's rule. Let the second stage test statistic for $H_{\{L_1,L_2\}}$ be $Z_{2,\{L_1,L_2\}} = f_{\{L_1,L_2\}}\{Z_{2,L_1}, Z_{2,L_2}\}$, with $f_{\{L_1,L_2\}}$ defined as follows. Under an adaptation such that only one element of $\{Z_{2,L_1},\ Z_{2,L_2}\}$ is present, $f_{\{L_1,L_2\}}$ returns that value; if both are present, $f_{\{L_1,L_2\}} = \Phi^{-1}(F(\max(Z_{2,L_1}, Z_{2,L_2})))$, where Φ is the standard normal cdf and F is the cdf of the maximum of the two second stage Z-values under $H_{\{L_1,L_2\}}$. The null distribution of $Z_{2,\{L_1,L_2\}}$ is then uniform conditional on each adaptation. The function $f_{\{L_1,L_2\}}$ generalizes in the obvious way to an f_D for any other set of doses, D.

To accommodate doses whose first stage data are not available, or vice versa, doses whose second stage data are not available is straightforward. Call a simple or intersection hypothesis *testable* if a Z- value corresponding to it or to any of its sub-hypotheses exists in the second stage. For instance, $H_{\{L_1,L_2\}}$ is testable iff Z_{2,L_1} and/or Z_{2,L_2} is available. Define the combined Z-value for any set of hypotheses D not tested in the first stage as its second stage Z value, $Z_{2,D}$. For instance, if H_{L_2} is not tested in the first stage, we set $Z_{L_2} = C_{L_2}(Z_{1,L_2}, Z_{2,L_2}) = Z_{2,L_2}$. We now use the restricted procedure: REJECT H_d IFF H_D IS TESTABLE AND H_D IS REJECTED ($Z_D > z_\alpha$), $\forall D \ni d$. We note that, by the construction above, if H_d is testable then so are H_D, $\forall D \ni d$, and that the closed testing principle still applies to protect the FWER.

A clearer way to view this procedure is as an extension of the conditional rejection probability (CRP) principle of Müller and Schäfer [29]. Let the data observed up to the adaptation point be expressed by the random vector $\mathcal{X}$. Then, the CRP principle states that the decision function ψ_1 may be changed to any other decision function ψ_2 for which the inequality $E(\psi_2|\mathcal{X}) \leq E(\phi_1|\mathcal{X})$ holds true under the null hypothesis.

From this point of view, our design adapts a prespecified closed testing procedure by using adaptations that preserve or reduce the conditional rejection probability of each test

within it. Specifically, in the design that promotes all doses to the second stage, the test statistic for an arbitrary hypothesis, H_D, is $Z_D = C_D(Z_{1,D}, Z_{2,D})$, with rejection occurring iff $Z_D > z_\alpha$. Given the first stage data $z_{1,D}$, the probability of rejection, $Pr(C_D(z_{1,D}, Z_{2,D}) > z_\alpha)$, may be expressed in terms of the associated *conditional error function* as $Pr(Z_{2,D} > A_D(z_{1,S}))$. Given the first stage data and a possible adaptation, there are two cases for the null distribution of the adapted $Z_{2,D}$, $Z^a_{2,D}$. Either $Z^a_{2,D} \preceq N(0,1)$, in which case $Pr(Z^a_{2,S} > A_S(z_{1,S})) \leq Pr(Z_{2,S} > A_S(z_{1,S}))$, or H_D is declared untestable, in which case $Pr(Z^a_{2,D} > A_D(z_{1,D}))$ is defined as zero. In either case, the conditional rejection probability principle guarantees the validity of the test and of the procedure as a whole. This argument will be developed in Chapter 5.

Müller and Schäfer [29] acknowledge that designing good conditional error preserving rules in the case of even a single hypothesis is not trivial. For a system of tests this could be even more complicated, but a desirable approach would be to begin with a good and efficient group sequential design that controls the FWER using the closed testing principle. Such a design could be optimized for a certain probability to correctly identify an effective treatment in a least favorable configuration, and to have a prespecified probability of stopping early under the global null. Potential adaptations from a restricted class could be specified and the performance under these adaptations could be verified. Two examples are given next. The first uses data dependent promotion of first stage doses. The second also allows data dependent dose addition.

Two Basic Examples

Forwarding More Than One Dose : Consider a two stage experiment with the prespecified set of doses $\{L_j : j = 1, \ldots, J\}$. Based on the first stage data, it is decided which, if any, of the first stage doses to promote and study further. Following Brannath, Posch, and Bauer [7], the second stage could be defined as a combination of substages, and so on, through recursion. The second stage could be redesigned following first stage results to achieve a

needed conditional power.

The base design could be optimized for power under a least favorable configuration under the assumption that all doses are promoted. Assuming a first stage sample size of n_1 per arm and a second stage sample size of n_2 per arm, an efficient combination rule for each hypothesis H_D would then be $C_D(Z_{1,D}, Z_{2,D}) = \sqrt{\frac{n_1}{n_1+n_2}}Z_{1,D} + \sqrt{\frac{n_2}{n_1+n_2}}Z_{2,D}$. Another possibility would be to optimize under the assumption that only the dose with the greatest first stage performance will be promoted. The resulting design can be made to be equivalent to the TSE Design, except that it also allows the promotion of more than one first stage dose.

An example of the possible results from a trial that studies five doses across two stages is given below.

Hypothesis (Dose #'s)	Stage 1 Statistic	Stage 2 Statistic	Combined Statistic
H_{L_5}	Z_{1,L_5}	$\emptyset$	$\emptyset$
H_{L_4}	Z_{1,L_4}	$\emptyset$	$\emptyset$
H_{L_3}	Z_{1,L_3}	Z_{2,L_3}	$Z_{L_3} = C_{L_3}(Z_{1,L_3}, Z_{2,L_3})$
H_{L_2}	Z_{1,L_2}	Z_{2,L_2}	$Z_{L_2} = C_{L_2}(Z_{1,L_2}, Z_{2,L_2})$
H_{L_1}	Z_{1,L_1}	$\emptyset$	$\emptyset$

$\emptyset$ denotes a missing value and/or an untestable dose.

Intersection hypotheses are not shown.

It was adaptively decided to forward doses L_2 and L_3 for further study. Assuming that H_{L_2} and H_{L_5} are true, and that H_{L_1}, H_{L_3}, and H_{L_4} are false, committing a familywise error requires the rejection of $H_{\{L_2,L_5\}}$. In our formulation:

$$C_{\{L_2,L_5\}}(Z_{1,\{L_2,L_5\}}, Z_{2,\{L_2,L_5\}}) = C_{\{L_2,L_5\}}\{f_{\{L_2,L_5\}}(Z_{1,L_2}, Z_{1,L_5}), f_{\{L_2,L_5\}}(Z_{2,L_2}, Z_{2,L_5})\}$$
$$> z_\alpha.$$

If, for instance, Z_{1,L_3} and Z_{2,L_3} are the maxima of their respective stages, the statistics required to test H_{L_3} have a convenient ordering. For the first stage statistics,

$$(Z_{1,L_3}) \geq f_{\{L_2,L_3\}}\{Z_{1,L_2}, Z_{1,L_3}\} \geq \cdots \geq f_{\{L_1,L_2,L_3,L_4,L_5\}}(Z_{1,L_1}, \ldots, Z_{1,L_5})$$

and similarly for the second stage statistics,

$$(Z_{2,L_3}) \geq f_{\{L_2,L_3\}}\{Z_{2,L_2}, Z_{2,L_3}\} \geq \cdots \geq f_{\{L_1,L_2,L_3,L_4,L_5\}}(Z_{2,L_1}, \ldots, Z_{2,L_5}),$$

so that rejection of H_{L_3} occurs if and only if

$$C_{\{L_1,L_2,L_3,L_4,L_5\}}\{f_{\{L_1,L_2,L_3,L_4,L_5\}}(Z_{1,L_1}, \ldots, Z_{1,L_5}), f_{\{L_1,L_2,L_3,L_4,L_5\}}(Z_{2,L_1}, \ldots, Z_{2,L_5})\} > z_\alpha.$$

This means that of all the 16 tests involving dose L_3, it is necessary and sufficient to only consider one.

It is assumed that the combination functions are the same for each hypothesis and, as in the case for the four combination rules stated previously, this function is nondecreasing in both of its arguments.

If doses other than the maximum are tested, the same criterion can be applied to any selected dose. This would be a conservative procedure. A small gain in power might be possible by individually considering each of the intersection tests required to reject a simple hypothesis. This is because the relevant combined Z_D values may be higher (more extreme) than the conservative lower bound. Still, not all of the test would need to be done.

Basic Dose Addition: Adaptive dose addition can also be accommodated. Let the first stage doses be $\{L_j : j = 1, \ldots, J\}$ and let a large collection of possible (new) second stage doses be $\{l_k : k = 1, \ldots, K\}$. For convenience we assume that one first stage dose (L_s) is selected for promotion and that a collection of second stage doses $\{l_k : k \in S\}$ is added adaptively.

In the evaluation of the closed testing procedure following the experiment, those members of $\{l_k : k = 1, \ldots, K\}$ that are not tested in the second stage may be totally ignored. Consider the test of any composite H_C, $C = F \cup G \cup H$, where F is a collection of first stage doses, G is a collection of second stage doses studied in the experiment, and H is the set of those second stage doses not actually used in the second stage. For convenience, we relabel the elements of G as $l_1, \ldots, l_{|G|}$. The second stage test statistic for H_C will be

$$Z_{2,C} = f_{\{F \cup G \cup H\}}(Z_{2,L_s}, Z_{2,l_1}, \ldots, Z_{2,l_{|G|}}).$$

This test, in the first and second stages, is then redundant with that of $H_{C'} = F \cup G$ and may be ignored; among the infinite number of potential second stage doses, only those actually studied have an influence on the final analysis.

The ordering properties discussed above continue to hold and they allow reductions in the collection of tests that is sufficient for conclusions on any dose of interest. For instance, if L_1 has the largest observed Z-value in the first stage,

$$Z_{1,L_1} \geq f_{\{L_1,L_2\}}(Z_{1,L_1}, Z_{1,L_2}) \geq f_{\{L_1,L_2,L_3\}}(Z_{1,L_1}, Z_{1,L_2}, Z_{1,L_3}) \geq \ldots$$

and if it again has the largest observed Z-value in the second stage, then only one test is necessary and sufficient to confirm its efficacy. If it ranks kth largest in the second stage, then k tests are necessary and sufficient to consider.

The tests of leaf doses offer more complications. Even if l^* has the greatest observed effect among the second stage doses, it is still necessary to consider J tests in order to confirm the effect at l^*. These are the tests of the hypotheses corresponding to the sets of doses:

$$\begin{aligned}
&\{L_1^*, l_1, \ldots, l_K\},\\
&\{L_1^*, L_2^*, l_1, \ldots, l_K\},\\
&\quad\vdots\\
&\{L_1^*, L_2^*, \ldots, L_J^*, l_1, \ldots, l_K\},
\end{aligned}$$

where L_1^*, L_2^*, ..., L_J^* denote the J doses under study in the first stage ranked in increasing order by their observed effect sizes. The first stage results on doses such as L_1^* could be very poor, and the failure to reject any of these hypotheses leads to a failure to reject overall. This is problematic, especially in the context of a Limb-Leaf Design where poor performance on some first stage limbs is to be expected. In the next section we give further attention to this issue and make a proposal for improvement.

Adaptive addition of doses is mentioned in the appendix of a paper by Bauer and Keiser [2]. Their development is similar, but we would like to claim that the presentation here (independently derived) is more elegant. The problem of forwarding an adaptively determined number of doses was discussed by Kelly, Stallard, and Todd [20] and by Stallard and Friede [37]. The first version achieved only weak control of the FWER. The second attempt has strong control only if the number of doses to be forwarded is prespecified, and otherwise it needs extensive simulations. The method given here does not have these drawbacks.

Chapter 4

Criticism of the Existing Tests and a Proposal for a New "Horizontal" Test

4.1 Shortcomings of the Basic Dose Addition Method

The clear danger in the basic version of dose addition is the tendency to fail to confirm the effects of added doses because of poor first stage performance at other, irrelevant dose levels.

For instance, consider a two stage study with one original dose, L_1, and one leaf dose to potentially be added, l_1. Assume that the Weighted Rule, $C_{\text{Weighted}}(Z_{1,D}, Z_{2,D}) = \sqrt{\frac{n_1}{n_1+n_2}}Z_{1,D} + \sqrt{\frac{n_2}{n_1+n_2}}Z_{2,D}$, is prespecified to test the global null hypothesis. This is a reasonable choice for a combination rule because it would lead to efficiency in the case that L_1 were promoted and no addition occurred. In the case of normally distributed data with known variance, the resulting test depends on the (sufficient) overall mean and coincides with the UMP test.

However, if first stage results on L_1 are poor (or worse, strongly negative), and the decision is made to discontinue it and instead use l_1, the test statistic for the global null hypothesis then equals $\sqrt{\frac{n_1}{n_1+n_2}}Z_{1,L_1} + \sqrt{\frac{n_2}{n_1+n_2}}Z_{2,l_1}$. In spite of a large value of Z_{2,l_1} (strong evidence for efficacy at l_1), the overall test statistic could be greatly reduced by a low or even negative Z_{1,L_1}. Failure to reject the global null then leads to failure to confirm the efficacy of l_1.

We stress that, because overall rejection of a hypothesis by a closed testing procedure requires rejection in the test of every composite hypothesis containing it, the tendency of any such test to fail due to poor first stage results on other dose levels (far from the most effective dose) limits the performance of the procedure as a whole. In a Limb-Leaf Design, it is assumed that treatment effects will be low, or even negative, relative to the active control, over large parts of the studied dose range. Therefore, we require any method of combining data across stages to be robust to poor first stage results at these dose levels.

It is possible to consider this problem at the level of combination functions (or equivalently, conditional error functions). A robust choice of combination rule is the Maximum Rule, $C_{\text{Max}}(Z_{1,D}, Z_{2,D}) = \Phi^{-1}[\max\{1 - \Phi(Z_{1,D}), 1 - \Phi(Z_{2,D})\}^{\frac{1}{2}}]$, where Φ is the standard normal CDF. As commented in 3.1, this is Tippett's rule, expressed in terms of stagewise Z-values.

Unfortunately, in the case that L_1 performed well and was promoted instead of replaced, there would be problems with efficiency and credibility. One full stage of data would be effectively ignored by this rule; this is inefficient and contrary to the principles of an adaptive seamless approach. Options between these two extremes– maximizing efficiency in the case of promotion and maximizing robustness in the case of a switch– are possible by modifications of the combination function (conditional error function). However, these rules lack a justification in terms of the fundamentals of the problem, involve data distortion, and can strain credibility. One would like to do better. Given that for bivariate normal data, with a single unknown parameter, $\mu \in \mathcal{R}^2$, there is no uniformly most powerful test of $H_0 : \mu = (0, 0)^T$ versus $H_1 : \mu \in \mathcal{R}^+ \otimes \mathcal{R}^+$, we cannot expect a uniformly most powerful test in our situation. What we can seek is an adaptive test that is well motivated, interpretable, and has a favorable tradeoff between efficiency and robustness.

In the following section we propose a "Horizontal Rule" with a fundamentally different structure that cannot be expressed as a combination rule for P-values. This rule is well suited to the needs of the Limb-Leaf Design. As shown in Section 4.3, in the case of good

first stage performance, the efficiency rivals that of the Weighted Rule, and in the case of poor first stage results, the the robustness is equal to that of the Maximum Rule. There is also a small additional contribution to the power of the test from its new structure. The Limb-Leaf Design using the Horizontal Rule is developed in Chapter 6.

4.2 Proposal for a "Horizontal" Test

The Horizontal Test respects the identities of individual doses. By keeping data from doses such as L_1 and l_1 distinct, expressions such as $\sqrt{\frac{n_1}{n_1+n_2}}Z_{1,L_1} + \sqrt{\frac{n_2}{n_1+n_2}}Z_{2,l_2}$ do not arise. The previous trade-off between robustness and performance is modified and some of the distortions of reducing stagewise data to P-values are avoided. The Horizontal Test is compatible with the full score-based method for handling multiple forms of data. We present it in a simple case.

Consider outcomes that are normally distributed with common known variance σ^2. Let the first stage of a two stage experiment use the set of limb doses $\mathcal{L}_{\text{full}} = \{L_j : j = 0, 1, \ldots, |\mathcal{L}_{\text{full}}|\}$, with L_0 as the control. Let the set of second stage leaf doses be $\ell_{\text{full}} = \{l_{jk} : k = 1, \ldots, M_j, \text{ for } j = 1, \ldots, |\mathcal{L}_{\text{full}}|\}$. For any dose d, we use $\bar{X}_d$ to denote the sample mean (or we use $\bar{X}_{i,d}$ when it is necessary to specify the stage, i). Let $Y_d = \bar{X}_d - \bar{X}_0$ (or $Y_{i,d} = \bar{X}_{i,d} - \bar{X}_{i,0}$) be the observed effect relative to the control. We let μ_d be the treatment mean for d and $\theta_d = \mu_d - \mu_0$ be the treatment effect. We assume there are n_1 patents per arm for stage 1 and n_2 per arm for stage 2, so that $V_1 = \sigma^2/n_1$ and $V_2 = \sigma^2/n_2$ are the variances of the first and second stage sample means. We write:

$$\begin{pmatrix} Y_{1,L_1} \\ Y_{1,L_2} \\ \vdots \\ Y_{1,L_{|\mathcal{L}_{\text{full}}|}} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \theta_{L_1} \\ \theta_{L_2} \\ \vdots \\ \theta_{L_{|\mathcal{L}_{\text{full}}|}} \end{pmatrix}, \begin{pmatrix} 2V_1 & V_1 & \cdots & V_1 \\ V_1 & 2V_1 & \cdots & V_1 \\ \vdots & \vdots & \ddots & \vdots \\ V_1 & V_1 & \cdots & 2V_1 \end{pmatrix}\right)$$

and

$$\begin{pmatrix} Y_{2,L_1} \\ Y_{2,L_2} \\ \vdots \\ Y_{2,L_{|\mathcal{L}_{\text{full}}|}} \\ Y_{l_{1,1}} \\ Y_{l_{1,2}} \\ \vdots \\ Y_{l_{|\mathcal{L}_{\text{full}}|,M_{|\mathcal{L}_{\text{full}}|}}} \end{pmatrix} \sim MVN \left(\begin{pmatrix} \theta_{L_1} \\ \theta_{L_2} \\ \vdots \\ \theta_{L_{|\mathcal{L}_{\text{full}}|}} \\ \theta_{l_{1,1}} \\ \theta_{l_{1,2}} \\ \vdots \\ \theta_{l_{|\mathcal{L}_{\text{full}}|,M_{|\mathcal{L}_{\text{full}}|}}} \end{pmatrix}, \begin{pmatrix} 2V_2 & V_2 & \cdots & V_2 \\ V_2 & 2V_2 & \cdots & V_2 \\ \vdots & \vdots & \ddots & \vdots \\ V_2 & V_2 & \cdots & 2V_2 \end{pmatrix} \right)$$

for the first and second stages, respectively. This simplification captures the key features of the score based method.

In general, let θ_{d_j} be a measure the superiority of treatment d_j over control d_0 and let it be expressible as $\mu_{d_j} - \mu_{d_0}$, where μ_{d_j} is the measure of the level of response in group j. Then, as shown by Scharfstein, Tsiatis, and Robins, [34], the score function for θ_{d_j}, $\mathcal{Z}_{d_j}$, can be expressed as $\mathcal{V}(\hat{\mu}_{d_j} - \hat{\mu}_{d_0})$ where $\hat{\mu}_{d_j}$ is the maximum likelihood estimate of μ_{d_j} and $\mathcal{V}$ is the observed information for the comparison. This gives an immediate interpretation of the score vector.

In large sample behavior, the relevant mathematical result is the weak convergence of the normalized score vector under a local alternative sequence. If $\theta_{\mathsf{n}} = \theta_0 + h_{\mathsf{n}}\mathsf{n}^{-1/2}, h_{\mathsf{n}} \to h \in R^k$, and appropriate regularity conditions hold (quadratic mean differentiable family), then under $P^{\mathsf{n}}_{\theta_{\mathsf{n}}}$, $\mathsf{Z}_{\mathsf{n}} \xrightarrow{d} N(I(\theta_0))$. For details and exact notation see Lehmann [21], Chapter 12. In the literature of sequential trials [14, 39, 45], this is often written less rigorously. In the case of equal observed information per group, with $\mathcal{V}_{L_0} = \mathcal{V}_{L_1} = \cdots = \mathcal{V}_{L_J} = \mathcal{V}$ as the observed information for each comparison, a common notation is:

$$\begin{pmatrix} \mathcal{Z}_{L_1} \\ \mathcal{Z}_{L_2} \\ \vdots \\ \mathcal{Z}_{L_J} \end{pmatrix} \overset{\text{approx}}{\sim} MVN\left(\begin{pmatrix} \theta_1\mathcal{V} \\ \theta_2\mathcal{V} \\ \vdots \\ \theta_J\mathcal{V} \end{pmatrix}, \begin{pmatrix} \mathcal{V} & \mathcal{V}/2 & \cdots & \mathcal{V}/2 \\ \mathcal{V}/2 & \mathcal{V} & \cdots & \mathcal{V}/2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{V}/2 & \mathcal{V}/2 & \cdots & \mathcal{V} \end{pmatrix}\right).$$

We give a first definition of the Horizontal Test.

Definition 1 *In the above setting, consider the set of doses $D = \mathcal{L} \cup \ell$, with $\mathcal{L} \subset \mathcal{L}_{\text{full}}$, and $\ell \subset \ell_{\text{full}}$. A horizontal test of $H_D : \theta_d = 0, \forall d \in D$ is given by these steps:*

1. *Set c_D according to*

$$Pr_{H_D}(\max\{n_1Y_{1,L} + n_2Y_{2,L} : L \in \mathcal{L},\ k_Dn_2Y_l : l \in \ell\} \geq c_D) = \alpha$$

   where k_D is the ratio of standard deviations $\sqrt{\frac{n_1+n_2}{n_2}}$.

2. *Following the first stage, decide which set $S' = \mathcal{L}' \cup \ell'$, with $\mathcal{L}' \subset \mathcal{L}$, and $\ell' \subset \ell$, to study in the second stage. Set c'_D such that:*

$$Pr_{H_D}(\max\{n_1Y_{1,L} + n_2Y_{2,L} : L \in \mathcal{L},\ k_Dn_2Y_l : l \in \ell\} \geq c_D | y_{1,L} : L \in \mathcal{L})$$

$$= Pr_{H_D}(\max\{n_1Y_{1,L} + n_2Y_{2,L} : L \in \mathcal{L}',\ k_Dn_2Y_l : l \in \ell'\} \geq c'_D | y_{1,L} : L \in \mathcal{L})$$

3. *Perform the second stage and reject H_D iff*

$$\max\{n_1y_{1,L} + n_2y_{2,L} : L \in \mathcal{L}',\ k_Dn_2y_l : l \in \ell'\} \geq c'_D$$

This test is valid by Müller and Schäfer's principle. Let R be the event of rejection of H_D in the procedure restricted to D'=D, and let R_a be the event of rejection of H_D in the procedure allowing full data dependent adaptation. Under H_D, with the first stage data written as $\mathcal{X}$:

$$Pr(R) = E\{Pr(R|\mathcal{X})\} = E\{Pr(R_a|\mathcal{X})\} = Pr(R_a) = \alpha. \tag{4.1}$$

Standard assumptions that are implicit in other adaptive designs, for instance that the second stage data be conditionally independent of the first stage given the adaptation decision, and that the adaptation rule used to select D' be measurable with a countable range must also be made here. These are discussed by Jennison and Turnbull, and by Liu, Proschan, and Pledger [15, 26]. The preferred method to calculate adapted c'_D values is by Monte Carlo simulation.

The name "Horizontal" refers to the way the test first pools data from different treatment groups (horizontally) across stages before taking a maximum. This is in contrast to the formulation of Section 3.2, where data were reduced (vertically) to stagewise P-values and later pooled (horizontally) using combination rules.

4.3 Characteristics of the Horizonal Test

Here we argue that the Horizontal Test is a good choice for use in a Limb-Leaf Design from two perspectives. One is the evaluation of its performance under fixed adaptation plans that are similar to those we would expect to follow in a Limb-Leaf Design. The other is the graph of its conditional error function. Both perspectives are informative in showing that the efficiency of the Horizontal Test is close to that of the efficient combination test based on the Weighted Rule and that the robustness of the Horizontal Test exceeds that of the the robust combination test based on the Maximum Rule.

Consider a two stage study with two doses, d_1, and d_2. Let d_1 be studied in the first stage with options to promote it to the second stage (sustain), promote it and add d_2 (sustain with addition), discontinue it (discontinue), or discontinue it and switch to d_2 in the second stage

(switch). Let the effects for d_1 and d_2 be θ_{d_1} and θ_{d_2}, respectively. We do not assume that d_2 is a leaf from limb d_1. We need to allow its effect to vary.

To be specific, let there be 25 patients per arm (including the control) in each stage. Outcomes are normally distributed with standard deviation σ. We will consider two adaptation plans. Plan 1 is to sustain d_1 if $Y_{1,d_1} \geq \sigma/4$ and otherwise to switch to d_2. Plan 2 is to sustain d_1 with addition if $Y_{1,d_1} \geq \sigma/4$, and otherwise to discontinue the study. These rules mimic what would occur in an actual trial, where adequate performance of a treatment arm would lead to further study, and poor performance would cause interest to shift to other, potentially more effective dose levels.

The first case we present is where d_1 is effective, $\mu_{d_1} \geq \sigma/2$. Plan 1 is used for the adaptation. We report the power to confirm the effect of d_1. It is seen that the Horizontal Rule (Horizontal Test) outperforms the robust Maximum Rule and does not suffer significantly in performance over the most important range of powers compared to the Weighted Rule.

Sustain after good first stage results for n = 25

Form of Rule	$\theta_{d_1} = \frac{\sigma}{2}$	$\theta_{d_1} = \frac{5\sigma}{8}$	$\theta_{d_1} = \frac{3\sigma}{4}$	$\theta_{d_1} = \sigma$
Horizontal	.68	.85	.94	1.00
Maximum	.58	.80	.92	0.99
Weighted	.74	.87	.95	1.00

The next case is where the first dose is ineffective and a switch is appropriate. Plan 1 is still followed. We report the results over a range of means. Note that lower powers are due to the effect of the lower cumulative sample size on d_2. This is correctable by increasing the second stage sample size to achieve any desired conditional power.

Switch after poor first stage results for $\mathbf{n = 25}$, $\theta_{\mathbf{d_1}} = \mathbf{0}$

Form of Rule	$\theta_{d_2} = \frac{\sigma}{2}$	$\theta_{d_2} = \frac{5\sigma}{8}$	$\theta_{d_2} = \frac{3\sigma}{4}$	$\theta_2 = \sigma$
Horizontal	.38	.51	.63	.77
Maximum	.35	.49	.61	.77
Weighted	.19	.30	.40	.57

The Horizontal Rule outperforms the robust Maximum Rule. The performance of the Weighted Rule is poor, and despite its ideal performance in the case of sustaining a limb dose, the power in case of a switch is so low that the Weighted Rule is not acceptable. Fortunately, the Horizonal Rule has similar efficiency and is far more robust.

We now allow more than one dose in the second stage under Plan 2. Assume both d_1 and d_2 are effective. We report the power to confirm either effect over a range of means.

Sustain and add after good first stage results for $\mathbf{n = 25}$, $\theta_{\mathbf{d_1}} = \theta_{\mathbf{d_2}} = \theta$

Form of Rule	$\theta = \frac{\sigma}{2}$	$\theta = \frac{5\sigma}{8}$	$\theta = \frac{3\sigma}{4}$	$\theta = \sigma$
Horizontal	.69	.87	.95	0.99
Maximum	.62	.81	.92	0.99
Weighted	.75	.88	.96	1.00

The Horizontal Rule again outperforms the robust Maximum Rule and does not suffer significantly in performance over the most important range of powers, relative to the Weighted Rule.

We also consider the power under Plan 2 to confirm the effect on d_1 when the added dose, d_2, is not effective. We set θ_{d_2} to 0 and allow θ_{d_1} to vary.

Sustain and add after good first stage results for $\mathbf{n = 25}$, $\theta_{\mathbf{d_2}} = \mathbf{0}$

Form of Rule	$\theta_{d_1} = \frac{\sigma}{2}$	$\theta_{d_1} = \frac{5\sigma}{8}$	$\theta_{d_1} = \frac{3\sigma}{4}$	$\theta_1 = \sigma$
Horizontal	.66	.84	.94	0.99
Maximum	.56	.76	.90	0.99
Weighted	.68	.85	.94	1.00

Favorable properties of the Horizontal Rule can be seen from the graph of its conditional error function in relation to the conditional error functions of the other rules (see Figure 1, Appendix). We recall that the value of the conditional error function CEF(P_1) for a P-value combination test is the greatest second stage P-value that could lead to rejection. Assuming a (conditional) uniform P_2, CEF(P_1) equals the conditional probability of a type 1 error, given the first stage P-value. The CEF can also be expressed in the Z_1, Z_2 plane. In general, the Horizontal Rule is not expressible in terms of a combination rule for P-values; rigorously its conditional error function is undefined. However, under Plan 1 of the previous simulation study, the conditional probability of a type 1 error given the first stage results can be computed. This conditional error function coincides with that of the robust Maximum Rule, when first stage results are poor. When first stage results are good, its conditional error function bends to approach that of the efficient Weighted Rule. This behavior is highly desirable.

From one point of view, the Horizontal Rule provides a well motivated and appropriate conditional error function for this application. However, further simulation studies (not shown) demonstrate that the new structure of the Horizontal Rule also adds to its performance. This can be seen even in the two treatment case. To demonstrate, we compared the Horizontal Rule to the combination rule of stagewise P-values whose conditional error function was set equal to the previously described conditional error function of the Horizontal Rule. The results were positive. For instance, in the case of sustaining and adding after

good first stage results, if $\theta_{d_1} = \frac{5\sigma}{8}$ and $\theta_{d_2} = 0$, the power of the Horizontal Rule was approximately .84 while the power of the stagewise constructed substitute was approximately .81.

Given these results, we now wish to formulate a Limb-Leaf Design using Horizontal Tests, and to investigate the performance. This will be done in Chapter 6. First we make some theoretical observations.

Chapter 5

Further Theory of Closed Testing Procedures

Here we give some results to justify our approach to produce the Limb-Leaf Design. Several issues to discuss are: the choice of a closed testing procedure as the means to control the FWER, what kind of optimality is achievable, and the validity of a closed testing procedure based upon the Horizontal Rule. As an aside, an alternate derivation of the Hölm Procedure is deduced from Proposition 1.

Proposition 1 *Any multiple testing procedure $\mathcal{B}$ concerning hypotheses $\{H_j : j = 1, \ldots, J\}$, with strongly controlled familywise error rate α is "dominated" by a closed testing procedure $\mathcal{C}$ with the same familywise error rate. (By this we mean that for any false H_j, the probability of rejecting H_j in $\mathcal{C}$ is greater than or equal to the probability of rejecting H_j in $\mathcal{B}$, regardless of the underlying parameter configuration.)*

Proof: For each $S \subset \{1, \ldots, N\}$ define the test C_S^* of $H_S = \bigcap_{j \in S} H_j$ as

$$\begin{cases} 1 & \text{if } H_j \text{ is rejected in } \mathcal{B}, \text{ for some } j \in S; \\ 0 & \text{otherwise} \end{cases}$$

and let H_j be rejected in the multiple testing procedure $\mathcal{C}^*$ iff C_S^* rejects H_S, $\forall S \ni j$. This means that $\mathcal{C}^*$ is a closed system of tests such that H_j is rejected by $\mathcal{C}^*$ iff H_j is rejected

by $\mathcal{B}$. To verify that each $C_S^* \in \mathcal{C}^*$ is of level α, we refer to the strong familywise error rate control of $\mathcal{B}$. Assume that for given S, all $j \in S$ are true. Then $Pr\{C_S^* = 1\} \leq Pr\{\text{Reject some true null hypothesis in } \mathcal{B}\} \leq \alpha$

Now the individual tests, $C_S^* : S \subset \{1, \ldots, N\}$, may be improved as follows: If $\sup_{\theta \in H_S} Pr\{C_S^* = 1\} \leq \alpha$ then we may add to its critical region, $R_{C_S^*}$, a (possibly null) set A_S of the sample space so that its size becomes α. Let the test C_S be constructed from C_S^* for each S in this manner, with rejection region $R_{C_S} = R_{C_S^*} \cup A_S$. The resulting tests form a closed procedure, $\mathcal{C}$, and since, for any false H_j, $\{\cap C_S^* \text{ rejects}, \forall S \ni j\} \subset \{\cap C_S \text{ rejects}, \forall S \ni j\}$, $Pr(\mathcal{C}^* \text{ rejects } H_j) \leq Pr(\mathcal{C} \text{ rejects } H_j)$. Thus $\mathcal{C}$ dominates $\mathcal{C}^*$, and also $\mathcal{B}$.

This observation allows us to think of the closed testing procedures in the multiple testing problem as analogous to an essentially complete class in a decision theory problem. We do not need to look beyond this class in our choice of test procedure and can instead focus on choosing an efficient and convenient procedure within this class.

A desirable property of a closed testing procedure is consonance.

Definition 2 *The closed testing procedure $\mathcal{C}$ of hypotheses $\{H_j : j = 1, \ldots, J\}$ using tests $\{C_S : S \in \mathcal{S} = 2^{\{1,\ldots,J\}}\}$ is called consonant if rejection of any (composite) null hypothesis H_S by $\mathcal{C}$ implies the further rejection of some hypothesis H_T, $T \subset S$.*

In a consonant procedure, rejection of the global null in its specific test ensures (by recursion) that at least one component hypothesis will be rejected in the overall procedure. A trivial example is as follows: Consider hypotheses $\{H_1, \cdots, H_J\}$ with associated P-values $\{P_1, \ldots, P_J\}$. Let a closed testing procedure be formed by testing composite hypotheses using the Bonferroni correction applied to each component. Then, since $\frac{\alpha}{J} < \frac{\alpha}{J-1} < \cdots < \frac{\alpha}{1}$, it is clear that this procedure is consonant.

Proposition 2 *This dominating procedure may be chosen to be consonant.*

Proof: Consider the procedure $\mathcal{C}$ given above. The previous construction may be applied again to generate a procedure $\mathcal{C}'$. It then follows that $\mathcal{C}'$ is consonant and rejects any H_j iff this hypothesis is rejected in $\mathcal{C}$.

It may not be possible (or useful) for an adaptive procedure to maintain consonance under every possible adaptation. However, this result is still interesting.

The next question is what kind of optimality is achievable. Proposition 3 shows that even in simple cases there does not exist a UMP multiple testing procedure.

Proposition 3 *If we define a UMP multiple testing procedure of familywise error rate α for a given set of hypotheses as one that dominates every other multiple testing procedure of level α, then no UMP multiple testing procedure exists (even) for testing the means of two independent normal populations.*

Proof: Consider $(Z_1, Z_2)' \sim \text{MVN}\left(\vec{\theta}, \text{I}_2\right)$, where $\vec{\theta} = (\theta_1, \theta_2)'$ and $\mathbf{I}_2$ is the identity matrix of dimension 2. The hypotheses of interest are $H_1 : \theta_1 \leq 0$ and $H_2 : \theta_2 \leq 0$. Let the corresponding alternatives be $H_{1a} : \theta_1 > 0$ and $H_{2a} : \theta_2 > 0$.

Assume that a UMP multiple testing procedure $\mathcal{B}$ of familywise error rate α exists. Then it may be chosen as a closed consonant testing procedure consisting of size α tests, $\psi_1, \psi_2,$ and $\psi_{\{1,2\}}$, of $H_1, H_2,$ and $H_{\{1,2\}} = H_1 \cap H_2$, respectively. Let ϕ_1 denote the UMP size α test of H_1 versus H_{1a}, and let ϕ_2 denote the UMP size α test of H_2 versus H_{2a}. Let the rival procedure $\mathcal{B}_1$ test H_1 and $H_{\{1,2\}}$ by ϕ_1 and H_2 by ϕ_2. Similarly, let the procedure $\mathcal{B}_2$ test H_2 and $H_{\{1,2\}}$ by ϕ_2, and test H_1 by ϕ_1. We produce a contradiction as follows:

Consider the alternative ($\theta_1 = \theta_1' > 0$, $\theta_2 = 0$). Then, since $\mathcal{B}$ dominates $\mathcal{B}_1$, $Pr(\psi_{\{1,2\}} = 1) = Pr(\phi_1 = 1)$. Then, from the uniqueness of UMP tests for normal models, $\psi_{\{1,2\}} = \phi_1$ a.s. must be satisfied. Now consider alternative ($\theta_1 = 0$, $\theta_2 = \theta_2' > 0$). For $\mathcal{B}$ to dominate $\mathcal{B}_2$, $Pr(\psi_{\{1,2\}} = 1) = Pr(\phi_2 = 1)$, and $\psi_{\{1,2\}} = \phi_2$ a.s. must be satisfied. These two necessary conditions on $\mathcal{B}$ are not consistent.

Seeing this limitation, our target will be to produce a closed testing procedure with a favorable tradeoff between efficiency and robustness. The performance should hold over a range of alternative configurations of interest and over plausible adaptations.

As a first step, we show that a closed testing procedure built from the Horizontal Rule will be valid.

Proposition 4 *Adaptations of a closed testing procedure that preserve the conditional type 1 error rate of each constituent test maintain the strong control of the familywise error rate of the procedure.*

Proof: Let S^* be the set of the indices of all true null hypotheses and let $H_{S^*} = \bigcap_{j \in S^*} H_j$. Given the closed structure, it still holds in the adapted version of the procedure that {Reject any true null hypothesis} $\subset$ {Reject H_{S^*}} and $P($ Reject any true null hypothesis) $\leq P$(Reject H_{S^*}). Equation 4.1, Section 4.2, holds for any adaptive test that preserves its conditional type 1 error rate. Therefore, the adaptive test of H_{S^*} is valid and the familywise error rate of the procedure as a whole is controlled at α.

Example: To show that these observations are useful beyond Limb-Leaf Designs, we give an example of an alternate derivation of a classic sequential test procedure from Proposition 1.

The Hölm Procedure [11] is a versatile and powerful procedure for the following situation: There are J null hypotheses, $H_1, H_2, \ldots, H_J$ that are tested separately with statistics $Y_1, Y_2, \ldots, Y_J$, respectively. These test statistics are assumed to tend toward greater values when their corresponding hypotheses are not true. The critical level $\hat{\alpha}_k(y_k)$ for the outcome y_k of statistic Y_k then equals $\sup_{H_k} Pr(Y_k > y_k)$. The observed levels (P-values) $R_1, R_2, \ldots, R_n$, defined by $R_k = \hat{\alpha}_k(Y_k)$, are ordered as $R^{(1)} \leq R^{(2)} < \cdots \leq R^{(J)}$. Let their corresponding hypotheses be $H^{(1)}, \ldots, H^{(J)}$. For given α, the procedure is as follows:

Step 1: Let $j = 1$.

Step 2: Is $R^{(j)} \leq \frac{a}{J-j+1}$?

Step 3: If no, accept $H^{(j)}, ..., H^{(J)}$, and end the procedure.

If yes, reject $H^{(j)}$.

Step 4: If $j = J$, end the procedure, otherwise let $j = j + 1$ and go to Step 2.

Proof: Consider the Bonferroni tests that compare each R_j to $\frac{\alpha}{J}$. For any $S \subseteq \{1, \ldots, J\}$ define the test C_S^* that rejects iff H_j is rejected in its Bonferroni test for some $j \in S$. Each C_S^* is then a level α test of H_S, and any H_j is rejected in the Bonferroni test iff C_S^* rejects $\forall S \ni j$. This defines $\mathcal{C}^*$, consisting of $\{C_S^* : S \subseteq \{1, \ldots, J\}\}$ as a closed testing procedure that rejects an individual hypothesis iff this hypotheses is rejected in the Bonferroni Procedure.

We note that the size of each $C_S^* \in \mathcal{C}^*$ is only $\frac{|S|}{J}\alpha$, where $|S| = m$. Following the method of Proposition 1, the size of each test can be increased to α in a way that increases the overall power. In this particular case, let the test be enlarged as follows: for any C_S^* with $|S| = m$, derive the Bonferroni tests with level $\frac{\alpha}{m}$ of each $H_j, j \in S$. Let C_S reject H_S if any of these m component tests rejects. The result is an improved closed testing procedure $\mathcal{C}$ that is strictly more powerful than the original procedure $\mathcal{C}^*$.

Finally, it is necessary to show that rejection of a component hypothesis in $\mathcal{C}$ is equivalent to its rejection in the sequential Hölm Procedure. Let us assume that $H^{(1)}, \ldots, H^{(j)}$ are rejected sequentially. Consider $H^{(1)}$, with obtained level $R^{(1)} < \frac{\alpha}{J}$. This forces the rejection of the global null, and of the null hypothesis by every C_S with $H_S \subset H^{(1)}$. Therefore $H^{(1)}$ is rejected in $\mathcal{C}$. Similar reasoning applies to $H^{(2)}$: the global null has already been rejected, and every remaining $H_S \subset H^{(2)}$ must satisfy $|S| \leq J - 1$. Since $R^{(2)} < \frac{\alpha}{J-1}$, each such hypothesis will be rejected in $\mathcal{C}$. By recursion, we continue to consider hypotheses in this order until $H^{(j)}$ is reached. At this last stage, every $H_S \subset H^{(j)}$ yet to be rejected now satisfies $|S| < J - j + 1$. Since $R^{(j)} < \frac{\alpha}{J-j+1}$, these hypotheses will all be rejected in $\mathcal{C}$.

Next, let us assume that $H^{(j+1)}$ is not rejected sequentially. This means that $R^{(j+1)} \geq \frac{\alpha}{J-j}$. So long as $j < J$, there exists an S such that $H_S \subset H^{(j+1)}$ and $|S| = J - j$. Since $R^{(j+1)} \leq R^{(j+2)} \leq \cdots \leq R^{(J)}$, the rejection of this hypothesis in $\mathcal{C}$ would require $R^{(j+1)} < \frac{\alpha}{J-j}$, contrary to our assumption. Therefore, rejection of a hypothesis in the sequential procedure occurs iff rejection also occurs in the closed procedure $\mathcal{C}$, and the two procedures are equivalent. The original derivation by Hölm [11] took a different and less intuitive approach.

Chapter 6

A Basic Version of the Limb-Leaf Design

Here we show a basic version of the Limb-Leaf Design using the Horizontal Rule. In Section 6.1 we present the plan of the experiment and make a restriction on the adaptation procedure. In Section 6.2 we characterize the dose response curves for which we intend to use the Limb Leaf Design as those with "locatable" effects. The advantages of using the Horizontal Rule in a Limb-Leaf Design are demonstrated by simulation in Section 6.3. Further components such as a template adaptation strategy, early stopping for futility, optimization of parameters, and sample size adjustments are discussed in Chapter 7.

6.1 Description of the Design

Consider that the set of limb doses $\mathcal{L}_{\text{full}}$ and the set of leaf doses ℓ_{full} are to be investigated. The Limb-Leaf experiment consists of four steps: (1) a first stage where n_1 patients are randomized to each limb and to the control; (2) an interim analysis where it is decided which, if any, limb to promote to the second stage and which leaf doses to add; (3) a second stage where n_2 patients are randomized to each limb or leaf dose of interest and to the control; and (4) a final analysis, where $\hat{d}^*$ is selected and Horizontal Tests are used in a closed testing

procedure to control the familywise error rate for the decision on the efficacy of $\hat{d}^*$ at level α. It should be the best performing limb that is promoted and the best performing second stage dose that is recommended as $\hat{d}^*$. This is a safe assumption; as in other designs, if a different selection is made, the resulting procedure is conservative.

We set a restriction on our search strategy to improve the convenience of the Horizontal Test in a Limb-Leaf Design: of the limbs, at most one can be promoted to the second stage, and of the leaves, at most a prespecified number are allowed to be added. To allow exploration on both sides of a promising region we prefer to set this number, $n_{|\ell|}$, to 2.

In the above setting, consider the set of doses $D = \mathcal{L} \cup \ell$, with $\mathcal{L} \subset \mathcal{L}_{\text{full}}$, and $\ell \subset \ell_{\text{full}}$. With $n_{|\ell|} = 2$ and the distributional assumptions of Section 4.2, the Horizontal Test Ψ_D of $H_D : \theta_d = 0, \forall d \in D$ is given by these steps:

Definition 3 *For $|\ell\,| \leq 2$:*

1. *Set the unadapted version Ψ_D as:*

$$I\left\{\max\left\{n_1 Y_{1,L^*} + n_2 Y_{2,L^*} - E_D\,, k_D \max_{l\in\ell} n_2 Y_l\right\} > c_D\right\}$$

with constants:

$$E_D = E_{H_D}\left[n_1 Y_{1L^*} + n_2 Y_{2L^*}\right], k_D = \sqrt{\frac{var(n_1 Y_{1,L^*} + n_2 Y_{2,L^*})}{var\,(n_2 Y_l)}},\ \textit{and}\ c_D.$$

Here L^* indicates the limb promoted to the second stage. To calculate c_D we assume $L^* = argmax_{L\in\mathcal{L}} Y_{1L}$ and set the rejection probability under the null to α. If a different promotion decision is made, the test will be conservative.

2. *Based on the first stage results, decide which $L \in \mathcal{L}$, if any, to promote and which subset $\ell' \subseteq \ell$ to include in the second stage. Calculate c'_D to preserve the conditional rejection probability at Step 3 as in Definition 1.*

3. *If a limb has been promoted, apply the rejection condition:*

$$I\left\{\max\left\{n_1 y_{1,L^*} + n_2 y_{2.L^*} - E_{\mathcal{L},\ell}\,, k_D \max_{l\in\ell'} n_2 y_l\right\} > c'_D\right\}$$

If no limb has been promoted, the corresponding term is left out.

For $|\ell| > 2$ the restriction on the number of leaves requires a modification. Let $\ell = \{l_1, l_2, ..., l_{|\ell|}\}$, ordered arbitrarily.

1. *The unadapted version Ψ_D is set as:*

$$I\left\{\max\left\{n_1 Y_{1,L^*} + n_2 Y_{2,L^*} - E_D, k_D \max_{l=l_1,l_2} n_2 Y_l\right\} > c_D\right\}.$$

   L^*, E_D, k_D, and $c_{\mathcal{L},\ell}$ are defined as above and the value of c_D does not depend on the arbitrary choices of l_1 and l_2.

2. *Based on the first stage results, decide which limb, if any, to promote and which set $\ell' \subseteq \ell$, $|\ell'| \leq 2$, to study. Calculate c'_D to preserve the conditional rejection probability at Step 3 as in Definition 1. When some $L^* \in \mathcal{L}$ is promoted and $|\ell'| = 2$, c'_D will be unchanged from c_D.*

3. *The rejection criterion is again*

$$I\left\{\max\left\{n_1 y_{1,L^*} + n_2 y_{2,L^*} - E_D\,, k_D \max_{l\in\ell'} n_2 y_l\right\} > c'_D\right\}$$

   (excluding the term involving L^* if no limb appears in the second stage).

Proposition 4, Chapter 5, allows the use of this test in a closed procedure of given familywise error rate α.

To aid in computation, we classify the tests involved in a rejection decision in the closed testing procedure and give a bound to the number of tests that are sufficient to consider. This bound is independent of what adaptations have been performed. Consider the confirmation of activity on a leaf l^* in the overall procedure, and all of the tests that need to be applied at the end of the experiment. Any of these corresponds to a set of doses that may be denoted as $D^* = \{\mathcal{L}^* \cup \ell^*\}$. Possible values $|\mathcal{L}^*|$ may take are $0, \ldots, |\mathcal{L}_{\text{full}}|$. Further, there exists the indicator $B \in \{0, 1\}$ of whether a promoted limb, L^*, if one exists, belongs to $\mathcal{L}^*$. Finally,

there is the number of leaves considered, $|\ell^*|$. We say that a test Ψ_{D_1} concerning the dose set D_1 dominates the test Ψ_{D_2} concerning D_2 if a rejection by Ψ_{D_1} implies a rejection by Ψ_{D_2}. Obviously, two tests may be equivalent by dominating each other. The bound is based on constructing a sufficient collection of tests associated with l^* such that any other test involved in the inference on l^* is dominated by a member of this collection. The size of this collection is then shown to be linear in $|\mathcal{L}_{\text{full}}|$ and $n_{|\ell|}$.

We note from Definition 3 that for any $D^* = \mathcal{L}^* \cup \ell^*$ with $|\ell^*| > 2$, Φ_{D^*} is dominated by (equivalent to) a $\Phi_{D^\dagger}$ for some $D^\dagger$ with the same $\mathcal{L}^*$ and B, but with $|\ell^*| = 2$. We therefore eliminate the tests of all D^* with $|\ell^*| > 2$ from our sufficient collection.

A class of sets $\{D^*\}_{|\mathcal{L}^*|,B,|\ell^*|}$ is determined by any chosen values of $|\mathcal{L}^*|$, B, and $|\ell^*|$. For a given D^* within that class, let τ be the number of elements of ℓ^* that appear in the second stage. Since the total collection of all hypotheses to consider in the closed testing procedure has the form of a Cartesian product between all collections of limb doses and all collections of leaf doses, it is possible to extremalize over each dimension independently. Thus, for $\{D^*\}_{|\mathcal{L}^*|,B,|\ell^*|}$ we may select a $D^+ = \mathcal{L}^+ \cup \ell^+$ such that τ is maximized and the first stage results, Y_{L_j} for $L_j \in \mathcal{L}^+$ are minimized. Then, by Definition 3, Ψ_{D^+} dominates every other member of this class.

There are at most $|\mathcal{L}_{\text{full}}| + 1$ possibilities for $|\mathcal{L}^*|$, at most two possibilities for B, and $n_{|\ell|}$ possibilities for $|\ell^*|$. Therefore, the bound on the size of a sufficient class of tests is $2(|\mathcal{L}_{\text{full}}| + 1)n_{|\ell|}$. The same reasoning applies to any selected limb, L^*, except that it is not necessary to consider the case $B = 0$, and the bound is $(|\mathcal{L}| + 1)n_{|\ell|}$.

It is possible to reduce the sufficient set of tests further, however the form of this bound already has an important property: it is independent of the total number of leaves $|\ell_{\text{full}}|$, and only depends on the largest number that can be added, $n_{|\ell|}$. A consequence is that $|\ell_{\text{full}}|$ can go to infinity, without affecting the bound on the number of tests in the sufficient collection. This means that "tuning" of second stage limbs may be performed: an arbitrarily dense collection of leaf doses may be specified, the second stage additions may be chosen to

lie anywhere within this collection, and the upper bound on the number of tests to perform for final inference is unaffected.

6.2 Locatable Effects in Dose Response Profiles

The assumption we will make in the Limb-Leaf Design is that the choice of the limbs, leaves, and initial parameters is appropriate for an adaptive selection strategy to access a dose level with the desired effect. We wish to state this more formally. Let the limb-leaf system $\mathcal{S} = \{\mathcal{L}_{\text{full}} \cup \ell_{\text{full}}\}$ be as described in Section 2.1 and let the parameter vector $\vec{\delta}$ be the vector of constants $(\delta_1, \delta_2, \delta_3, \delta_4)$, with $\delta_1 < \delta_2 < \delta_3 < \delta_4$. Let $\mathcal{D}$ denote the underlying dose response curve and let the effect for any specific dose d be θ_d. The need of the Limb-Leaf Design is that the desired level of effect (δ_4) be locatable with respect to $\mathcal{S}$ and $\vec{\delta}$.

Definition 4 *The dose response curve $\mathcal{D}$ is said to have a locatable effect with respect to $\mathcal{S}$ and the vector $\vec{\delta}$ if two conditions are satisfied:*

1. *Existence of a promising region: The limb doses may be classified as "promising" or "unpromising", according to whether $\theta_{L_i} > \delta_2$ or $\theta_{L_i} < \delta_1$ for each $i \in \{1, ..., |\mathcal{L}_{\text{full}}|\}$, with at least one $\theta_{L_i} > \delta_1$.*

2. *Existence of desired effects within promising regions: For any promising limb dose L_i, all $d \in \{L_i, l_{i,1}, \ldots, l_{i,M_i}\}$ can be classified as having the desired effect, if $\theta_d > \delta_4$, or lacking the desired effect, if $\theta_d < \delta_3$. For any promising L_i, there is at least one $d \in \{L_i, l_{i,1}, \ldots, l_{i,M_i}\}$ with the desired effect.*

Given the seeming strength of these assumptions, we illustrate several of the common situations in which they apply. Figure 2a (see Appendix) shows examples of locatable effects on either limb or leaf doses, with the limb-leaf system and the vector of effects held constant. Figure 2b shows departures from locatability caused by misspecifications of limb and leaf locations or by misspecifications of parameters. It is important to note that any of these

departures could be put in a form with a locatable effect by a different choice of the limb-leaf system and/or a different choice of parameter vector. In Section 8.2.2 we examine the effects of departures from these assumptions or misspecifications of parameters in a fully implemented Limb-Leaf Design and show that the consequences are often mild.

6.3 Simulation Results

Here we investigate the performance of the basic version of the Limb-Leaf Design using Horizontal Tests and compare it with procedures using the Weighted Rule or Maximum Rule. The goal is to show that the properties of the Horizontal Test seen in Section 4.2 lead to a favorable balance between robustness and efficiency in the multiple testing procedure. There are many aspects of the performance to consider: power on limb doses as well as leaf doses must be investigated under different numbers of limbs and leaves, parameter configurations, and adaptation decisions. The simulation studies reported here consist of several cases. These have been chosen because they are the most plausible in our intended use of the Limb-Leaf design and the most revealing of the performance characteristics. These examples all deal with locatable effects.

Specifically, there are two dimensions to explore in terms of power: power to confirm the effect on a limb, and power to confirm the effect on a leaf dose. There are three possible types of adaptations: to forward a chosen limb without addition of leaves, to forward with addition of leaves, and to promote leaves without the corresponding limb. These adaptations are to be made data dependently; however, we consider fixed adaptation strategies for this section in order to isolate the performance of the multiple test procedure from that of the adaptation rule.

The eight simulation studies presented in Table 1a cover different adaptations with $|\mathcal{L}_{\text{full}}| = 2,$ or 3, and $n_\ell = 2$. The generic parameters chosen for this study are $n_1 = n_2 = 100$, and $V_1 = \sigma^2/n_1 = V_2 = \sigma^2/n_2 = .01$; parameter optimization and other components are deferred to Chapter 7. The powers reported here allow us to compare the multiple testing

procedures using the Horizontal (H), Weighted (W), and Maximum (M) Rules and to confirm that the results are consistent with those seen in Section 4.3

Scenario I (Power on a Leaf with Promotion): This scenario investigates the power to confirm an effect on a leaf dose in the case of promotion of a first stage limb and addition of two leaves. The situation for $|\mathcal{L}_{\text{full}}| = 2$ is as follows: $\mathcal{L}_{\text{full}} = \{\mathcal{L}_1, \mathcal{L}_2\}$, and $\ell_{\text{full}} = \{l_{1,1}, l_{1,2}, l_{2,1}, l_{2,2}\}$. The corresponding vectors of effects are $\{\delta_1, \delta_2\}$, and $\{\delta_1, \delta_1, \delta_2, \delta_4\}$, with $\delta_1 = 0, \delta_2 = .2 = 2\sqrt{V_1}$, and δ_4 allowed to vary. The selection rule is the following: At the first stage, the limb with greater observed effect is promoted and its leaves are added. At the end of the second stage, the arm with the greatest observed second stage effect is selected and a multiple testing procedure is conducted to confirm the effectiveness of that dose. Power is defined as the probability to correctly identify and confirm the true effect on leaf $l_{2,2}$.

Similarly, for $|\mathcal{L}_{\text{full}}| = 3$, $\mathcal{L}_{\text{full}} = \{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3\}$ and $\ell_{\text{full}} = \{l_{1,1}, l_{1,2}, l_{2,1}, l_{2,2}, l_{3,1}, l_{3,2}\}$. The vector of limb effects is $\{\delta_1, \delta_1, \delta_2\}$ and the vector of leaf effects is $\{\delta_1, \delta_1, \delta_1, \delta_1, \delta_2, \delta_4\}$, with $\delta_1 = 0, \delta_2 = .2$, and δ_4 allowed to vary. At the first stage, the limb with greatest observed effect is promoted and its leaves are added; at the end of the second stage, the arm with the greatest observed second stage effect is selected and a multiple testing procedure is conducted to confirm the effectiveness of that dose. The power is the probability to correctly identify and confirm the effect on l_6.

This scenario corresponds to one of the most likely implementations of a Limb-Leaf Design: after first stage results, it is decided to promote the best performing limb and to explore two leaves around it.

Scenario II (Power on a Leaf Without Promotion): This scenario covers the case in which leaves are explored but the underlying limb is discarded. The arms and associated parameters are as in Scenario I, and the adaptation rule is that two leaf doses are added

upon the best performing limb without promotion of that limb. The power reported is the probability to correctly identify and confirm the true effect on leaf $l_{2,2}$ (or $l_{3,2}$).

This scenario is relevant because first stage results may not be sufficiently promising on the best performing limb to promote it, but further exploration may still be worthwhile. This strategy represents the belief that one of the leaves outperforms the limb, and accepts the risk of only exploring these leaves in order to conserve resources.

Scenario III (Power on a Limb with Promotion): This Scenario concerns the confirmation of an effect on a limb dose when promotion occurs along with the addition of two leaves. For $|\mathcal{L}_{\text{full}}| = 2$, $\mathcal{L}_{\text{full}}$ and ℓ_{full} have vectors of effects $\{\delta_1, \delta_4\}$, and $\{\delta_1, \delta_1, \delta_2, \delta_2\}$, respectively, with $\delta_1 = 0, \delta_2 = .2$, and δ_4 allowed to vary. The adaptation rule is as in Scenario I: At the first stage, the limb with greater observed effect is promoted and its leaves are added. At the end of the second stage, the arm with the greatest observed second stage effect is selected and a multiple testing procedure is conducted to confirm the effect of that dose. The power given here is the power to correctly identify and confirm the desired effect on $\mathcal{L}_2$.

For $|\mathcal{L}_{\text{full}}| = 3$, $\mathcal{L}_{\text{full}}$ and ℓ_{full} have vectors of effects $\{\delta_1, \delta_1, \delta_4\}$ and $\{\delta_1, \delta_1, \delta_1, \delta_1, \delta_2, \delta_2\}$, respectively, with $\delta_1 = 0$, $\delta_2 = .2$, and δ_4 allowed to vary. The selection rule is as in Scenario I and the power explored is the power to correctly identify and confirm the effect on $\mathcal{L}_3$.

This scenario is a likely outcome of a Limb-Leaf Design: the desired signal was on one of the original limbs, however additional exploration was undertaken. While this kind of unnecessary exploration should be avoided with high probability, good performance in this case will be required.

Scenario IV (Power on a Limb Without Addition): Here we consider confirmation of an effect on a limb dose when leaves are not added. For $|\mathcal{L}_{\text{full}}| = 2$ (or $|\mathcal{L}_{\text{full}}| = 3$) the vector of the effects corresponding to $\mathcal{L}_{\text{full}}$ is $\{\delta_1, \delta_4\}$ ($\{\delta_1, \delta_1, \delta_4\}$), with $\delta_1 = 0$ and δ_4 allowed

to vary. The selection rule is that the best performing limb is selected at the end of the first stage, and promoted without addition to the second stage. The powers of the different multiple testing procedures to confirm the desired effect are reported in the table.

This is one of the most relevant possibilities, because if sufficiently promising activity is seen on a limb dose, no further exploration may be required. The goal would be to confirm the desired level of activity on the selected limb while using as few additional resources as possible.

Discussion and Conclusions: The conclusion of this simulation study is that the efficiency and robustness properties of the Horizontal Test seen in Section 4.3 carry over to the Limb-Leaf procedure as a whole. The procedure based on the Horizontal Test is an advantageous compromise; it has far better power for confirming an effect on a leaf than the procedure based on the Weighted Rule, and it suffers a smaller loss in power for confirming an effect on a limb than the procedure based on the Maximum Rule.

In Scenario I, the procedure using the Weighted Rule has unacceptably low power (maximum of .68). The procedure based on the Horizontal Rule has the desired robustness and outperforms the procedure based on the Maximum Rule by a small margin. The reported powers plateau before reaching 1 because they are limited by the correct selection probability. We consider it important to report power that combines the effects of the chosen multiple testing procedure and the correct selection probability as our primary results. However, since selection may occur by other criteria (cost, toxicity, or other factors) we also report power figures in which the correct selections are imposed from outside (Table 1b). The conclusions are the same.

Scenario II shows the same pattern as seen in Scenario I: the procedure based on the Weighted Rule has a dramatic loss in power and the procedure based on the Horizontal Rule outperforms the procedure based on the Maximum Rule. The influence of correct selection can be seen because the powers do not reach 1, and by comparison with the results under

enforced selection in Table 1b. As reported here, an inversion of the ordering of rules occurs because the Maximum Rule seems to outperform the Horizontal Rule for some large values of δ_4. However, at these parameter values both procedures have already attained the maximum power allowed by the correct selection probability, and the reported differences lie beneath the margin of error based on 10,000 runs of the simulation.

In Scenarios I and II, the choice of $\delta_1 = 0$ is consistent with the search strategy, and with the expectation that poor performance would exist on some dose levels. If $\delta_1 < 0$ were chosen, the loss of power for the procedure based on the Weighted Rule would be exacerbated. This loss of power diminishes with increasing δ_1.

In Scenario III, the procedure using the Weighted Rule has the best power, and the performance of the procedure using the Horizontal Rule is comparable. There is a modest loss of power from the use of the Maximum Rule, particularly in the range of powers between .7 and .9. The reported powers approach 1 because they are not limited by the correct selection probability in the same way as they were in Scenarios I and II (correct selection probability approaches 1 with large δ_4). It is also important to note the large difference in power between the limb and leaf doses. Further comments will be made below.

In Scenario IV, the procedure using the Weighted Rule again has the best power, and the procedure using the Maximum Rule suffers a meaningful loss of power in the range of powers between .7 and .9. The procedure using the Horizontal Rule comes closer to the efficiency of the procedure using the Weighted Rule than the procedure using the more robust Maximum Rule. As in Scenario III, there is a large difference between powers for limb and leaf doses.

Some discussion is needed of the cases presented and their simplifications. The case of power to confirm a limb or a leaf when one leaf is added is not shown because it is less common (though still possible) and the pattern of behavior is not qualitatively different from Scenarios I and III; one can expect the power on a leaf to be slightly greater than that reported in Scenario I, and the power on a limb to be slightly greater than that reported in Scenario III. The parameter choices of $n_1 = n_2 = 100$, $V_1 = \sigma^2/n_1 = V_2 = \sigma^2/n_2 = .01$ are

arbitrary; however, in any study it would be necessary to allow adequate first and second stage sample sizes to achieve a high correct selection probability. Since selection can occur in both stages, a severe imbalance between n_1 and n_2 would diminish overall performance for all procedures.

A general feature of Limb-Leaf Designs is that the power to confirm an effect depends on whether this effect is on a limb or a leaf and on how many leaves have been added. The disparity in power seen in these simulations is foreseeable given the imbalance in total sample size between the limb and leaf arms.. This imbalance is remedied in Chapter 7. However, instead of proposing one-size-fits-all sample sizes and parameter values, this study suggests that a conditional power implementation may be useful: initial values can be chosen and possibly readjusted based on which scenario is carried out. The selection of initial parameter values, their adjustment following an interim analysis, and other features such as early stopping are discussed in the next chapter.

TABLE 1a: PERFORMANCE CHARACTERISTICS

	Scenario I						Scenario II					
	$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$			$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$		
δ_4	W	M	H	W	M	H	W	M	H	W	M	H
.20	.081	**.112**	**.136**	.051	**.099**	**.130**	.099	**.160**	**.185**	.065	**.139**	**.168**
.25	.140	**.213**	**.245**	.095	**.190**	**.243**	.180	**.268**	**.285**	.111	**.234**	**.280**
.30	.221	**.312**	**.362**	.142	**.304**	**.361**	.257	**.396**	**.421**	.163	**.363**	**.405**
.35	.321	**.464**	**.502**	.223	**.430**	**.493**	.356	**.524**	**.563**	.253	**.497**	**.526**
.40	.442	**.614**	**.634**	.299	**.568**	**.620**	.444	**.651**	**.691**	.317	**.611**	**.640**
.45	.540	**.716**	**.741**	.407	**.666**	**.715**	.523	**.757**	**.775**	.381	**.712**	**.742**
.50	.608	**.802**	**.827**	.487	**.757**	**.785**	.580	**.833**	**.848**	.467	**.776**	**.795**
.55	.639	**.862**	**.880**	.484	**.811**	**.826**	.666	**.874**	**.890**	.493	**.817**	**.832**
.60	.673	**.893**	**.904**	.528	**.832**	**.844**	.695	**.901**	**.904**	.536	**.847**	**.853**
.65	.683	**.910**	**.912**	.505	**.854**	**.854**	.690	**.916**	**.912**	.535	**.857**	**.852**
	Scenario III						Scenario IV					
	$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$			$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$		
δ_4	W	M	H	W	M	H	W	M	H	W	M	H
.20	**.189**	.157	**.180**	**.169**	.150	**.157**	**.535**	.408	**.481**	**.468**	.375	**.400**
.25	**.346**	.291	**.331**	**.325**	.278	**.310**	**.721**	.594	**.677**	**.675**	.562	**.610**
.30	**.535**	.471	**.526**	**.520**	.449	**.509**	**.861**	.760	**.823**	**.831**	.732	**.782**
.35	**.713**	.654	**.703**	**.700**	.636	**.693**	**.938**	.879	**.921**	**.930**	.863	**.903**
.40	**.842**	.806	**.841**	**.831**	.791	**.832**	**.981**	.945	**.976**	**.973**	.938	**.961**
.45	**.925**	.902	**.922**	**.921**	.899	**.922**	**.995**	.981	**.994**	**.993**	.979	**.987**
.50	**.967**	.961	**.966**	**.967**	.956	**.962**	**.999**	.994	**.999**	**.998**	.994	**.994**
.55	**.987**	.985	**.986**	**.985**	.983	**.987**	**1.00**	.999	**1.00**	**1.00**	.998	**1.00**
.60	**.996**	.994	**.995**	**.996**	.994	**.994**	**1.00**	1.00	**1.00**	**1.00**	.999	**1.00**
.65	**.998**	.998	**.999**	**.999**	.999	**.999**	**1.00**	1.00	**1.00**	**1.00**	1.00	**1.00**

Based on 10,000 simulated experiments. Margin of error ≈ .006. Horizontal Rule and its competitor are shown in bold.

TABLE 1b: ENFORCED SELECTION

	Scenario I						Scenario II					
	$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$			$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$		
δ_4	W	M	H	W	M	H	W	M	H	W	M	H
.20	.211	**.306**	**.228**	.153	**.313**	**.265**	.218	**.338**	**.265**	.154	**.338**	**.243**
.25	.279	**.385**	**.363**	.187	**.378**	**.373**	.274	**.406**	**.401**	.199	**.419**	**.352**
.30	.365	**.486**	**.480**	.243	**.486**	**.541**	.344	**.515**	**.522**	.247	**.513**	**.489**
.35	.426	**.566**	**.627**	.326	**.603**	**.653**	.445	**.661**	**.650**	.328	**.636**	**.629**
.40	.549	**.714**	**.757**	.391	**.720**	**.781**	.547	**.755**	**.777**	.400	**.739**	**.754**
.45	.632	**.809**	**.856**	.479	**.818**	**.869**	.659	**.839**	**.865**	.492	**.837**	**.852**
.50	.666	**.896**	**.914**	.532	**.896**	**.932**	.706	**.912**	**.934**	.532	**.912**	**.923**
.55	.713	**.946**	**.963**	.575	**.938**	**.967**	.737	**.955**	**.971**	.580	**.952**	**.960**
.60	.752	**.977**	**.981**	.667	**.972**	**.986**	.741	**.979**	**.985**	.600	**.979**	**.982**
.65	.762	**.989**	**.993**	.631	**.989**	**.995**	.763	**.991**	**.995**	.630	**.991**	**.992**
	Scenario III						Scenario IV					
	$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$			$\lvert\mathcal{L}_{\text{full}}\rvert = 2$			$\lvert\mathcal{L}_{\text{full}}\rvert = 3$		
δ_4	W	M	H	W	M	H	W	M	H	W	M	H
.20	**.195**	.152	**.185**	**.390**	.334	**.350**	**.561**	.451	**.496**	**.524**	.424	**.437**
.25	**.353**	.287	**.336**	**.597**	.489	**.548**	**.738**	.626	**.690**	**.699**	.570	**.617**
.30	**.546**	.471	**.530**	**.753**	.658	**.726**	**.871**	.771	**.841**	**.855**	.737	**.784**
.35	**.711**	.651	**.708**	**.892**	.806	**.862**	**.946**	.877	**.932**	**.933**	.863	**.905**
.40	**.839**	.804	**.835**	**.955**	.900	**.947**	**.983**	.950	**.976**	**.980**	.945	**.968**
.45	**.920**	.903	**.915**	**.985**	.954	**.984**	**.996**	.981	**.995**	**.995**	.978	**.990**
.50	**.965**	.958	**.968**	**.997**	.986	**.995**	**.999**	.995	**.998**	**.999**	.993	**.997**
.55	**.986**	.985	**.987**	**.999**	.996	**.999**	**1.00**	.999	**1.00**	**.999**	.999	**.999**
.60	**.996**	.995	**.995**	**.999**	1.00	**1.00**	**1.00**	1.00	**1.00**	**1.00**	1.00	**1.00**
.65	**.999**	.999	**.999**	**1.00**	1.00	**1.00**	**1.00**	1.00	**1.00**	**1.00**	1.00	**1.00**

Based on 10,000 simulated experiments. Margin of error ≈ .006. Horizontal Rule and its competitor are shown in bold.

Chapter 7

Additional Details of the Limb-Leaf Design

The basic Limb-Leaf Design was described in Section 6.1. Here we introduce further components that are necessary to implement it. We discuss (1) a reasonable exploration strategy, (2) the use of an early stoping boundary for futility, (3) the choice of initial parameter values, and (4) sample size adjustments at interim. These features are implemented in the example of Section 8.2.

7.1 An Exploration Strategy

Assuming a dose response curve with a locatable effect as in Section 6.2, we propose a strategy to explore the curve in order to identify and confirm a dose with this desired effect. This is basic guidance; the actual strategy could differ based on the needs or characteristics of any particular trial. Let θ^* be the desired effect, and $\hat{\theta}_{1,L_j}, j = 1, \ldots, |\mathcal{L}_{\text{full}}|$, be the estimated effects on limb doses from the first stage data. In the previous case of normal data with known variance, $\hat{\theta}_{1,L_j} = Y_{1,L_j} = X_{1,L_j} - X_{1,0}$. In the general score method, $\mathcal{Z}_{1,L_j} \overset{approx}{\sim} N(\theta_{1,L_j}\mathcal{V}_1, \mathcal{V}_1)$, and we may use $\hat{\theta}_{1,L_j} = \mathcal{Z}_{1,L_j}/\mathcal{V}_1$. Let $\hat{\theta}^*$ be the effect estimate on the selected limb L^*, and let $\vec{a} = (a_1, a_2, a_3)$, with $a_1 < a_2 < a_3$, express three levels of

evidence as follows. The first component, a_1, is the minimum threshold for the plausibility of the desired effect within a neighborhood of L^*; if $\hat{\theta}^* < a_1$ further study is not indicated. The second, a_2, is the threshold for likely activity in the neighborhood of L^*; if $a_1 \leq \hat{\theta}^* < a_2$, exploration in the neighborhood of L^* may be justified, without investigating L^* itself. In this case we may, for instance, discontinue L^* but add two of its nearby leaves (as considered in Scenario II, Section 6.3). The third level, a_3, is the threshold for a strong indication of the desired activity on L^*. If $a_2 \leq \hat{\theta}^* < a_3$ then we may choose to promote L^* with addition of two leaves (Scenario I/III). If $\hat{\theta}^* > a_3$, we may consider the evidence of activity on L^* strong enough to not do any further exploration and concentrate on confirming this activity with the minimum additional effort (Scenario IV). This proposed strategy is summarized in the table below.

TABLE 2: SEARCH STRATEGY

I:	$\hat{\theta}^* \geq a_3$	Strong indication- Continue with L^* without adding leaves.
II:	$a_2 \leq \hat{\theta}^* < a_3$	Likely indication- Promote L^* and add two adjacent leaves.
III:	$a_1 \leq \hat{\theta}^* < a_2$	Weak indication- Do not promote L^*, and add two nearby leaves (further exploration).
IV:	$\hat{\theta}^* < a_1$	Insufficient indication- Stop the study early and do not spend additional resources.

Criteria for selecting $\vec{a}$ and the initial sample sizes per arm are discussed in Section 7.3. Each decision (I-III) may be taken with desired conditional power as discussed in Section 7.4.

7.2 Early Stopping

Early stopping contributes to the efficiency of the Limb-Leaf Design. Termination of an ineffective treatment after a crude examination of the dose response is a main benefit of this approach. To integrate this early stopping feature into the test procedure, the Horizontal Test may be modified as follows.

Definition 5 *(Horizontal Test with early stopping)*

For $|\ell\,| \leq 2$*, set the unadapted version* $\Psi_{\mathcal{D}}$ *as:*

$$I\left(\left\{\max\left[n_1 Y_{1,L^*} + n_2 Y_{2,L^*} - E_D\,, k_D \max_{l\in\ell} n_2 Y_l\right] > c_D^1\right\} \cap \left\{n_1 Y_{1,L^*} > c_D^0\right\}\right).$$

Similarly, for $|\ell\,| \leq 2$*, the unadapted version* Ψ_D *is set as:*

$$I\left(\left\{\max\left[n_1 Y_{1,L^*} + n_2 Y_{2,L^*} - E_D\,, k_D \max_{l\in\{l_1,l_2\}} n_2 Y_l\right] > c_D^1\right\} \cap \left\{n_1 Y_{1L^*} > c_D^0\right\}\right).$$

Here L^* *indicates the promoted limb;* E_D*,* k_D*,* l_1*, and* l_2 *are as in Definition 3; and* c_D^1 *and* c_D^0 *are such that the type I error probability is* α *under the assumption that* $L^* = argmax_{L\in\mathcal{L}} Y_{1L}$*. The test then proceeds as in Definition 3.*

We set the test of the global null according to Definition 5, with $c_D^0 = n_1 a_1$, and c_D^1 determined by the type 1 error constraint. This modification enforces the early stopping given in Table II. Specifically, when the maximum of the first stage estimated effects is below a_1, the global test fails. Then, it will not be possible to reject any individual hypothesis in the overall procedure and the study may end.

It is also possible to modify other tests in the overall procedure according to Definition 5. However, we do not follow this route because gains in power from such modifications are minimal. Such early stopping of tests for certain combination hypotheses can also be questioned: Consider a test $\Psi_{D^*=\mathcal{L}^*\cup\ell^*}$ when some $l \in \ell^*$ is not associated with any $L \in \mathcal{L}^*$. Given that there is poor first stage evidence of effect on the limbs, it may or may not be considered appropriate to terminate the test of the combination hypothesis. In the worst case, a heavy bias towards early stopping based on the limbs could prevent the confirmation of a true effect on l^*. In the case of the global null hypothesis, though, the early stopping decision is well motivated.

7.3 Initial Parameter Values

The initial parameters that need to be set are $\vec{a}$, n_{1L}, n_{2L}, and n_{2l}. Here n_{1L} is the sample size per arm in the first stage, and n_{2L} and n_{2l} are the sample sizes per arm on the limb/control and leaf doses in the second stage, respectively. In general we allow these sample sizes to be different. We assume that the number of limbs, $|\mathcal{L}_{\text{full}}|$, is given. The choice of $|\mathcal{L}_{\text{full}}|$ is discussed in Section 8.1.

The choice of a_1 is based on statistical and clinical concerns. Clinically, a_1 must be a meaningful threshold such that the investigators are willing to declare lack of efficacy if no first stage estimated effect exceeds a_1. Statistically, a_1 should be such that the probability of early stopping under the global null is sufficiently high ($\approx .4 - .6$). For $|\mathcal{L}_{\text{full}}| = 3$, a reasonable value for this parameter would be $.2\,a_3$. Once this value is set and incorporated into the test procedure, it may not be changed.

Conceptually, a_2 is analogous to δ_2 in Section 6.2. It is a level of response that is sufficiently promising on a given limb for the desired level of activity to be plausible on that limb (as opposed to only being plausible in the neighborhood). This parameter is tentative, and may be revised following the first stage as part of the flexible adaptation strategy. Reasonable initial values are in the range of $.6\,a_3$. An overestimate of a_2 may

throw away a promising limb, an underestimate could lead to the use of resources on a limb whose performance is beneath the desired level.

The target level of activity, a_3, should be chosen on clinical grounds. Statistically, an overly optimistic initial value could lead to an underpowered study. An overly pessimistic estimate could lead to an inefficiently large first stage sample size and a willingness to accept a meaningfully lower treatment effect that the true maximum. This parameter is not fixed, it can be modified using the first stage results.

For fixed choices of a_1, a_2, and a_3, initial values for the parameters n_{1L}, n_{2L}, and n_{2l} may be set based on desired performance characteristics. In setting these initial values, we can simplify by imposing additional constraints. We have set n_{1C}, the sample size of the control arm in the first stage, equal to n_{1L}, and n_{2C}, the sample size of the control arm in the second stage, equal to n_{2L}. We may also set n_{2L} equal to $\rho\, n_{2l}$, with $\rho = 2$, so that the total number of patients per arm on a leaf dose equals that on a limb. The initial design then depends on only one parameter (n).

This difference in sample size per arm between limb and leaf doses in the second stage should be reflected in the form of the horizontal test. In Definition 3 (equal samples per arm), Y_{2L^*} and Y_{2l} were both weighted by $n_2 = n_{2L} = n_{2l}$. Now, to reflect that the variance of Y_{2l} is proportional to $\frac{1}{n_{2l}} + \frac{1}{n_{2L}}$, we set $n'_{2l} = \frac{2n_{2L}n_{2l}}{n_{2L}+n_{2l}}$, and set the unadapted test as

$$I\left\{\max\left\{n_1 Y_{1L^*} + n_2 Y_{2L^*} - E_D\,, k_D \max_{l\in\ell'} n'_{2l} Y_{2l}\right\} > c_D\right\}.$$

We make the corresponding change of n_{2l} to n'_{2l} in the adapted test as well.

The initial performance target can be expressed as the power to identify and confirm locatable effects under two parameter configurations. The first places the desired effect on one of the limbs; for $|\mathcal{L}_{\text{full}}| = 2$ and $|\ell_{\text{full}}| = 4$, the vectors of effects on $\{L_1, L_2\}$ and $\{l_{1,1}, l_{1,2}, l_{2,1}, l_{2,2}\}$ could be chosen as $(0, a_3)$ and $(0, 0, a_2, a_2)$, respectively. Similarly, for $|\mathcal{L}_{\text{full}}| = 3$ and $|\ell_{\text{full}}| = 6$, the vectors of effects for $\{L_1, L_2, L_3\}$ and $\{l_{1,1}, l_{1,2}, l_{2,1}, l_{2,2}, l_{3,1}, l_{3,2}\}$ could be $(0, 0, a_3)$ and $(0, 0, 0, 0, a_2, a_2)$. The second places the desired effect on the leaves:

For $|\mathcal{L}_{\text{full}}| = 2$ and $|\ell_{\text{full}}| = 4$, the vectors of effects on $\{L_1, L_2\}$ and $\{l_{1,1}, l_{1,2}, l_{2,1}, l_{2,2}\}$ could be $(0, a_2)$ and $(0, 0, a_2, a_3)$. For $|\mathcal{L}_{\text{full}}| = 3$ and $|\ell_{\text{full}}| = 6$, possible vectors of effects on $\{\mathcal{L}_1, \mathcal{L}_2, \mathcal{L}_3\}$ and $\{l_{1,1}, l_{1,2}, l_{2,1} l_{2,2} l_{3,1}, l_{3,2}\}$ are $(0, 0, a_2)$ and $(0, 0, 0, 0, a_2, a_3)$. Reasonable choices for power in these configurations are between .8 and .9, and the desired power on a limb may be chosen to be greater than the power on a leaf (.9 versus .85, for instance). In Section 8.2, we identify the "least favorable configurations with locatable effects" and set the power in each case to be .9.

The power to identify and confirm the desired effect under the search strategy of Table II is then an increasing function of n. To set an initial value, two approaches are suggested. The simpler is to choose the minimum n, $n_{\min}$, that meets the specified power conditions; minimizing n is equivalent to minimizing the maximum of the total sample size, N. Since a closed form solution is impractical, a direct simulation can be used to find this minimal value. We call this a "minimax-style" parameter selection because it is not strictly a minimax solution. A formal solution uses decision theory.

Assume a finite set $\Theta = \{\vec{\theta}_t, t \in \mathcal{T}\}$ of possible dose-response vectors, where each $\vec{\theta}_t$ represents a vector of effects across the limbs and leaves of the experiment. Let the prior Π assign probability p_t to each $\vec{\theta}_t$. For specific n, denote a decision rule corresponding to the search strategy in Table II by δ_n^*; a value of this rule then corresponds to an adaptation decision based on the first stage data. The loss function is equal to the total sample size expended in the entire experiment.

We may express the Bayes' Risk of δ_n^*, $\mathcal{R}_{\delta_n^*}$, as the risk adjusted expected total sample size:

$$\mathcal{R}_{\delta_n^*} = E_\Pi(N) = \sum_{t \in \mathcal{T}} p_t E_{\vec{\theta}_t}(N),$$

where $E_{\vec{\theta}_t}$ is the expectation with respect to the parameter vector $\vec{\theta}_i$, and each term of the sum is evaluated by simulation. We may select n to minimize the Bayes' Risk subject to the above power constraints (restricted Bayes solution). For given δ_n^* we may also evaluate the

maximum risk over Θ and find the (restricted) minimax solution.

The optimality of any initial choice of n depends on prior assumptions. Sample sizes, conditional power, and the relative balance between limb and leaf arms may be adjusted after the first stage as shown in Section 7.4.

7.4 Sample Size Adjustment

Adjustment of sample sizes per arm based on conditional power is a useful feature of the Limb-Leaf Design. One reason is that the sample size requirements to achieve desired conditional power under the different adaptation decisions may differ substantially. For example, in case IV of Table II there is only one dose level of interest for the second stage, whereas in case III there are three. There is no need for correct selection in the second stage of case IV and a lower n_{2L} could be used to achieve a similar conditional power. Other reasons to use conditional power include the desire to balance the powers to confirm limb and leaf effects, and the need to adjust tests based on first stage performance. For example, if interim results indicate that the choice of sample sizes based on the performance criteria stated in Section 7.3 is overly optimistic or overly pessimistic, sample sizes could be increased or decreased respectively. Finally, it may be desired to alter the relative balance between limbs and leaves from the default value $\rho = 2$. For these needs, we offer two types of modifications.

The first modification preserves the structure of the tests and the associated cutoff values. Specifically, we refer to the rejection criterion of the adapted test as shown in Item 3 of Definition 3:

$$I\left\{\max\left\{n_1 Y_{1L^*} + n_2 Y_{2L^*} - E_D\,,\, k_D \max_{l\in\ell'} n'_{2l} Y_{2l}\right\} > c'_D\right\}.$$

The limb and leaf sample sizes may be increased by a constant factor c, and the resulting test statistics can be labeled Y'_{2L^*} and Y'_{2l}, for $l \in \ell'$. Then, substitution of $\sqrt{c}\,Y'_{2L^*}$ for Y_{2L^*}, and $\sqrt{c}\,Y'_{2l}, l \in \ell'$ for $Y_{2l}, l \in \ell'$ in the original test does not alter the multivariate null distribution or the resulting cutoff values. The alternative distribution changes in predictable

way- each effect size θ is multiplied by $\sqrt{c}$. Simulation would be used to set the appropriate sample size for desired conditional power. In this case only the parameter c needs to be increased and the tests themselves do not need to be modified, until the desired conditional power is reached.

In the second option it would also possible to modify ρ (or even change the balance between limb and control sample sizes). Such a modification requires the computation of new cutoff values based on the new covariance structure. One computation is sufficient, though, because after new cutoff values are found, additional increases in sample size can be expressed using a scaling constant, and the test statistics may be modified as previously (without further changes to the covariance structure).

Chapter 8

Comparison with a Standard Approach

We have proposed that a Limb-Leaf Design can further improve on the benefits of adaptive seamless designs in terms of the speed, efficiency, and streamlining of a drug development program. In Section 8.1 we present some guidance on when to consider the Limb-Leaf Design. In Section 8.2 we evaluate the performance and the assumptions of a limb-leaf strategy using an explicit example. The benefits over a standard TSE-type design are seen to be large under the global null hypothesis and also under the alternative when the first stage limbs are well chosen to detect the desired level of effect.

8.1 Considerations for the Limb-Leaf Option

The Limb-Leaf Design is suited to cases in which it is necessary to identify an effect of the desired level that may only exist in a narrow dose range, or when extracting the additional benefits from a closer study of the dose response curve is considered worthwhile. This could be the case in studying a new treatment for a severe disease if existing therapies are ineffective or marginally effective, the benefit from any new treatment may only be small, and the assumption of a monotonic dose response curve is not appropriate. Section 2.2

identifies ALS research as an area in which this study design may be useful.

The appropriate use of a Limb-Leaf Design assumes an alternative $\mathcal{D}$ with a locatable effect with respect to the chosen limb-leaf system $\mathcal{S}$ and parameter vector $\vec{\delta}$. Sources of information on which to base this assumption include clinical judgement, animal models, experience with related drugs, existing literature and case reports, pharmacodynamic models, and Phase I studies within the same drug development program. Since there is greater efficiency when the desired effect exists on a limb of $\mathcal{S}$, it is further hoped that the prior knowledge would give a significant chance of capturing a desired effect within $\mathcal{L}$.

We do not foresee that the logistical and operational needs of a Limb-Leaf Design will be substantially different from those of other adaptive seamless designs (as discussed by Maca et al. [28]), except that the rate of recruitment may need to change depending on how many doses are of interest in the second stage.

8.2 A Comparison by Simulation with a TSE-type Design

We now show a comparison between a Limb-Leaf Design and a TSE-type adaptive seamless design. For equivalent power, the Limb-Leaf Design has considerably lower risk adjusted expected sample size, with savings possible under both the null and alternative configurations. The robustness of the Limb-Leaf Design's performance to departures from assumptions and misspecifications of parameters is investigated in Section 8.2.2.

8.2.1 Designs and Choices of Parameters

Consider that a new drug is under development and that the dose response curve may not be assumed to be monotonic. It is hypothesized that the dose response curve is unimodal and that the drug effect only exists in a narrow dose range. However, strictly increasing, plateau, or constant profiles are also considered plausible.

We will choose the simplest possible assumptions to emphasize the generality of this example. Let the outcome x at any dose level d_i be distributed as $N(\theta_{d_i}, 1)$, where θ_{d_i} is a level of effect with respect to the control and the desired level of effect is 1. Desired power, defined as the probability to select and confirm an effective dose, is .9. The familywise error rate with respect to the hypotheses of no effect at each dose will be controlled at $\alpha = .05$. As part of the specification of power, it is required to distinguish the desired effect from effects that are 30 percent lower; an effect of .7 or lower will be considered unacceptable and if it is mistakenly chosen when the desired effect (≥ 1) is present, the experiment will not be considered a success.

Let a collection of 9 (generally unequally spaced) doses be considered appropriate to identify the desired level of effect. Let these doses be organized as the limb-leaf system $\mathcal{S} = \{l_{1,1}, L_1, l_{1,2}, l_{2,1}, L_2, l_{2,2}, l_{3,1}, L_3, l_{3,2}\}$, with the doses listed in increasing order. We are willing to assume that under the alternative the desired effect of 1 is locatable, and that the level of .625 corresponds to a promising effect as in Definition 4.

To finish the specification of the class of alternatives, we set $\vec{\delta} = (.125, .625, .625, 1)$. Although the choice $\delta_1 = 0$ would favor the Limb-Leaf Design, we are cautious in setting a low "background level" for the effect of the treatment. The assumption of $\delta_3 \downarrow \delta_2$ is a simplification and a conservative choice. It would be correct to set δ_3 to any value between .625 and .7; however, the sample size requirements of the standard design were found to increase rapidly with δ_3. The lowest allowable value of δ_3 is chosen in favor of the standard design.

Examples of possible response levels $\vec{\theta} = (\theta_{l_{1,1}}, \theta_{L_1}, \theta_{l_{1,2}}, \theta_{l_{2,1}}, \theta_{L_2}, \theta_{l_{2,2}}, \theta_{l_{3,1}}, \theta_{L_3}, \theta_{l_{3,2}})$ corresponding to these ordered doses include $(.625, 1, .625, 0, 0, 0, 0, 0, 0)$, $(.125, .625, 1, .625, .125, .25, 0, 0, 0)$, and $(.125, .125, .125, .400, .625, 1, 1, 1, 1)$, among others such as $(0, 0, 0, .1, .25, .5, .625, 1, 1.25)$. The robustness of the Limb-Leaf Design to departures from these assumptions and/or misspecifications of parameters is given in Section 8.2.2. The global null is for no effect at any dose.

The TSE design [42] (described in Section 1.2) is modified to use normal outcomes as described by Jennison and Turnbull in [18]. It has two stages; the first stage assigns subjects to all nine treatments plus the control, and the second stage studies only the best performing dose from the first stage against the control. There is the option to stop for futility using a cutoff value after the first stage, and the final decision is made by whether the combined measure of effect on the selected treatment exceeds a second cutoff value. The parameters to set are: the first stage sample size per arm, n_1; the second stage sample size per arm, n_2; the first stage futility boundary, y_1; and the second stage cutoff value for efficacy, y_2.

A strict implementation of the TSE design would use a least favorable configuration of $\vec{\theta}_{\text{lf}}= (.625, .625, .625, .625, .625, .625, .625, .625, 1)$, to set the power. However, given our assumption that the alternative is in the subclass with locatable effects, this criterion should be modified; to neglect the assumption of locatable effects leads to a sample size that is an order of magnitude larger. By the argument given in [18], among all alternatives with a single locatable effect within $\mathcal{S}$ with respect to $\vec{\delta}$, the power is (non-uniquely) minimized by the vector of effects $\vec{\theta}_{\text{loc}} = (.125, .125, .125, .125, .125, .125, .625, .625, 1)$, which we call the "least favorable configuration with a locatable effect with respect to $\vec{\delta}$". It is clear (and easily verified by simulation) that the minimizers of power for two and three locatable effects, (.125,.125,.125,.625,.625,1,.625,.625,1), and (.625,.625,1,.625,.625,1,.625,.625,1), lead to higher power. We therefore set the power to .9 in the least favorable (locatable effect) configuration.

We set this design's parameters according to the original paper by minimizing the risk adjusted expected sample size $E_\pi N = \pi E_{H_A} N + (1 - \pi) E_{H_0} N$. Here H_A represents the alternative hypothesis of the least favorable (locatable effect) configuration, H_0 represents the global null hypothesis, and π is a prior probability assigned to the alternative hypothesis. Specifically, for given n_1 and n_2, y_1 and y_2 are set by the type 1 error and power constraints (when solutions exist). Then $E_\pi N$ can be minimized over n_1 and n_2. Optimization using simulation of expected sample sizes for $\pi = .2$ yields $n_1 = 36$ and $n_2 = 11$.

We implement the Limb-Leaf Design as discussed in Chapter 7. The search strategy in Section 7.1 is followed: the three limbs begin the first stage with options of early stopping, addition, promotion with addition, or promotion without addition into the second stage depending on the maximum observed effect at interim. In the case of promotion without addition (when the desired effect is seen at the first stage), we adapt the second stage sample size for the selected dose and the control according to Section 7.4. We let the adapted second stage sample size be $n_{2L_{adapted}} = \frac{1}{2}n_{2L}$; a more exact modification could potentially lead to greater efficiency.

Among all alternatives with a single locatable effect with respect to $\mathcal{S}$ and $\vec{\delta}$, the power for an effect on a leaf is (non-uniquely) minimized by the vector of effects $\vec{\theta}_{\text{leaf}} = (.125, .125, .125, .125, .125, .125, .625, .625, 1)$, which is the "least favorable configuration with a locatable leaf effect with respect to $\vec{\delta}$". As verified by simulation, the minimizers of power for two and three locatable leaf effects, (.125,.125,.125,.625,.625,1,.625,.625,1), and $(.625, .625, 1, .625, .625, 1, .625, .625, 1)$, respectively, lead to higher power. Similarly, the least favorable configuration with a locatable limb effect with respect to $\vec{\delta}$ is $\vec{\theta}_{\text{limb}} = (.125, .125, .125, .125, .125, .125, .625, 1, .625)$. By enforcing power of .9 in the least favorable configurations for limb and leaf effects, we meet the same power requirement that was used in the previous design.

The parameters to set for the Limb-Leaf Design are n_1, n_{2L}, and $\vec{a}$. The choice of values will again be based on minimizing the risk adjusted expected sample size, here expressed as $E_\pi N = \pi_1 E_{H_{A_L}} N + \pi_2 E_{H_{A_l}} N + (1 - (\pi_1 + \pi_2)) E_{H_0} N$, with H_{A_L} and H_{A_l} as the least favorable configurations for effects on the limbs and leaves, respectively.

In this study, we have divided $\pi = Pr(\text{least favorable locatable effect}) = .2$ as $\pi_1 + \pi_2$, where $\pi_1 = Pr(\text{least favorable effect on limb}) = .1$ and $\pi_2 = Pr(\text{least favorable effect on leaf}) = .1$. In the standard design, there is no distinction between the doses that are labeled as limbs and those that are labeled as leaves. There this expression reduces to the previous form with $\pi = .2$. The comparison measures for these two designs are therefore consistent. We consider $\pi_1 = .1$ to be a conservative choice because a wise selection of limb doses could

result in a higher π_1.

We set the parameters of the Limb-Leaf Design by minimization under constraints: We set $n_1 = n_{2L} = \frac{1}{2}n_{2l}$ as suggested in Section 7.3, and choose $a_3 = \delta_4$ because it is reasonable to promote a limb without further exploration when the desired activity is already seen at interim. According to the advice in Section 7.3, we select .6 as the probability for early stopping under the global null. This determines a_1 as a function of n_1. Over the remaining parameters, n_1 and a_2, we minimize the risk adjusted expected sample size subject to the needs of .9 power for locatable effects on both the limbs and leaves (with respect to $\mathcal{S}$ and $\vec{\delta}$).

The results of numerical optimization are $n_1 = 42$, $n_2 = 85$, and $\vec{a} = (.175, .656, 1)$. This restricted optimum is sufficient to demonstrate the favorable performance of the Limb-Leaf Design. The performance of the competing designs given the above parameters is shown below.

TABLE 3: COMPARISON OF E_N VALUES

Least Favorable Locatable Configuration	**Limb-Leaf**	**Standard (TSE) Design**
Limb Effect	333.1	382.0
Leaf Effect	388.0	382.0
No Effect	254.3	364.8
Risk Adjusted	275.6	368.2

Results based on 10,000 simulated experiments.

The savings in risk adjusted expected sample size from using the limb-leaf approach is approximately 25%. The greatest savings occurs under the global null, where the investigation of the dose response curve beginning with three limb doses rather than all 9 possibilities is clearly efficient. We note that, in the case where the desired effect is on a leaf, the Limb-Leaf Design's expected total sample size slightly exceeds that of the standard design. The further

exploration consumes approximately the same number of patients as were saved in the first stage. From our point of view, this is an acceptable allocation of resources on an as-needed basis.

8.2.2 Robustness to Deviations

It is necessary to study the robustness of the Limb-Leaf Design to deviations from the assumed dose response profile. These results are condensed in Table III. As mentioned in Section 6.2, such deviations can often be considered either as perturbations of a specified dose response curve, or as misspecifications of the components of the vector $\vec{\delta}$. While it is impossible to do a complete exploration of all deviations, robust behavior in several plausible scenarios is important to show. Here we will fix the design parameters and vary the dose response profile.

The first case we consider is the Step Function dose response with $\vec{\theta}_S = (.125, .125, .125, .625, .625, .625, 1, 1, 1)$. The effect is not locatable because Condition 2 of Definition 4 is not satisfied; however the effect would be locatable for a different choice of $\vec{\delta}$ such as $(.625, .65, .7, 1)$. The performance is good under this alternative, with power=.963 and $E_N = 334.3$.

Two related cases are the Plateau dose response profile, with $\vec{\theta}_P = (0, .16, .33, .5, .66, .83, 1, 1, 1)$, and the Monotone Increasing profile with $\vec{\theta}_M = (0, .125, .250, .375, .500, .625, .750, .825, 1)$. The first has power .945 and expected sample size 333.6, the second has power .947 and expected sample size of 367.2.

Of greater concern is the Bimodal profile represented by $\vec{\theta}_{B_L} = (.125, .225, .125, .625, 1, .625, .125, .125, .125)$ and $\vec{\theta}_{B_l} = (.125, .225, .625, 1, .625, .125, .125, .125, .125)$. Here the decrease in power is greater when the desired effect is on a leaf rather than a limb. This is because of the risk of misdirecting the search strategy to explore the wrong region of the dose response curve. Results for power and expected sample size for $\vec{\theta}_{B_L}$ are .925 and 335.2, respectively; for $\vec{\theta}_{B_l}$ the power and expected sample size are .885 and 388.7, respec-

tively. A further increase in the magnitude of the secondary maximum, to $\vec{\theta}_{B_L} = (.125, .325, .125, .625, 1, .625, .125, .125, .125)$ and to $\vec{\theta}_{B_l} = (.125, .325, .625, 1, .625, .125, .125, .125, .125)$ yields power and expected sample size for $\vec{\theta}_{B_L}$ of .926 and 332.6, and power and expected sample size for $\vec{\theta}_{B_l}$ of .847 and 389.9.

Also of concern are response profiles that are "Unimodal with Background" such as $\vec{\theta}_{UB_L} = (.225, .225, .225, .625, 1, .625, .225, .225, .225)$ and $\vec{\theta}_{UB_l} = (.225, .225, .625, 1, .625, .225, .225, .225, .225)$. Power under $\vec{\theta}_{UB_L}$ is .927 and expected sample size is 335.0. For $\vec{\theta}_{UB_l}$, power is .870 and expected sample size is 388.5. With further distortion, represented by $\vec{\theta}_{UB_L} = (.325, .325, .325, .625, 1, .625, .325, .325, .325)$ and $\vec{\theta}_{UB_l} = (.325, .325, .625, 1, .625, .325, .325, .325, .325)$, performance to detect the signal on the leaf suffers further but power for a signal on the limb is robust. Power under $\vec{\theta}_{UB_L}$ is then .924 and expected sample size is 334.6. For $\vec{\theta}_{UB_l}$ the power is .792 and expected sample size is 389.4.

We also consider a decrease in the maximum effect. This could also be considered a misspecification of δ_4. The first case is when the decreased effect exists on a limb: for $\vec{\theta}_{D_L} = (.125, .225, .125, .625, .9, .625, .125, .125, .125)$ power is .805 and expected sample is 359.7. This is in keeping with the performance of the standard design, whose power is .790 and whose expected sample size is 381.8. When the depressed effect is on a leaf, the decay in power is not so great. For $\vec{\theta}_{D_l} = (.125, .125, .625, .9, .625, .125, .125, .125, .125)$, the power is .842 and the expected sample size is 387.6. This exceeds the performance of the standard design.

TABLE 4: ROBUSTNESS OF LIMB-LEAF DESIGN PERFORMANCE

Type of Response Curve/ Parameter Vector	Power	$E(N)$
Step Function [(1)]		
$\vec{\theta}_S = (.125, .125, .125, .625, .625, .625, 1, 1, 1)$	.963	334.3
Plateau [(1)]		
$\vec{\theta}_P = (0, .16, .33, .5, .66, .83, 1, 1, 1)$	.945	333.6
Monotone Increasing [(1)]		
$\vec{\theta}_M = (0, .125, .250, .375, .500, .625, .750, .875, 1)$	.947	367.2
Bimodal [(2)]		
$\vec{\theta}_{B_L} = (.125, .225, .125, .625, 1, .625, .125, .125, .125)$	.925	335.2
$\vec{\theta}_{B_l} = (.125, .225, .625, 1, .625, .125, .125, .125, .125)$	.885	388.7
$\vec{\theta}_{B_L} = (.125, .325, .125, .625, 1, .625, .125, .125, .125)$	.926	332.9
$\vec{\theta}_{B_l} = (.125, .325, .625, 1, .625, .125, .125, .125, .125)$	.847	389.9
Unimodal with Background [(2)]		
$\vec{\theta}_{UB_L} = (.225, .225, .225, .625, 1, .625, .225, .225, .225)$	.927	335.0
$\vec{\theta}_{UB_l} = (.225, .225, .625, 1, .625, .225, .225, .225, .225)$	.870	388.5
$\vec{\theta}_{UB_L} = (.325, .325, .325, .625, 1, .625, .325, .325, .325)$	.924	334.6
$\vec{\theta}_{UB_l} = (.325, .325, .625, 1, .625, .325, .325, .325, .325)$	.792	389.4
Depressed Maximum [(3)]		
$\vec{\theta}_{D_L} = (.125, .125, .125, .625, .9, .625, .125, .125, .125)$	.805	359.7
$\vec{\theta}_{D_l} = (.125, .125, .625, .9, .625, .125, .125, .125, .125)$	.842	387.6

Notes: (1) Power exceeds design power for these cases. (2) Power is unaffected when the effect is on the limb. There is a decrease in power for an effect on the leaf as it becomes more difficult to identify the promising region. (3) The decrease in power for an effect on a limb dose is in keeping with that of a standard design (see discussion). The performance for an effect on the leaf dose is less impacted.

8.2.3 Discussion and Conclusions

This study demonstrates the Limb-Leaf Design in a specific situation where the assumption of a locatable effect in a nonmonotonic dose response curve is satisfied. A significant benefit is seen over a standard seamless design of the TSE-type. In the chosen example with the need to identify a dose having 70% or more of the target effect, the savings in risk adjusted expected sample size over the standard design reaches 25%. The Limb-Leaf Design allocates resources on an as needed basis rather than committing early to a thorough search.

Given that prior information can be imprecise, it is necessary to study the robustness of the design under deviations from the assumed dose response curve and/or misspecifications of the parameters. The results are reassuring but not perfect. The standard cases of Step Function, Monotone, and Plateau alternatives perform well, with greater power than designed. Degradation of power is seen in the Bimodal case and in the case we labeled "Unimodal with a Background." This degradation is only noticeable when the target effect is on a leaf dose. The cause is misdirection of the design to explore the wrong area of the dose response curve as it grows harder to identify the promising region based on the first stage limb performance. Degradation of the power also occurs when the target effect is depressed. This decline is either similar to that seen in the standard design, for a target effect on a limb, or somewhat less, for a target effect on a leaf dose.

As seen in the Bimodal and Unimodal with Background cases, as δ_1 grows in relation to δ_2, it becomes harder to resolve a region of promising effect in the first stage and the performance begins to suffer. If such an alternative is anticipated by design, a large sample size could be required that could negate the benefits of the limb-leaf strategy. In such a situation, one could "front load" the Limb-Leaf Design by removing the constraint that $n_1 = n_{2L}$. For example, with anticipated alternative $\vec{\theta} = (.325, .325, .325, .325, .325, .325, .625, .625, 1)$, having $\delta_1 = .325$ and $\delta_2 = .625$, leaf power of 90% and limb power exceeding 90% can be achieved with $n_1 = 68$, $n_{2L} = 42$, $n_{2l} = 85$, and a search strategy using $\vec{a} = (.2, .75, 1)$; $E_\pi(N)$ is then 353.5. The risk adjusted expected sample size for the optimized standard design in this situation

is 368.2. On the other hand, "back loading" the Limb-Leaf Design is possible when the assumed alternative would make second stage selection more difficult. Under an anticipated Depressed Maximum alternative, with $\vec{\theta} = (.125, .125, .125, .625, .9, .625, .125, .125, .125)$, the choice $n_1 = 60, n_{2L} = 80$, and $n_{2l} = 160$, and the search strategy using $\vec{a} = (.2, .656, .9)$ results in $E_\pi(N) = 403.5$ and meets the desired power conditions. The risk adjusted expected sample size of the optimized standard design for this situation exceeds 500. Other simulations show that if one can allow a difference in power between the cases of an effect on a limb and an effect on a leaf of up to 5 %, a smaller risk adjusted expected sample size can be achieved than when the power must be more evenly balanced.

We conclude that the benefits of Limb-Leaf Designs make them attractive for situations where their assumptions are acceptable. We suggest that adaptive exploration through a limb-leaf strategy is one answer among many to calls for innovation in clinical trial design by the FDA [44] and by specialists (such as Schoenfeld and Cudkowicz, [36]) for a diversity of designs to serve their needs in clinical research.

Chapter 9

Future Work

Further work on Limb-Leaf Designs could take several directions. One is to explore the best use of Limb-Leaf Designs within a drug development program. Enhancements to the design, such as group sequential stopping boundaries and the ability to switch endpoints, could be incorporated by applying known techniques. Statistically, the properties of point and interval estimates derived from Limb-Leaf Designs could be investigated by known methods. We would like to give a brief discussion of these directions.

The needs of a Limb-Leaf Design may call for changes in the standard drug development plan. For instance, Phase I trials could be modified to collect preliminary data on efficacy endpoints and to assess the assumptions of the Limb-Leaf Design. A comprehensive literature review, the use of animal studies, pharmacodynamic models, and the meta-analysis of existing evidence on the drug of interest as well as related drugs could all contribute to the study design process for a Limb-Leaf Design. The composition and operation of the team that designs and conducts the study could be modified with a greater role and a more integrated role for the statistician. This follows the recommendations by Rockhold [32].

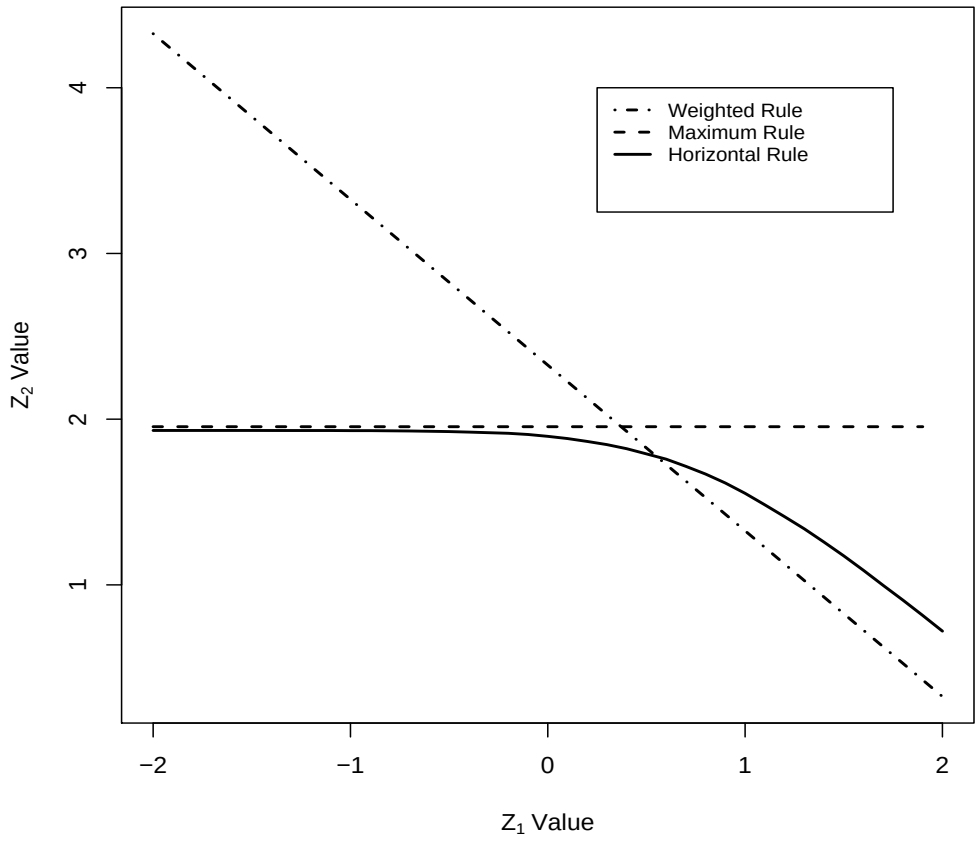
The advantages of group sequential boundaries to stop early for efficacy or futility are described by many authors, including O'Brien and Fleming [31]. In our case, the advantages may be somewhat less because the criterion of distinguishing the desired effect from adjacent promising effects may be more stringent than that of confirming the effect on the selected

dose. Once the sample size is sufficient to allow correct selection with high probability at the second interim analysis, very little additional sample size may be required for high probability to confirm the effect of the selected dose. If needed, group sequential boundaries could be constructed by one of two methods. The first would be, for each the hypothesis H_D (associated with the set of doses D), to generalize the cutoff c_D to a sequence of $c_D^1 \dots c_D^I$ according to an α spending function $\alpha(t) : 0 \leq t \leq 1$. Adaptations could preserve the conditional type 1 error rate at each interim analysis rather than the conditional type 1 error rate across all future analyses in order to maintain a correspondence with the original α spending function. A second method would consider the first two (selection) stages of a Limb-Leaf Design as their own study and to extract a multiplicity adjusted P-value for the effect of the selected treatment. Further stages could stand independently and be combined with the previous evidence using a prespecified group sequential combination rule as described by Müler and Schäfer [29].

Selection at the first and second interim analyses may be based on outcomes other than the primary endpoint. The resulting multiple testing procedure is conservative. An improvement could be made by taking account of the relationship between endpoints. The correlation between score statistics for different endpoints can be estimated and the use of the bivariate distribution can lead to a less conservative procedure. This was done by Stallard and Todd [39] and their method could be applied to a Limb-Leaf Design with a change of endpoint.

Estimation is important for regulatory and medical needs. An established estimation method in selection studies uses bias adjusted estimates. The performance of these estimates and the associated confidence regions in selection studies is studied by Stallard and Todd in [40]. Similar results are likely in the case of Limb-Leaf Designs. Another method for interval construction, the repeated confidence interval method discussed by Brannath, Posch, and Bauer [7], may also be relevant.

FIGURE 1: Comparison of Conditional Error Functions.



Note: The Conditional Error Function of a combination test represents the boundary between the acceptance and rejection regions and provides insight into the robustness and efficiency of the associated combination rule. The conditional error functions associated with the Maximum, Weighted, and Horizontal rules under adaptation Plan 1 of Section 4.3 are graphed above. For discussion see Section 4.3.

FIGURE 2a: EXAMPLE PROFILES WITH LOCATABLE EFFECTS.

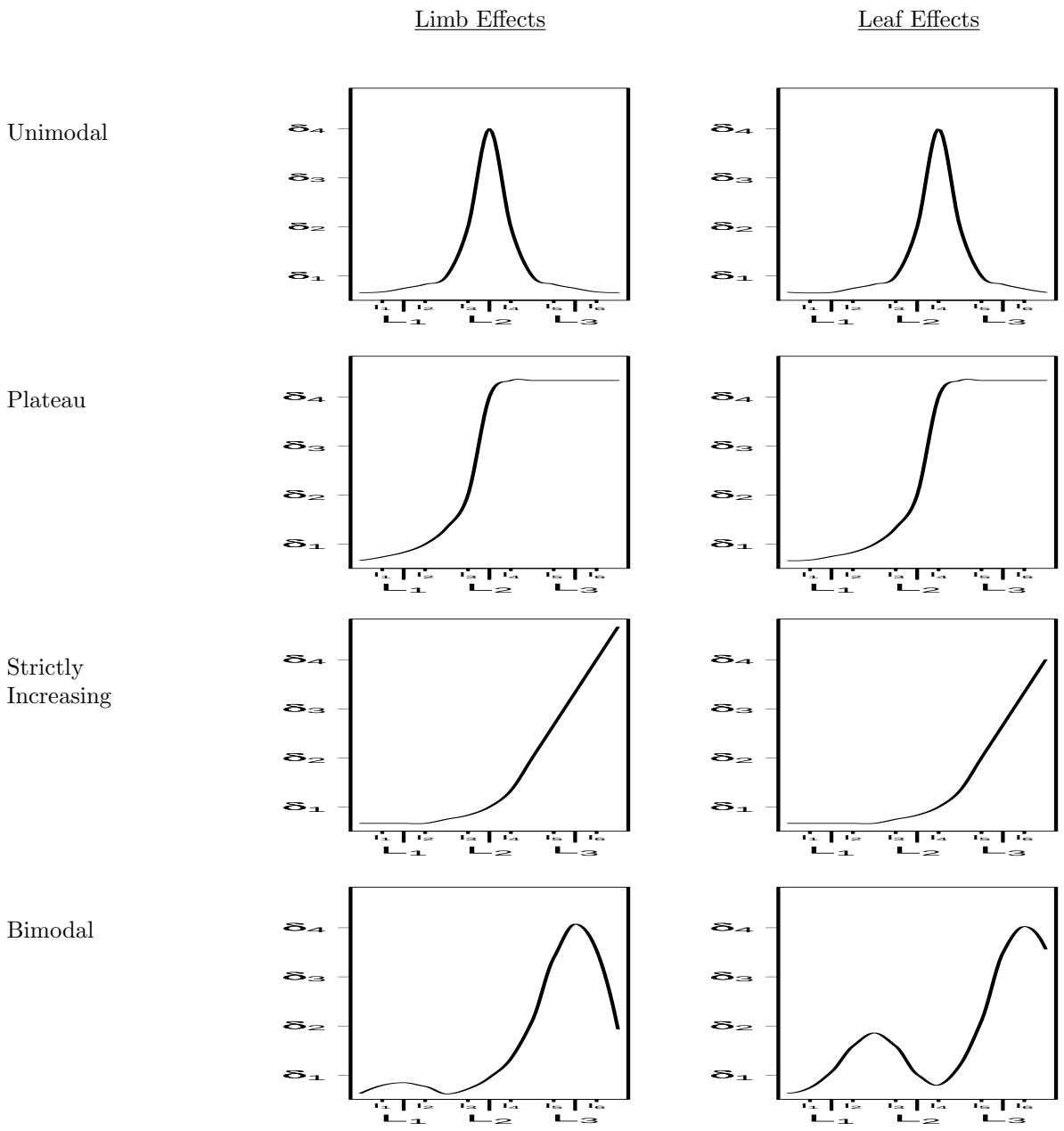
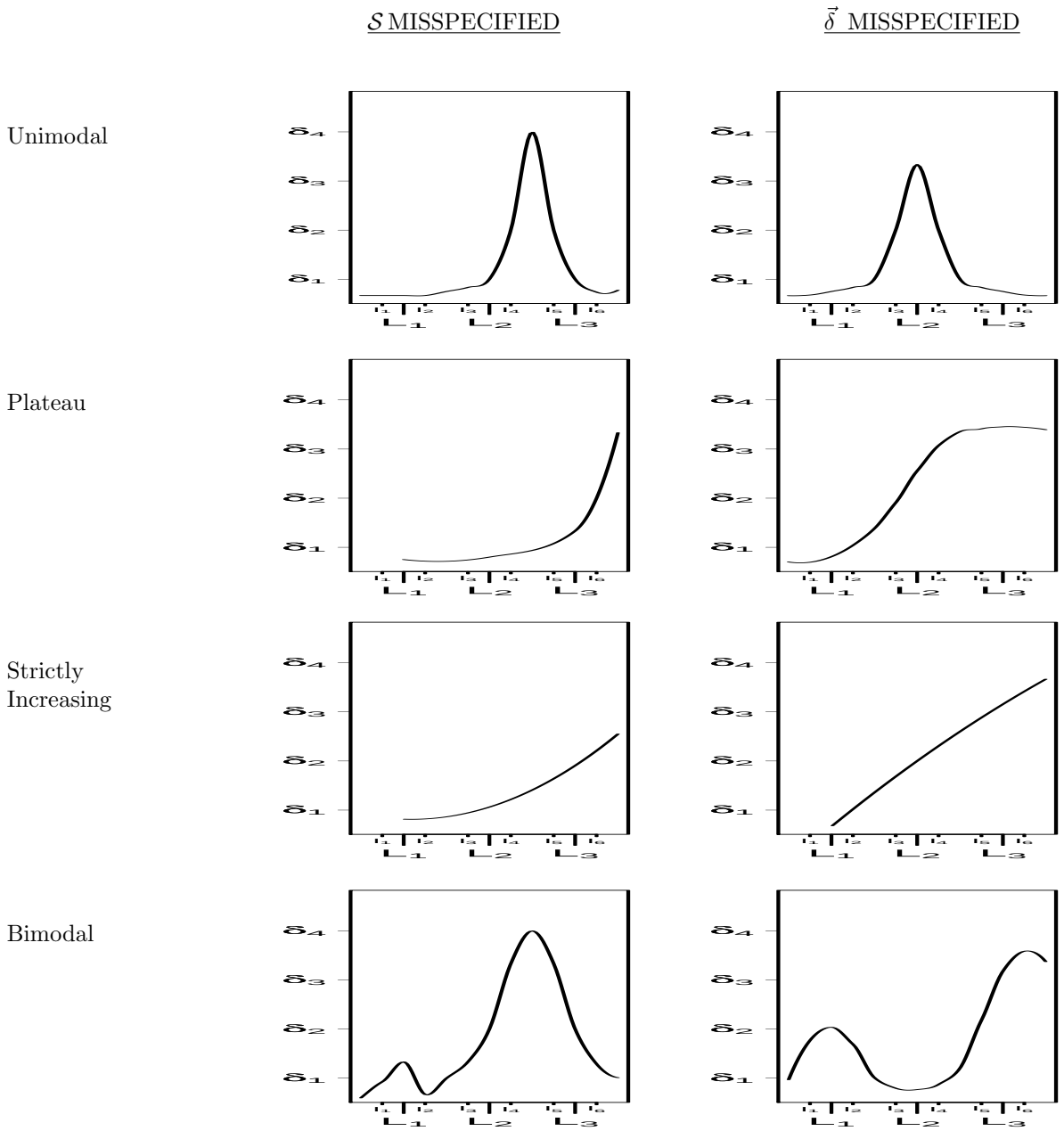


FIGURE 2b: VIOLATIONS OF LOCATABILITY. MISSPECIFICATIONS OF $\mathcal{S}$ AND $\vec{\delta}$.



Bibliography

[1] Anderson, K., and Liu, Q. (2004), "Optimal Adaptive Versus Optimal Group Sequential Designs," *Biopharmaceutical Applied Statistics Symposium XI,* Savannah, Georgia, Available at: http://bass.georgiasouthern.edu, Accessed September 29, 2011.

[2] Bauer, P., and Keiser, M.(1999), "Combining Different Phases in the Development of Medical Treatments Within a Single Trial," *Statistics in Medicine*, 18, 1833-1848.

[3] Bauer, P., and Köhne, K. (1994), "Evaluation of Experiments with Adaptive Interim Analyses," *Biometrics* 50, 1029-1041.

[4] Beal, M.F., Lang, A. E., Ludolph, A.C. (2005), *Neurodegenerative Diseases: Neurobiology, Pathogenesis and Therapeutics,* Cambridge: Cambridge University Press.

[5] Bechhofer, R. E., Kiefer, J., and Sobel, M. (1968), *Sequential Identification and Ranking Problems*, Chicago: University of Chicago Press.

[6] Bensimon, G., Lacomblez, L., Meininger, V., and the ALS/Riluzole Study Group (1994), "A Controlled Trial of Riluzole in Amyotrophic Lateral Sclerosis," *New England Journal of Medicine*, 330, 585-591.

[7] Brannath, W., Posch, M., and Bauer, P. (2002), "Recursive Combination Tests," *Journal of the American Statistical Association*, 97, 236-244.

[8] Bretz, F., Schmidli, H., König, F., Racine, A., and Maurer, W. (2006), "Confirmatory Seamless Phase II/III Clinical Trials with Hypothesis Selection at Interim: General Concepts," *Biometrical Journal*, 48, 623-634.

[9] Follman, D.A., Proschan, M.A., Geller, N.L. (1994), "Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials," *Biometrics*, 50, 325-336.

[10] Goutis, C., Casella, G., Wells, M. (1996), "Assessing Evidence in Multiple Hypotheses," *Journal of the American Statistical Association*, 1996, 1268-1277.

[11] Hölm, S. (1979), "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Satistics*, 6, 65-70.

[12] Hubert, J.P., Delumeau J.C., Glowinski, J., Prémont, J., Doble, A. (1994), "Antagonism by Riluzole of Entry of Calcium Evoked by NMDA and Veratridine in Rate Cultured Granule Cells: Evidence for a Dual Mechanism of Action," *British Journal of Pharmacology*, 113, 261-267.

[13] Inoue, L. Y. T., Thall, P. F., Berry, D.A. (2002), "Seamlessly Expanding a Randomized Phase II Trial to Phase III", *Biometrics*, 58, 823-831.

[14] Jennison, C. (2007), "Introduction to Adaptive Clinical Trial Designs," Presentation at Föreningen för Medicinsk Statistik, Göteberg, Sweden, Available at: http://people.bath.ac.uk/mascj/talks_2007/cj_fms_part2.pdf, Accessed September, 29, 2011.

[15] Jennison, C., and Turnbull, B. (2007), "Adaptive Seamless Designs: Selection and Prospective Testing of Hypotheses," *Journal of Biopharmaceutical Statistics*, 17, 1135-1161.

[16] ———— (2000), *Group Sequential Methods with Applications to Clinical Trials*, Boca Raton: Chapman and Hall/CRC.

[17] ———— (2005), "Meta-Analysis and Adaptive Group Sequential Designs in the Clinical Development Process," *Journal of Biopharmaceutical Statistics*, 15, 537-558.

[18] ———— (2006), "Confirmatory Seamless Phase II/III Clinical Trials with Hypothesis Selection at Interim: Opportunities and Limitations," *Biometrical Journal*, 48, 650-655.

[19] Kaufmann, P., Levy, G., Thompson, J. L. P., DelBene, M. L., Battista, V., Gordon, P. H., Rowland, L. P., Levin, B., and Mitsumoto, H. (2005), "The ALSFRSr Predicts Survival Time in an ALS Clinic Population," *Neurology*, 64, 38-43.

[20] Kelly, P. J., Stallard, N., and Todd, S. (2005), "An Adaptive Group Sequential Design for Phase II/III Clinical Trials that Select a Single Treatment from Several," *Journal of Biopharmaceutical Statistics*, 15, 641-658.

[21] Lehmann, E., and Romano, J. (2005), *Testing Statistical Hypotheses* (3rd ed.), New York: Springer.

[22] Levin, B. (2005), "The Utility of Futility (Editorial)," *Stroke*, 36, 2231-2232.

[23] Levin, B., Thompson, J.L.P., Levy, G., Mitsumoto, H., and Kaufmann, P. (2006), "Pentoxifylline in ALS: A Double-Blind, Randomized, Multicenter, Placebo-Controlled Trial (Letter)," *Neurology*, 66, 1786-1787.

[24] Levy, G., Kaufmann, P., Buchsbaum, R., Montes, J., Barsdorf, A., Arbing, R., Battista, V., Zhou, X., Mitsumoto, H., Levin, B., Thompson, J.L.P. (2006), "A Two-Stage Design for a Phase II Clinical Trial of Coenzyme Q10 in ALS," *Neurology*, 66, 660-663.

[25] Liu, Q., and Pledger, G. (2005), "Phase 2 and 3 Combination Designs to Accelerate Drug Development," *Journal of the American Statistical Association*, 100, 493-502.

[26] Liu, Q., Proschan, M. A., and Pledger, G. W. (2002), "A Unified Theory of Two-Stage Adaptive Designs," *Journal of the American Statistical Association*, 97, 1034-1041.

[27] Ludolph, A. C., Jesse, S. (2009), "Review: Evidence-Based Drug Treatment in Amyotrophic Lateral Sclerosis and Upcoming Clinical Trials," *Therapeutic Advances in Neurological Disorders*, 2, 319-326.

[28] Maca, J., Bhattacharya, S., Dragalin, V., Gallo, P., and Krams, M. (2006), "Adaptive Seamless Phase II/II Designs– Background, Operational Aspects and Examples," *Drug Information Journal*, 40, 463-475.

[29] Müller, H., H., and Schäfer, H. (2001), "Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches," *Biometrics*, 57, 886-891.

[30] Noh, K. M., Hwang, J. Y., Shin, H. C., and Koh, J. Y. (2000), "A Novel Neuroprotective Mechanism of Riluzole: Direct Inhibition of Protein Kinase C," *N*eurobiological Disorders, 7, 375-383.

[31] O'Brien, P., and Fleming, T. (1979), "A Multiple Test Procedure for Clinical Trials," *Biometrics*, 25, 549-556.

[32] Rockhold, F. (2000), "Strategic Use of Statistical Thinking in Drug Development," *Statistics in Medicine*, 19,3211-3217.

[33] Schaid, D. J., Wieand, S., and Therneau, T. M. (1990), "Optimal Two-Stage Screening Designs for Survival Comparisons," *Biometrika*, 77, 659-663.

[34] Scharfstein, D. O., Tsiatis A. A., and Robins, J. M. (1987), "Semiparametric Efficiency and Its Implications on the Design and Analysis of Group-Sequential Studies," *J*ournal of the American Statistical Association, 92, 1342-1350.

[35] Schmidli, H., Bretz, F., Racine, A., and Maurer, W. (2006), "Confirmatory Seamless Phase II/III Clinical Trials with Hypothesis Selection at Interim: Practical Considerations," *Biometrical Journal*, 48, 635-643.

[36] Schoenfeld, D., and Cudkowicz. M. (2001), "Design of Phase II ALS Clinical Trials," *Amyotrophic Lateral Sclerosis*, 9, 16-23.

[37] Stallard, N., and Friede, T. (2008), "A Group Sequential Design for Clinical Trials with Treatment Selection," *Statistics in Medicine*, 27, 6209-6227.

[38] Stallard N, and Todd S. (2003), Sequential Designs for Phase III Clinical Trials Incorporating Treatment Selection," *Statistics in Medicine*, 22, 689-703.

[39] ———— (2005), "A New Clinical Trial Design Combining Phases II and III: Sequential Designs with Treatment Selection and a Change of Endpoint," *Drug Information Journal*, 39, 109-118.

[40] ———— (2005), "Point Estimates and Confidence Regions for Sequential Trials Involving Selection," *Journal of Statistical Planning and Inference*, 135, 402-419.

[41] Tamhane, A., Hochberg, Y., and Dunnett, C. (1996), "Multiple Test Procedures for Dose Finding," *Biometrics*, 52, 21-37.

[42] Thall, P. F., Simon R., and Ellenberg S. S. (1988), "Two-Stage Selection and Testing Designs for Comparative Clinical Trials," *Biometrika*, 75, 303-310.

[43] Tsiatis, A., and Mehta C. (2003), "On the Inefficiency of the Adaptive Design for Monitoring Clinical Trials," *Biometrika*, 90, 367-378.

[44] U.S. Department of Health and Human Services (2010), *Guidance for Industry: Adaptive Design Clinical Trials for Drugs and Biologics*, Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation /Guidances/ucm201790.pdf, Accessed September 2011.

[45] Whitehead, J. (1997), *The Design and Analysis of Sequential Clinical Trials* (rev. 2nd ed.), Chichester: Wiley.