Contemporary Clinical Trials 64 (2018) 210–218

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Limb-Leaf designs for adaptive exploration of the dose-response curve

John Spivack[a,*], Bin Cheng[b], Bruce Levin[b]

[a] Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA
[b] Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York 10032, USA

ARTICLE INFO

Keywords:
Adaptive design
Dose-response curve
Nonmonotonicity
Dose addition
Closed testing principle

ABSTRACT

We propose a two-stage strategy, called the Limb-Leaf method, to explore the dose-response curve using dose promotion and addition in the context of adaptive seamless Phase II/III trials. Strong control of the overall type 1 familywise error rate of the proposed method is enforced by the closed testing principle. The design constants are determined to minimize the risk-adjusted expected total sample size while maintaining a target power. In the case of a nonmonotonic dose response curve where more doses are required to adequately explore the curve, substantial savings in sample size are achieved compared with a traditional strategy which offers only selection and promotion from among initial first stage doses.

1. Introduction

The traditional process of drug development generally consists of four 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, i.e., the Phase 2 results are 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 merge phases together and to eliminate the gaps and delays between them. In general, a seamless design combines the objectives of multiple phases of the development process into a single trial. In particular, it is often possible to meet the objectives of Phases 2 and 3 within one less costly, combined study.

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 while maintaining control over the type 1 error rate. A landmark two-stage adaptive seamless design was proposed by Thall et al. [19] (henceforth, the TSE Design). There a first stage is used to select the best (or apparently 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 comparison between control and selected arm. This design in its original formulation applies only to binary outcomes, however the TSE template is easily modified to other outcome distributions, for instance normal outcomes as described by Jennison and Turnbull [7,8]. An important generalization that includes multiple stages and the use of a test based on the score statistic was proposed by Stallard and Todd [16], and Todd and Stallard [20]. This method accommodates a general endpoint, which, for instance, could be normal, binary, or a time to an event.

Another route for the development of adaptive seamless designs has been through the adaptive P-value combination tests used by Bauer and Köhne [2]. 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 familywise error rate for the multiple hypotheses under study. The method of adaptive combination tests is very general and [14] shows how it includes group sequential tests, the two stage TSE design, and further generalizations as special cases.

2. A proposal for adaptive exploration

Experiments can benefit from a more structured exploration strategy when the dose-response curve is not assumed to be monotonic, particularly so in the case where it is still assumed to be unimodal. This

* Corresponding author.
E-mail address: john.spivack@mountsinai.org (J. Spivack).

http://dx.doi.org/10.1016/j.cct.2017.10.002
Received 12 June 2017; Received in revised form 27 September 2017; Accepted 4 October 2017
Available online 06 October 2017
1551-7144/ © 2017 Elsevier Inc. All rights reserved.

possibility is especially plausible when dealing with a combined efficacy/toxicity or benefit/cost endpoint, or when dealing with a combination therapy such that no single ordering of the doses may be possible. We present the following example of a real-world setting in which such an exploration strategy might be applied.

2.1. An example based on the QALS trial

The U.S. Food and Drug Administration has approved only one drug, Riluzole, for the treatment of the devastating neurodegenerative disease Amyotrophic Lateral Sclerosis (ALS). Benefits in patient function and survival are small and safety and tolerability of Riluzole especially in regard to liver toxicity are major considerations. There is hope that the progression of ALS may one day be slowed or stopped by new medications or combinations of drugs.

Methods of action of Riluzole or other possible therapies are not fully elucidated; they are complex and may involve multiple mechanisms as reported, for instance, by Hubert et al. [5], Noh et al. [15], and Beal et al. [3]. In such settings the expectation that an investigational drug would have a monotonic dose response curve with increasing patient benefit up to a well-defined maximum tolerated dose is especially problematic. Indeed an incorrect assumption of monotonicity in the dose response is a favored explanation for the failure of a recent major ALS study as reported by Ludolph and Jesse [12].

The QALS trial published by Kaufmann et al. [9] was undertaken to investigate high doses of Coenzyme Q10 as a possible therapy for ALS. It was not assumed that the dose response would be monotonic and toxicity was carefully monitored. The study was designed in two stages, a selection stage followed by a futility test, also known as a non-superiority test (see, e.g., [10]). Two doses of CoQ10 (1800 and 2700 mg per day) together with a placebo arm began the first stage of the study. After an interim analysis the apparently better performing dose was selected to continue and additional recruitment to that arm and the placebo arm took place in the second stage. The final test statistics involved data pooled over both stages under appropriate control for selection bias and type 1 error.

The outcome of the first stage of the trial was that the higher dose did better than the lower dose on the outcome measure (ALSFRSr, the ALS functional rating scale, revised) and had high tolerability. After continuation into the second stage the test statistic associated with this higher dose was nominally sufficient to avoid a declaration of futility. The investigators nevertheless did not consider the evidence promising enough to give it full endorsement. Further details are given in Kauffman et al. [9].

Notwithstanding the importance of certain key differences of the QALS trial, especially its aim to test a futility hypothesis rather than a superiority hypothesis, it is easy to imagine that in this or a similar study, a conventional superiority hypothesis might instead be the goal. In such a case, for example, it might have been considered worthwhile to allow further exploration in the second stage around the apparently better first stage dose. Specifically, had there been an option to add higher doses beyond 2700 mg per day or to explore both above and below this dose level, this freedom and flexibility might have been attractive to investigators. It is perhaps possible that an efficacious dose might have been among those added, and good enough to earn a full endorsement. We would like to make such further options available to investigators by design in especially difficult disease areas like ALS.

2.2. Stagewise adaptive exploration

In exploring the dose-response relationship, particularly a non-monotonic one, it is important to distinguish between d^*, a dose with the maximum possible effect, and $\hat{d}^*$, its estimated value. The corresponding effects of these doses, θ_{d^*} and $\theta_{\hat{d}^*}$, say, could be different in a meaningful way, with $\theta_{\hat{d}^*} < \theta_{d^*}$. Similarly, if d^* denotes a dose with a given desired effect (not necessarily maximum), $|\theta_{\hat{d}^*} - \theta_{d^*}|$ could be undesirably large. On the other hand, at least heuristically, the more closely $\hat{d}^*$ approximates d^* the greater the chance that the study will reject the global null hypothesis, $H_0 : \theta_d \leq 0$ for all doses d, and more importantly, the closer the final recommended dose will be to that which gives patients the desired (or maximum) benefit. These considerations provide strong motivation to better explore the dose-response curve in adaptive seamless designs.

One possibility for better exploration of a non-monotonic dose-response curve in a single-stage selection strategy like the TSE Design would be to include a large number of closely spaced doses in the first stage. However, there are reasons to expect performance to suffer with this approach. First, 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. Second, 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 ethically undesirable. Some of these issues are mentioned by Thall et al. [19], but they play less of a role in the case where there are only a few possible doses to consider with broad spacing between them.

In this paper we introduce a two-stage selection procedure called the "Limb-Leaf" design in which second stage doses not only are promoted from a modest number of first stage candidates but also may be added in response to first stage results. We aim to improve the estimation of d^* and to use resources more efficiently, this being particularly so under the global null hypothesis, where the probability of early stopping should be large. Such promotion and addition decisions can be based on all the available information, including efficacy and toxicity, whether it comes from within the study or from an outside source.

Although the Limb-Leaf design we present here achieves pre-specified performance characteristics in the case of a non-monotonic dose-response curve, it actually comprises a more general approach to structured exploration of response functions.

3. Method

Below we assume without essential loss of generality that the data are normal with variance σ^2 known. Extensions to other data types are mentioned with further detail given in the Appendix.

3.1. The TSE method

The TSE design has two stages; the first stage assigns subjects to all candidate doses plus the control, and the second stage studies only the best performing dose from the first stage against the control. There is an 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 of the selected treatment exceeds a second cutoff value.

A version of the TSE design using normal outcomes is described as follows. Let the test doses in the experiment be denoted as $d_1,\ldots,d_I$, with effects relative to control dose d_0 of $\theta_1,\ldots,\theta_I$. We assume that at either stage, the outcomes at a given dose d_j are independent and identically distributed (i.i.d.) as a normal random variable with mean μ_j and variance σ^2, $j = 0,1,\ldots,I$. The design proceeds in two stages:

Stage 1. Randomize $(I+1)n_1$ patients equally to $d_0,d_1,\ldots,d_I$. Let $T_1 = \max_{1\leq i\leq I} T_{1,i}$, where for each i, $T_{1,i} = (\overline{X}_{1,i} - \overline{X}_{1,0})/\sqrt{2\sigma^2}$, $\overline{X}_{1,0}$ is the sample mean for the control d_0, and $\overline{X}_{1,i}$ is the sample mean for dose d_i, $i = 1,\ldots,I$ at stage 1. If $T_1 > y_1$, then continue by selecting the treatment, d_{i^*} having the greatest observed effect, T_{1,i^*}, into a second stage. If $T_1 \leq y_1$ then stop and accept H_0 of no effect on any dose.

Stage 2. Randomize $2n_2$ additional patients equally to d_{i^*} and d_0. Let

$$T_2 = \frac{n_1}{n_1+n_2}\frac{\overline{X}_{1,i^*}-\overline{X}_{1,0}}{\sqrt{2\sigma^2}} + \frac{n_2}{n_1+n_2}\frac{\overline{X}_{2,i^*}-\overline{X}_{2,0}}{\sqrt{2\sigma^2}}.$$

If $T_2 > y_2$ then reject H_{0,i^*}: $\theta_{i^*} \le 0$ and conclude that $\theta_{i^*} > 0$; if $T_2 \le y_2$ then do not reject H_{0,i^*}.

Notably, the design can stop early for futility but allows a new treatment to be judged superior to the control only after a second stage, based upon data from $2(n_1 + n_2)$ patients. The design constants n_1,n_2,y_1, and y_2 are determined by minimizing the risk-adjusted expected total sample size and maintaining a target power under a least favorable dose-response scenario. More details will be given in Section 5.

3.2. The Limb-Leaf design

3.2.1. Limb-Leaf structure and locatable effect

Let $\mathscr{D} = \{d_1, ..., d_I\}$ be a prespecified collection of test doses to be investigated as in a TSE design. A Limb-Leaf structure for $\mathscr{D}$ is a relabelling of the doses in $\mathscr{D}$ as

$$\mathscr{D} = \{L_1, l_{1,1}, ..., l_{1,m_1}; ...; L_K, l_{K,1}, ..., l_{K,m_K}\},$$

where for each "limb" dose L_k, $k = 1,\ldots,K$, there is an associated neighborhood also including m_k leaf doses, $l_{k,1}, ..., l_{k,m_k}$. Let $\Theta_{\mathscr{D}}$, called the dose-response configuration, be the collection of effects associated with Limb-Leaf system $\mathscr{D}$. That is,

$$\Theta_{\mathscr{D}} = \{\theta_{L_1}, \theta_{l_{1,1}}, ..., \theta_{l_{1,m_1}}; ...; \theta_{L_K}, \theta_{l_{K,1}}, ..., \theta_{l_{K,m_K}}\}.$$

Generally, the effects $\theta_{l_{k,1}}, ..., \theta_{l_{k,m_k}}$ on the respective leaves $l_{k,1}, ..., l_{k,m_k}$ would be assumed to be somewhat similar but not identical to that of their associated limb dose L_k.

In this paper, we assume that the dose-response configuration under investigation has a "locatable" effect, whose definition is as follows.

Definition 1. *A dose-response configuration$\Theta_{\mathscr{D}}$is defined to have a locatable effect with respect to the Limb-Leaf System$\mathscr{D} = \{L_1, l_{1,1}, ..., l_{1,m_1}; ...; L_K, l_{K,1}, ..., l_{K,m_K}\}$and the response levels$\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$, with $\delta_1 < \delta_2 < \delta_3$, where $\delta_3$is considered to be the smallest desired level of effect, if the following two conditions hold:*

1. *The effects on each limb,θ_{L_k}, for $k = 1,\ldots,K$, are either less than or equal to δ_1, or greater than or equal to δ_2, with at least one k such that$\theta_{L_k} \ge \delta_2$.*
2. *For any limb $L_k$with$\theta_{L_k} \ge \delta_2$, each effect in its neighborhood is either greater than or equal to $\delta_3$or less than or equal to $\delta_2$with at least one effect greater than or equal to δ_3.*

Intuitively, a dose-response configuration $\Theta_{\mathscr{D}}$ with a locatable effect means that the shape of the dose response curve should permit exploration in stages, the first on a coarse level and the second for finer level adjustments, in order to successfully identify a dose of desired effect level δ_3 or better.

Proposition 1. *Any given dose-response configuration$\Theta_{\mathscr{D}}$has a locatable effect with respect to the Limb-Leaf System$\mathscr{D} = \{L_1, l_{1,1}, ..., l_{1,m_1}; ...; L_K, l_{K,1}, ..., l_{K,m_K}\}$and some vector of response levels $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$, with $\delta_1 < \delta_2 < \delta_3$.*

The proof of this result is given in the Appendix. It follows that for prespecified $\mathscr{D}$, $\Theta_{\mathscr{D}}$, and $\boldsymbol{\delta}$, failure to meet the definition of locatability may be considered as a misspecification of $\boldsymbol{\delta}$, which is a useful perspective in examining the performance of the Limb-Leaf approach under violations of assumptions. This does not, however, mean that selection decisions for $\mathscr{D}$ and $\boldsymbol{\delta}$ are unimportant; a selection of $\mathscr{D}$ and $\boldsymbol{\delta}$ that are not only formally correct but appropriate to the underlying dose-response relationship is key for a design to have good performance characteristics. Guidance on how $\mathscr{D}$ and $\boldsymbol{\delta}$ may be chosen and sources of information for this decision are offered in Sections 5 and 6.

3.2.2. The plan of a Limb-Leaf design

A Limb-Leaf method proceeds as follows:

Step 1. Prespecify a vector $\boldsymbol{c} = (c_1,c_2)$, $c_1 < c_2$, representing different levels of a test statistic, and a vector of weights (w_1,w_2) such that $w_1,w_2 \ge 0$, with $w_1^2 + w_2^2 = 1$.

Step 2. Randomize n_{1L} patients to each limb and control. Let $\hat{\theta}_{1,L_k} = \overline{X}_{1,L_k} - \overline{X}_{1,L_0}$ denote the first stage estimate and $Z_{1,L_k} = \hat{\theta}_{1,L_k} \Big/ \sqrt{\frac{2\sigma^2}{n_{1L}}}$ denote the first stage test statistic of the effect of dose L_k, $k = 1,\ldots,K$. The limb with the greatest first stage statistic, denoted by L_{k^*}, will have an estimate denoted by $\hat{\theta}_{1,L_{k^*}}$ and a test statistic denoted by $Z_{1,L_{k^*}}$. Note that k^* is a random variable.

Step 3. There are 3 possibilities.

(i) If $\hat{\theta}_{1,L_{k^*}} \le c_1$, then the study stops for futility.

(ii) If $c_1 < \hat{\theta}_{1,L_{k^*}} \le c_2$ the experiment continues to Stage 2 with L_{k^*}, its leaves $l_{k^*,1}, ..., l_{k^*,m_{k^*}}$, and the control dose L_0. Randomize n_{2L} patients to each of L_{k^*} and control L_0, and n_{2l} patients to each of $l_{k^*,1}, ..., l_{k^*,m_{k^*}}$.

Let $\hat{\theta}_{2,L_{k^*}} = \overline{X}_{2,L_{k^*}} - \overline{X}_{2,L_0}$, $Z_{2,L_{k^*}} = \frac{\hat{\theta}_{2,L_{k^*}}}{\sqrt{\frac{2\sigma^2}{n_{2L}}}}$; $\hat{\theta}_{2,l_{k^*,l}} = \overline{X}_{2,l_{k^*,l}} - \overline{X}_{2,L_0}$ and $Z_{2,l_{k^*,l}} = \frac{\hat{\theta}_{2,l_{k^*,l}}}{\sqrt{\frac{\sigma^2}{n_{2l}} + \frac{\sigma^2}{n_{2L}}}}$ $l = 1, ..., m_{k^*}$; and $\hat{\theta}_{L_{k^*pooled}} = \frac{n_{1L}\ \hat{\theta}_{1,L_{k^*}} + n_{2L}\ \hat{\theta}_{2,L_{k^*}}}{n_{1L}+n_{2L}}$.

If $\hat{\theta}_{L_{k^*pooled}} = \max\{\hat{\theta}_{L_{k^*pooled}}, \hat{\theta}_{2,l_{k^*,l}}, l = 1, ..., m_{k^*}\}$, then we reject $H_{0,L_{k^*}}$: $\theta_{L_{k^*}} \le 0$ and estimate $\hat{d}^*$ as L_{k^*} if and only if

$$w_1\Phi^{-1}\{F_K(Z_{1,L_{k^*}})\} + w_2\Phi^{-1}\Big(\min_{0\le k\le m_{k*}}\{G_{\mathrm{Limb+k}}[\max(Z_{2,L_{k^*}}, Z_{2,l_{k^*,(k)}})]\}\Big) > z_\alpha$$

Here $Z_{2,l_{k^*,(k)}}$ is the kth smallest order statistic of $\{Z_{2,l_{k^*,1}}, ..., Z_{2,l_{k^*,m_{k^*}}}\}$ and it is understood that for $k = 0$ the maximum in the second term reduces to that over $Z_{2,L_{k^*}}$ alone. The function F_K is the cdf of $\max\{Z_{1,L_1}, ..., Z_{1,L_K}\}$ under H_0: $\theta_{L_1} = \cdots = \theta_{L_K} = 0$; G_{Limb+k} is the cdf of $\max\{Z_{2,L_{k^*}}, Z_{2,l_{k^*,1}}, ..., Z_{2,l_{k^*,k}}\}$ under H_0: $\theta_{L_{k^*}} = \theta_{l_{k^*,1}} = \cdots = \theta_{l_{k^*,k}} = 0$, conditional on the present adaptation decision (ii); and z_α is the $(1-\alpha)\times 100$-th percentile of the standard normal distribution.

If, for some l^*, $\hat{\theta}_{2,l_{k^*,l^*}} = \max\{\hat{\theta}_{L_{k^*pooled}}, \hat{\theta}_{2,l_{k^*,l}}, l = 1, ..., m_{k^*}\}$, then we reject $H_{0,l_{k^*,l^*}}$: $\theta_{l_{k^*,l^*}} \le 0$ and claim $\hat{d}^* = l_{k^*,l^*}$ if and only if

(a) $w_1\Phi^{-1}\{F_K(Z_{1,L_{k^*}})\} + w_2\Phi^{-1}\{G_{Limb+m_{k^*}}[\max(Z_{2,L_{k^*}}, Z_{2,l_{k^*,l^*}})]\} > z_\alpha$,

(b) $w_1\Phi^{-1}\{\min_{1\le k\le K-1}[F_k(Z_{1,L_{(k)}})]\} + w_2\Phi^{-1}\{G_{m_{k^*}}(Z_{2,l_{k^*,l^*}})\} > z_\alpha$, and

(c) $\Phi^{-1}\{G_{m_{k^*}}(Z_{2,l_{k^*,l^*}})\} > z_\alpha$,

where $Z_{1,L_{(k)}}$ is the kth smallest order statistic of $\{Z_{1,L_1}, ..., Z_{1,L_K}\}$ and $G_{m_{k^*}}$ is the cdf of $\max\{Z_{2,l_{k^*,1}}, ..., Z_{2,l_{k^*,m_{k^*}}}\}$ under H_0: $\theta_{L_{k^*}} = \theta_{l_{k^*,1}} = \cdots = \theta_{l_{k^*,m_{k^*}}} = 0$, conditional on the present adaptation decision (ii).

(iii) If $\hat{\theta}_{1,L_{k^*}} > c_2$ then L_{k^*} and the control dose only should proceed into the second stage. Randomize n_{2L}' patients to L_{k^*} and to the control. Let $Z_{2,L_{k^*}} = \frac{\overline{X}_{2,L_{k^*}} - \overline{X}_{2,L_0}}{\sqrt{\frac{2\sigma^2}{n'_{2L}}}}$. We reject $H_{0,L_{k^*}}$: $\theta_{L_{k^*}} \le 0$ and claim $\hat{d}^* = L_{k^*}$ if and only if

$$w_1\Phi^{-1}\{F_K(Z_{1,L_{k^*}})\} + w_2 Z_{2,L_{k^*}} > z_\alpha.$$

We note that the selection of the final second stage dose using

$\hat{\theta}_{L_{k^*pooled}}$ is consistent with the seamless quality of the design: information across both stages is used for this decision. Hypothesis testing, however, may not be done directly in terms of $\hat{\theta}_{L_{k^*pooled}}$. Stagewise combination rules, as described in the next section, are needed to make valid inference.

The Limb-Leaf design is not limited to normal outcomes with known variance. Extensions to practical cases such as those of a large sample with consistently estimated nuisance parameters, and normally distributed data with unknown variance are presented in the Appendix. The design extends similarly to cases such as that of testing a location-shift hypothesis for continuous distributions using rank sum tests, and that of testing a difference in proportions for binary data using exact tests. These will be presented in later work.

3.2.3. Closed testing procedure and justification of the Limb-Leaf method

The search strategy outlined above entails selection over multiple dose levels, which calls for control of the familywise error rate (FWER) with respect to the collection of hypotheses $\{H_{0,d}\colon \theta_d \le 0,\ d \in \mathscr{D}\}$, where FWER is the probability of rejecting any true null hypothesis. Control of the FWER for a given procedure means that FWER $\le \alpha$ where the significance level α is prespecified. Furthermore, as argued by Tamhane et al. [18] among others, the appropriate form of familywise type 1 error control should be strong control, such that the FWER $\le \alpha$ regardless of which hypotheses or how many hypotheses from the collection $\{H_{0,d}\colon \theta_d \le 0,\ d \in \mathscr{D}\}$ are true. In other words we require that given a Limb-Leaf System $\mathscr{D}$, $\mathrm{FWER}(\Theta_{\mathscr{D}}) \le \alpha$ regardless of the shape of the underlying dose-response configuration $\Theta_{\mathscr{D}}$. We emphasize that in a design that allows addition of doses as well as their promotion, control of the FWER only under the global null (weak control) does not imply the needed strong FWER control and is insufficient.

The closed testing procedure of Marcus et al. [13] guarantees strong control of the FWER with respect to a prespecified family of hypotheses which is closed under intersection. For a given set of hypotheses $\{H_{0,j}: j = 1,\ldots,J\}$ the construction is as follows: For each subset S of $\{1,\ldots,J\}$ define the intersection hypothesis $H_{0,S} = \cap_{j\in S} H_{0,j}$ with corresponding level α test ϕ_S assumed to exist for each S. Then the closed testing principle requires that any hypothesis $H_{0,j}$ be rejected if and only if $H_{0,S}$ is rejected by ϕ_S for every set S that contains j. The proof of strong control of the FWER is immediate: Let S^* be the set of the indices of all true null hypotheses and assume S^* is non-empty (or else there is nothing to prove). For a familywise error to be committed, H_{0,S^*} must be rejected at level α, which occurs with probability no greater than α.

The Limb-Leaf procedure uses the following tests for hypotheses $H_{0,D}$. If D contains any limb, then reject the null if and only if

$$Z_D = w_1\Phi^{-1}\{F_D(\max_{d\in D}\{Z_{1,d}\})\} + w_2\Phi^{-1}\{G_D(\max_{d\in D}\{Z_{2,d}\})\} \ge z_\alpha.$$

Otherwise, reject the null if and only if

$$Z_D = \Phi^{-1}\{G_D(\max_{d\in D}\{Z_{2,d}\})\} \ge z_\alpha.$$

Here $Z_{i,d}$ is the test statistic for $H_{0,d}: \theta_d \le 0$ based on the ith stage data from dose d and control if dose d appears at Stage i, $i = 1,2$. F_D is the cdf of $\max_{d\in D}\{Z_{1,d}\}$ under $\theta_d = 0$, $d \in D$, and G_D is the cdf of $\max_{d\in D}\{Z_{2,d}\}$ under $\theta_d = 0$, $d \in D$, conditional on the first stage data and thus on adaptation decisions based on them. If no dose $d \in D$ appears in the second stage, we set $Z_D \equiv -\infty$, as their corresponding doses have been deemed irrelevant after review of first stage data. It is emphasized that w_1 and w_2, $w_1^2 + w_2^2 = 1$, must be prespecified at the design stage in order for the test to be valid, and that the second form given is a special case of the first with $w_1 = 0$.

The following result provides a theoretical justification of the Limb-Leaf procedure. The proof is given in the Appendix.

Theorem 1. *For any dose-response configuration $\Theta_{\mathscr{D}}$, the Limb-Leaf method specified in Section 3.2.2 yields strong control of the FWER at level α.*

Further flexibility in the Limb-Leaf design may be allowed. Selection of a dose other than the best performing at either stage does not undermine the validity of the familywise type 1 error rate control but may incur a power penalty that can be partly offset by a more complicated evaluation of all relevant intersection hypotheses. Dose selections can incorporate other information such as toxicity, cost or external data. Interim re-calculation of sample sizes, for instance to increase conditional power, would also be possible without undermining the validity of the trial. Finally, another potentially important type of flexibility would be in the definition of leaf doses; their exact dose levels or other aspects of their formulations could be finalized at the interim analysis without compromising the conditional sub-uniform distribution of their associated P-values and the procedure's validity.

The validity of the Limb-Leaf design under other types of data is addressed by Theorem 2, also proved in the Appendix.

Theorem 2. *The Limb-Leaf design extends to practical cases such as: 1) Normally distributed data with unknown variance σ^2, and 2) Large sample inference with consistently estimated nuisance parameters.*

4. Power and optimization

For a given Limb-Leaf System $\mathscr{D}$ and a vector of effect thresholds $\boldsymbol{\delta}$, we must prespecify the parameters $n_{1L}, n_{2L}, n_{2L}', n_{2l}$, $\boldsymbol{c} = (c_1, c_2)$, and weight w_1. To do this we minimize a risk-adjusted expected sample size under constraints on the power to identify and confirm an effect under specified alternatives.

4.1. Unfavorable configurations

Similar to Thall et al. [19], we choose certain unfavorable configurations (test configurations) in which we want to enforce adequate power. These are instances of locatable effects with respect to $\boldsymbol{\delta}$ that are unfavorable to identification and confirmation of an effect on the correct target. One such configuration, the unfavorable limb effect configuration, Θ_{Limb}, would satisfy the conditions: $\theta_{L_k} = \theta_{l_{k,1}} = \cdots = \theta_{l_{k,m_k}} = \delta_1$ for $k \ne k^*$, $\theta_{L_{k^*}} = \delta_3$, and $\theta_{l_{k^*,1}} = \cdots = \theta_{l_{k^*,m_{k^*}}} = \delta_2$ for some limb dose L_{k^*}. Analogously, we define an unfavorable leaf effect configuration Θ_{Leaf} as satisfying the conditions $\theta_{L_k} = \theta_{l_{k,1}} = \cdots = \theta_{l_{k,m_k}} = \delta_1$ for $k \ne k^*$, $\theta_{L_{k^*}} = \delta_2$, $\theta_{l_{k^*,l^*}} = \delta_3$, and $\theta_{l_{k^*,l}} = \delta_2$ for $l \ne l^*$, for some limb dose L_{k^*} and leaf dose l_{k^*,l^*}.

The intuition of these choices is clear but a full characterization of the configurations that minimize the power would depend on assumptions concerning the construction of stagewise tests, combination rules, and interim dose selection rules. We do not address a full solution in this paper. Thus we call Θ_{Limb} and Θ_{Leaf}" unfavorable" rather than "least favorable" configurations.

4.2. Power and risk adjusted expected sample size

Let the risk-adjusted expected sample size, a form of Bayes' risk, be defined as

$$E_\pi(N) = \pi_0 E_0(N) + \pi_{\Theta_{Limb}} E_{\Theta_{Limb}}(N) + \pi_{\Theta_{Leaf}} E_{\Theta_{Leaf}}(N),$$

where $E_0(N)$, $E_{\Theta_{Limb}}(N)$, and $E_{\Theta_{Leaf}}(N)$ denote the expected total sample sizes under the global null, unfavorable limb, and unfavorable leaf configurations, respectively, and π_0, $\pi_{\Theta_{Limb}}$, and $\pi_{\Theta_{Leaf}}$ denote their associated prior probabilities. Conservative values for early to mid-stage drug development would be $\pi_0 = 0.8$, $\pi_{\Theta_{Limb}} = 0.1$, and $\pi_{\Theta_{Leaf}} = 0.1$, for instance.

The design constants $n_{1L}, n_{2L}, n_{2L}', n_{2l}$, $\boldsymbol{c} = (c_1, c_2)$, and w_1 are selected to minimize $E_\pi(N)$, subject to power constraints

$$P_{\Theta_{Limb}}(\text{Confirm the treatment effect on limb } L_{k^*}) \ge 1 - \beta,$$

and

$$P_{\Theta_{Leaf}}(\text{Confirm the treatment effect on leaf } l_{k^*,l^*}) \ge 1 - \beta.$$

In general a nonlinear optimization procedure is required to find a possibly non-unique optimizer, or an approximate minimum can be found by a grid search over potential parameter values.

5. Simulation studies

We consider two schemas for comparison of a Limb-Leaf approach with a traditional TSE-style design. Both schemas are simple cases that could be used in practice. We compare performance characteristics over ranges of the $\delta_1,\delta_2,\delta_3$ parameters. In addition, robustness of the Limb-Leaf type design is considered in the following subsection. Without loss of generality we assume that the outcome of a patient on dose d is distributed as $N(\theta_d,1)$, and the outcome of a control patient is distributed as $N(0,1)$. All optimizations were conducted using computationally intensive grid-search algorithms. More efficient computational methods under development will be described in Section 6.

5.1. Schema A: A single limb

One of the simplest possible Limb-Leaf schemas (hence forth, "Schema A") is as follows. There are 3 testing doses of interest, organized as: limb L_1 and leaves $l_{1,1},l_{1,2}$. Stage 1 compares L_1 to control L_0 where each arm would have sample size n_{1L}. Depending on the results of this first stage comparison, we might: i) terminate the study, ii) continue the study with further recruitment to L_0 and L_1 of n_{2L}' subjects each in the second stage, or iii) continue with recruitment of n_{2L} subjects each to L_0 and L_1 as well n_{2l} subjects each to leaves $l_{1,1}$ and $l_{1,2}$. In case iii) we will allow selection of the optimal dose by comparison of overall sample means; although the inclusion of the first stage subjects in the sample mean of L_1 is not completely unbiased, in practice this information may improve the probability of correct selection of the best dose. Evaluation of the results would be by the closed testing procedure utilizing pre-specified combination tests as described in Section 3. This could correspond, for instance, to a situation where investigators want the option to introduce interim modifications of their initial test dose L_1 in case first stage results lead them to believe that either an increment or a decrement to the dose level might be necessary to optimize performance.

The corresponding TSE design would include L_0 and $L_1,l_{1,1},l_{1,2}$ as first stage doses, each with sample size n_1. If the first stage result of the best performing test dose $\hat{d}^*$ exceeds the preset threshold y_1, d^* would continue recruitment in the second stage along with control dose L_0 with n_2 subjects per arm. The efficacy of $\hat{d}^*$ relative to L_0 would then be determined by whether the overall sample mean of subjects assigned to $\hat{d}^*$ exceeds the pre-specified y_2.

The design constants to be set for the Limb-Leaf design are then $n_{1L},n_{2L},n_{2L}',n_{2l}$, $\boldsymbol{c} = (c_1,c_2)$, as well as the combination rule weights w_1 and w_2. Their values will be determined to minimize $E_\pi(N)$ defined in Section 4.2 with $\pi_0 = 0.8$, $\pi_{\Theta_{Limb}} = 0.1$, and $\pi_{\Theta_{Leaf}} = 0.1$. Specifically, for given $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$, we let Θ_{Limb} be given by $\{\theta_{L_1} = \delta_3, \theta_{l_{1,1}} = \delta_2, \theta_{l_{1,2}} = \delta_2\}$, and Θ_{Leaf} be given by $\{\theta_{L_1} = \delta_2, \theta_{l_{1,1}} = \delta_2, \theta_{l_{1,2}} = \delta_3\}$. Numerical optimization of the criterion $E_\pi(N)$ subject to contraints of 0.9 power in both Θ_{Limb} and Θ_{Leaf} configurations determines the design constants.

The corresponding TSE-type design will require the design constants n_1,n_2,y_1, and y_2. Since the TSE-type design does not recognize a distinction between limb and leaf doses, both Θ_{Limb} and Θ_{Leaf} may be expressed as Θ_{TSE}, given by $\{\theta_{L_1} = \delta_3, \theta_{l_{1,1}} = \delta_2, \theta_{l_{1,2}} = \delta_2\}$. The previous criterion, $E_\pi(N)$ then reduces to $0.8E_0(N) + 0.2E_{\Theta_{TSE}}(N)$. The power constraints similarly reduce to the single restriction of 0.9 power in the Θ_{TSE} configuration.

Below we present two tables, of optimized parameter values and performance characteristics for the two designs over ranges of $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$ for Schema A (Tables 1 and 2). Where this simple case assigns the same values of $\Theta_{Limb},\Theta_{Leaf}$, and Θ_{TSE} for several values of $\boldsymbol{\delta}$, the optimization is unaffected by which one is chosen and the results are combined.

Table 1
TSE design constants under Schema A.

δ_1	δ_2	δ_3	n_1	n_2	y_1	y_2	$E_\pi(N_{TSE})$
0.0	0.2	1.0	20	5	0.575	0.525	82.7
0.2	0.4	1.0	20	10	0.500	0.500	125.0
0.4	0.6	1.0	40	5	0.625	0.125	162.1
0.6	0.8	1.0	125	5	0.700	0.300	502.0

Table 2
Limb-Leaf design constants under Schema A.

δ_1	δ_2	δ_3	n_{1L}	n_{2L}	n_{2L}'	n_{2l}	c_1	c_2	w_1	$E_\pi(N_{LL})$
0.0	0.2	1.0	13	33	27	15	−0.475	1.275	0.025	132.6
0.2	0.4	1.0	12	34	35	5	−0.350	1.500	0.150	138.8
0.4	0.6	1.0	14	34	41	30	−0.150	1.700	0.100	136.8
0.6	0.8	1.0	35	107	154	24	0.275	1.925	0.275	225.0

5.2. Schema B: Two limbs

A slightly more complex Limb-Leaf schema could include 6 doses of interest: limbs L_1 and L_2 each with two associated leaves, $l_{1,1},l_{1,2}$ and $l_{2,1},l_{2,2}$, respectively. Stage 1 includes L_1,L_2 and control L_0 where each arm would have sample size n_{1L}. Depending on the results of this first stage comparison, we might: i) terminate the study, ii) continue the study with further recruitment to L_0 and L_{k^*} (the best performing limb in the first stage) of n_{2L}' subjects each in the second stage, or iii) continue with recruitment of n_{2L} subjects each to L_0, and L_{k^*} as well n_{2l} subjects each to leaves $l_{k^*,1}$ and $l_{k^*,2}$. In case iii) we again allow final selection of the optimal dose by comparison of overall sample means, possibly across both stages. Evaluation of the results would be by the closed testing procedure utilizing pre-specified combination tests as described in Section 3. This could correspond, for instance, to a situation where investigators have a prior belief that one of two chosen doses may provide sufficient efficacy, however, they also want the option to further explore around the dose level that appears promising in order to fine tune towards an optimal dose.

The corresponding TSE design includes $L_0,L_1,l_{1,1},l_{1,2},L_2,l_{2,1},l_{2,2}$ as first stage doses, each with sample size n_1. If the first stage result of the best performing test dose d^* exceeds the preset threshold y_1, d^* would continue recruitment in the second stage along with control dose L_0 with n_2 subjects per arm. The efficacy of d^* relative to L_0 would then be determined by whether the overall sample mean of subjects assigned to d^* exceeds the pre-specified y_2.

The design constants to be set for the Limb-Leaf design are as before $n_{1L},n_{2L},n_{2L}',n_{2l}$, $\boldsymbol{c} = (c_1,c_2)$, and the weights w_1 and w_2. We minimize $E_\pi(N)$ defined in Section 4.2 with $\pi_0 = 0.8$, $\pi_{\Theta_{Limb}} = 0.1$, and $\pi_{\Theta_{Leaf}} = 0.1$. In this case, for given $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$, we let Θ_{Limb} be given by

$$\{\theta_{L_1} = \delta_1, \theta_{l_{1,1}} = \delta_1, \theta_{l_{1,2}} = \delta_1, \theta_{L_2} = \delta_3, \theta_{l_{2,1}} = \delta_2, \theta_{l_{2,2}} = \delta_2\},$$

and Θ_{Leaf} be given by

$$\{\theta_{L_1} = \delta_1, \theta_{l_{1,1}} = \delta_1, \theta_{l_{1,2}} = \delta_1, \theta_{L_2} = \delta_2, \theta_{l_{2,1}} = \delta_2, \theta_{l_{2,2}} = \delta_3\}.$$

Numerical optimization of the criterion $E_\pi(N)$ subject to contraints of 0.9 power in both Θ_{Limb} and Θ_{Leaf} configurations determines design constants.

The corresponding TSE-type design constants n_1,n_2,y_1, and y_2 are set as before. The previous criterion, $E_\pi(N)$ then reduces to $0.8E_0(N) + 0.2E_{\Theta_{TSE}}(N)$, where Θ_{TSE} is either Θ_{Limb} or Θ_{Leaf} given in the previous paragraph. The power constraints similarly reduce to the single restriction of 0.9 power in the Θ_{TSE} configuration. Numerical optimization subject to the power constraint then determines values of

Table 3
TSE design constants under Schema B.

δ_1	δ_2	δ_3	n_1	n_2	y_1	y_2	$E_\pi(N_{TSE})$
0.0	0.2	1.0	18	12	0.525	0.550	134.9
0.0	0.4	1.0	20	12	0.475	0.550	149.4
0.0	0.6	1.0	36	2	0.575	0.400	252.9
0.0	0.8	1.0	122	2	0.650	0.575	854.8
0.2	0.4	1.0	22	22	0.425	0.525	165.7
0.2	0.6	1.0	36	2	0.575	0.375	292.9
0.2	0.8	1.0	124	2	0.650	0.375	868.8
0.4	0.6	1.0	38	4	0.550	0.150	267.9
0.4	0.8	1.0	122	4	0.675	0.325	855.6
0.6	0.8	1.0	122	4	0.675	0.150	855.6

Table 4
Limb-Leaf design constants under Schema B.

δ_1	δ_2	δ_3	n_{1L}	n_{2L}	n_{2L}'	n_{2l}	c_1	c_2	w_1	$E_\pi(N_{LL})$
0.0	0.2	1.0	85	52	42	22	−0.225	0.675	0.050	426.0
0.0	0.4	1.0	33	40	37	14	−0.050	1.100	0.050	216.3
0.0	0.6	1.0	28	36	48	31	0.125	1.350	0.125	178.7
0.0	0.8	1.0	37	112	156	65	0.350	1.750	0.275	266.5
0.2	0.4	1.0	97	40	59	37	0.125	0.825	0.200	367.3
0.2	0.6	1.0	44	51	55	44	0.200	1.300	0.1125	218.7
0.2	0.8	1.0	39	111	148	12	0.325	1.575	0.175	275.5
0.4	0.6	1.0	90	54	88	36	0.200	1.050	0.325	352.0
0.4	0.8	1.0	50	108	178	6	0.350	1.625	0.075	294.5
0.6	0.8	1.0	143	130	188	28	−0.450	1.300	0.275	1064.7

n_1,n_2,y_1, and y_2.

Below we present two tables of optimized parameter values and performance characteristics for the two designs over ranges of $\boldsymbol{\delta}$ for schema B (Tables 3 and 4).

5.3. Interpretation

Interpretation of the comparison between designs is aided by the following tables. The key figure is $E_\pi(N_{LL})/E_\pi(N_{TSE})$, the ratio of risk adjusted expected total sample size between the Limb-Leaf and TSE designs. Where this ratio is less than 1, the comparison favors the Limb-Leaf approach. As in Section 5.1, cells of Schema A which are redundant in terms of Θ_{Limb},Θ_{Leaf}, and Θ_{TSE} are combined (Tables 5 and 6).

We observe such a favorable ratio across both schemas for $\boldsymbol{\delta} = (0.0, 0.6, 1.0)$, $(0.0, 0.8, 1.0)$, $(0.2, 0.6, 1.0)$, $(0.2, 0.8, 1.0)$, and $(0.4, 0.8, 1.0)$. These schemas correspond to situations where a Limb-Leaf strategy as presented in Section 3 would be appropriate: The initially chosen limb dose achieves 60% or more of the desired signal; a further fine-tuning of the dose level is then possible to maximize performance. A subtle departure from this pattern occurs in Schema B with $\boldsymbol{\delta} = (0.6,0.8,1.0)$. If a first stage selection is necessary between two limbs, too small a separation between their response levels will require a large first stage sample size, disproportionately hurting the performance of the Limb-Leaf approach.

We acknowledge some limitations to the proper setting for the Limb-Leaf design, especially that prior knowledge or informed judgement are needed to guide the selection of limbs, leaves, and to support the selection of design constants appropriate for the locatable effect assumption.

Table 5
Sample size comparison of Limb-Leaf design and TSE design under Schema A.

δ_1	δ_2	δ_3	$E_\pi(N_{TSE})$	$E_\pi(N_{LL})$	$\frac{E_\pi(N_{LL})}{E_\pi(N_{TSE})}$
0.0	0.2	1.0	82.7	132.6	1.6
0.2	0.4	1.0	125.0	138.8	1.1
0.4	0.6	1.0	162.1	136.8	0.8
0.6	0.8	1.0	502.0	225.0	0.4

Table 6
Sample size comparison of Limb-Leaf design and TSE design under Schema B.

δ_1	δ_2	δ_3	$E_\pi(N_{TSE})$	$E_\pi(N_{LL})$	$\frac{E_\pi(N_{LL})}{E_\pi(N_{TSE})}$
0.0	0.2	1.0	134.9	426.0	3.2
0.0	0.4	1.0	149.4	216.3	1.4
0.0	0.6	1.0	252.9	178.7	0.7
0.0	0.8	1.0	854.8	266.5	0.3
0.2	0.4	1.0	165.7	367.3	2.2
0.2	0.6	1.0	292.9	218.7	0.7
0.2	0.8	1.0	868.8	275.5	0.3
0.4	0.6	1.0	267.8	352.0	1.3
0.4	0.8	1.0	855.6	294.5	0.3
0.6	0.8	1.0	855.6	1064.7	1.2

In practice, we would expect to use a maximum of three limbs and no more than two leaves per limb. Sufficient basis should exist from preliminary clinical and pre-clinical studies, pharmacokinetic and pharmacodynamics considerations, and/or expert judgement from related research to guide the choices of the limb doses and to motivate the leaf doses as clinically meaningful modifications of their respective limbs.

Such considerations are equally important in the selection of the vector $\boldsymbol{\delta}$. As in any trial, δ_3, as the effect level desired to detect, should be plausible and planners should avoid over-optimism. The component δ_2 should be chosen from a balanced judgment of possibilities such that an effect of level δ_2 should indicate a region of promising activity for finding an effect within it of magnitude δ_3. On the other hand, δ_1 should represent a baseline level of effect that would not itself identify an especially promising region. We advise that in planning to use a Limb-Leaf design one should give careful explanation in detail of these design choices and provide justification and citation of sources of information on which these design decisions are based. Guidance for making conservative choices that are robust to moderate misspecifications is given in Section 5.4.

5.4. Robustness against distortions for selected configurations

Recall that in Sections 5.1 and 5.2, for any given $\boldsymbol{\delta} = (\delta_1,\delta_2,\delta_3)$ two unfavorable configurations, Θ_{Limb} and Θ_{Limb}, are chosen as alternatives. We denote them here as $\Theta_{Limb}(\boldsymbol{\delta})$ and $\Theta_{Leaf}(\boldsymbol{\delta})$, respectively, to emphasize their dependence on $\boldsymbol{\delta}$. Under these alternatives the Limb-Leaf design constants are calculated to achieve a 90% power. In this section, we study the robustness of the Limb-Leaf design against misspecifications of these two unfavorable alternatives. Specifically, for any given $\boldsymbol{\delta}$ in Tables 2 and 4, we investigate how far we can modify each component into a possible $\boldsymbol{\delta}'$ such that the designs specified in Tables 2 and 4 maintain at least 85% or 80% power under the two modified configurations $\Theta_{Limb}(\boldsymbol{\delta}')$ and $\Theta_{Leaf}(\boldsymbol{\delta}')$. We focus on five favorable cases for both schemas with $\boldsymbol{\delta}$ values of (0.0,0.6,1.0),(0.0,0.8,1.0),(0.2,0.6,1.0), (0.2,0.8,1.0), and (0.4,0.8,1.0). As in Section 5.1, under Schema A redundant cells are combined (Tables 7 and 8).

We remark that in all cases where power was maintained, the overall change in risk adjusted expected sample size was within 10% of the unperturbed value (results not shown).

The design is seen to be robust to variation in δ_1', and especially so for δ_1' less than δ_1. Since the true underlying $\boldsymbol{\delta}$ is unknown to the investigator, a conservative design specification would set a value for δ_1 close to the upper limit of what is assumed possible. The design is sensitive at least in some scenarios to values of δ_2' greater than δ_2, however it is robust to values of δ_2' that are less than δ_2; a conservative specification would choose δ_2 near the upper limit of its assumed plausible range. Instances where δ_3' is less than δ_3 represent the greatest threat to the power of the design. For example, with the

Table 7
Allowed ranges of $\boldsymbol{\delta}' = (\delta_1', \delta_2', \delta_3')$ for the Limb-Leaf designs under schema A.

			Ranges maintaining 85% power			Ranges maintaining 80% power		
δ_1	δ_2	δ_3	δ_1'	δ_2'	δ_3'	δ_1'	δ_2'	δ_3'
0.2	0.6	1.0	$(-\infty, 0.60]$	[0.29,0.67]	$[0.95, \infty)$	$(-\infty, 0.60]$	[0.21,0.72]	$[0.91, \infty)$
0.4	0.8	1.0	$(-\infty, 0.80]$	[0.60,0.83]	$[0.97, \infty)$	$(-\infty, 0.80]$	[0.55,0.86]	$[0.95, \infty)$

Table 8
Allowed ranges of $\boldsymbol{\delta}' = (\delta_1', \delta_2', \delta_3')$ for the Limb-Leaf designs under schema B.

			Ranges maintaining 85% power			Ranges maintaining 80% power		
δ_1	δ_2	δ_3	δ_1'	δ_2'	δ_3'	δ_1'	δ_2'	δ_3'
0.0	0.6	1.0	[−0.58,0.18]	[0.47,0.69]	$[0.95, \infty)$	[−0.68,0.28]	[0.41,0.73]	$[0.90, \infty)$
0.0	0.8	1.0	[−0.50,0.43]	[0.61,0.83]	$[0.98, \infty)$	[−0.56,0.52]	[0.56,0.86]	$[0.95, \infty)$
0.2	0.6	1.0	[−0.48,0.33]	[0.50,0.73]	$[0.91, \infty)$	[−0.57,0.38]	[0.45,0.77]	$[0.88, \infty)$
0.4	0.8	1.0	[−0.49,0.44]	[0.59,0.83]	$[0.98, \infty)$	[−0.57,0.52]	[0.55,0.86]	$[0.95, \infty)$
0.4	0.8	1.0	[−0.43,0.51]	[0.65,0.84]	$[0.97, \infty)$	[−0.53,0.57]	[0.62,0.86]	$[0.95, \infty)$

assumed value of $\boldsymbol{\delta} = (0.0, 0.8, 1.0)$, a δ_3' value corresponding to a 2.5% reduction in δ_3 is associated with a 5% reduction in power while a 5% reduction in δ_3 is associated with a 10% reduction in power. On the other hand, values of δ_3' greater than δ_3 are not associated with negative impacts on power. For design purposes, an initial estimate of δ_3 on the low end of its assumed range would be a conservative choice.

6. Discussion

The proposed Limb-Leaf procedure is a structured exploration strategy that builds on earlier adaptive designs such as those proposed in Bauer and Keiser [1] and Bauer and Köhne [2]. It is a generalization of the design by Thall et al. [19] targeted to improve performance under a possibly non-monotonic dose-response relationship. These have become more relevant in recent years as new therapies often entail increasingly complex mechanisms of action and combination therapies may not allow a clear ordering among their dosage levels. When $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$ is chosen appropriately, for example when δ_2 is at least 60% of δ_3, the saving in total sample size using a Limb-Leaf design is substantial compared with the TSE design. Limitations on the proper setting for the Limb-Leaf design and the prior knowledge and judgement needed to implement it were discussed in Section 5.3. Guidance on robust parameter choices was given in Section 5.4.

The Limb-Leaf design can be generalized to allow greater flexibility. For example, it can permit different numbers of leaves for different limbs or leaves shared by more than one limb. Furthermore, the precise location (or formulation) of leaves need not be specified in full detail ahead of the second stage; error is controlled and a valid design is achieved as long as we account for their existence. Theoretically, at least, this allows the Limb-Leaf design to consider a larger class of doses than could ever be tested in a TSE-style design — in fact a countably infinite class. In addition, other combination rules besides the weighted inverse normal rule used in this paper for stage-wise Z-values are possible. Fisher's combination rule, for instance, may be a robust choice if some doses are expected to perform especially poorly and one wants to limit their impact on tests of effect of unrelated leaf doses. Further study on combination rules for robustness and efficiency is warranted. Along with different combination rules, different rules for stochastic curtailment may be employed. The efficiency gain of group sequential continuation [14] as well as dropping poorly performing treatment arms at potentially earlier interim points is worth investigation as is the comparison of relative benefit in a TSE-style approach.

To accomplish point and interval estimation in the context of Limb-Leaf designs existing methods based on bias adjusted estimators [17], test inversion to form confidence regions, and median unbiased point estimators are applicable [4,6,8,11]. The specific implementation of these methods and comparison of their performance will be treated in future research.

A final issue is the computational complexity of parameter optimization in a Limb-Leaf type design. An R program developed by Mr. X. Lu using simulated annealing for nonlinear optimization appears to offer substantially improved speed over a basic grid search method such that a design can be optimized in less than one hour on a Windows desktop computer with 3 GHz processing speed. A preliminary version of this software is available from the authors upon request. A full version is planned for public release as an R package.

Acknowledgments

We thank Mr X. Lu for offering his expertise in statistical programming.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A

Proof of Proposition 1. For simplicity assume that the effects of no two doses are identical. By renumbering the limbs as necessary, we may take the limb effects θ_{L_k}, $k = 1, \ldots, K$, as forming an increasing sequence. Choose δ_1 such that $\theta_{L_{K-1}} < \delta_1 < \theta_{L_K}$. Then, the effects of all leaf doses associated with θ_{L_K} may be written as an increasing sequence$\theta_{l_{k,1}}, \ldots, \theta_{l_{k,m_k}}$

There are then two subcases. The first applies if $\theta_{l_{k,m_k}} > \theta_{L_K}$. Choose δ_2 such that $\delta_1 < \delta_2 = \theta_{L_K}$. Let $i^* = \min\{i\colon \theta_{l_{k,i}} > \theta_{L_K}\}$, the lowest leaf dose whose effect exceeds that of L_K. We may then choose δ_3 such that $\theta_{L_K} < \delta_3 < \theta_{l_{K,i^*}}$. The conditions for a locatable effect with respect to $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$ are then satisfied.

If $\theta_{l_{k,m_k}} < \theta_{L_K}$ then let $\gamma = \max\{\delta_1, \theta_{l_{k,m_k}}, \theta_{L_{K-1}}\}$. Choose δ_2 and δ_3 such that $\gamma < \delta_2 < \delta_3 < \theta_{L_K}$. The conditions for a locatable effect with respect to $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)$ are similarly satisfied.

This construction may be modified to produce other solutions and to allow perfect ties between dose effects although the notation becomes more complicated.

Proof of Theorem 1. We demonstrate that when $H_{0,\hat{d}^*}$ is rejected, so are all the intersection hypotheses that contain it. Given data, we use $z_{i,d}$ to denote the observed $Z_{i,d}$ and relabel doses as necessary for convenience of notation.

Suppose $c_1 < \hat{\theta}_{1,L_{k^*}} \le c_2$ and $\hat{d}^* = L_{k^*}$. A composite hypothesis including $H_{0,L_{k^*}}$ may be written as $H_{0,D}$ where

$$D = \{L_{k^*}\} \cup \left\{L_1,...,L_{K'}\right\} \cup \left\{l_{k^*,1},...,l_{k^*,m'_{k^*}}\right\} \cup \left\{l_{1,1},...,l_{1,m'_1}\right\} \cup ...\cup\left\{l_{K',1},...,l_{K',m'_{K'}}\right\}.$$

Here $L_j \ne L_{k^*}$, $K' \le K-1$, and $0 \le m'_j \le m_j$, $j = 1,\ldots,K'$. For K' or m'_j equal to zero, the corresponding terms in braces are counted as empty. Let $k'' \in \{0...m_{k^*}\}$ denote the number of leaves in D that were utilized (randomized to) after the interim decision. Then the relations

$$\begin{aligned} z_D &= w_1\Phi^{-1}\{F_{1+K'}(z_{1,L_{k^*}})\} + w_2\Phi^{-1}\{G_{Limb+k''}[\max(z_{2,L_{k^*}}, z_{2,l_{k^*,l}}, l = 1,...,k'')]\} \\ &\ge w_1\Phi^{-1}\{F_K(z_{1,L_{k^*}})\} + w_2\Phi^{-1}(\min_{0\le k\le m_{k^*}}\{G_{Limb+k}[\max(z_{2,L_{k^*}}, z_{2,l_{k^*,(k)}})]\}) > z_\alpha \end{aligned}$$

imply the rejection of H_D. It follows from their definitions that $F_{1+K'}(x) \ge F_K(x) \quad \forall\, x$, and that $G_{Limb+k''}(x) \ge G_{Limb+m_{k^*}}(x) \quad \forall\, x$, given any first stage data. The second inequality is assumed from the rejection of $H_{0,L_{k^*}}$

Suppose $c_1 < \hat{\theta}_{1,L_{k^*}} \le c_2$ and $\hat{d}^* = l_{k^*,l^*}$ for some k^*, l^*. We consider D in three cases.

Case 1. If D contains L_{k^*}, using the same notations as above,

$$\begin{aligned} z_D &= w_1\Phi^{-1}\{F_{1+K'}(z_{1,L_{k^*}})\} + w_2\Phi^{-1}\{G_{Limb+k''}[\max(z_{2,L_{k^*}}, z_{2,l_{k^*,l^*}})]\} \\ &\ge w_1\Phi^{-1}\{F_K(z_{1,L_{k^*}})\} + w_2\Phi^{-1}\{G_{Limb+m_{k^*}}[\max(z_{2,L_{k^*}}, z_{2,l_{k^*,l^*}})]\} > z_\alpha, \end{aligned}$$

where the last inequality is a rejection condition for $H_{0,l_{k^*,l^*}}$.

Case 2. If D does not contain L_{k^*} but contains some other limb, we write

$$D = \{L_1, ...,L_{K'}\} \cup \{l_{k^*,1}, ...,l_{k^*,m'_{k^*}}\} \cup \{l_{1,1}, ...,l_{1,m'_1}\} \cup ...\cup\{l_{K',1}, ...,l_{K',m'_{K'}}\}$$

using the previous notations. Then,

$$\begin{aligned} z_D &= w_1\Phi^{-1}\{F_{K'}(\max\{z_{1,L_1},...,z_{1,L_{K'}}\})\} + w_2\Phi^{-1}\{G_{k''}(z_{2,l_{k^*,l^*}})\} \\ &\ge w_1\Phi^{-1}\{\min_{1\le k\le K-1}[F_k(z_{1,L_{(k)}})]\} + w_2\Phi^{-1}\{G_{m_{k^*}}(z_{2,l_{k^*,l^*}})\} > z_\alpha. \end{aligned}$$

The first inequality holds since, given any first stage data, $G_{k''}(x) \ge G_{m_{k^*}}, \quad \forall\, x$, from their definitions. The second inequality is a rejection condition for $H_{0,l_{k^*,l^*}}$.

Case 3. If D does not contain any limbs, we write

$$D = \{l_{k^*,1}, ...,l_{k^*,m'_{k^*}}\} \cup \{l_{1,1}, ...,l_{1,m'_1}\} \cup ...\cup\{l_{K',1}, ...,l_{K',m'_{K'}}\}.$$

Then,

$$z_D = \Phi^{-1}\{G_{k''}(z_{2,l_{k^*,l^*}})\} \ge \Phi^{-1}\{G_{m_{k^*}}(z_{2,l_{k^*,l^*}})\} > z_\alpha,$$

The second inequality is a rejection condition for $H_{0,l_{k^*,l^*}}$.

Finally, suppose $\hat{\theta}_{1,L_{k^*}} > c_2$. We write

$$D = \{L_{k^*}\} \cup \left\{L_1,...,L_{K'}\right\} \cup \left\{l_{k^*,1},...,l_{k^*,m'_{k^*}}\right\} \cup \left\{l_{1,1},...,l_{1,m'_1}\right\} \cup ...\cup\left\{l_{K',1},...,l_{K',m'_{K'}}\right\}.$$

Since

$$z_D = w_1\Phi^{-1}\{F_{1+K'}(z_{1,L_{k^*}})\} + w_2 z_{2,L_{k^*}} \ge w_1\Phi^{-1}\{F_K(z_{1,L_{k^*}})\} + w_2 z_{2,L_{k^*}} > z_\alpha,$$

where the second inequality is assumed from the rejection of $H_{0,L_{k^*}}$, $H_{0,D}$ is rejected. This concludes the proof of the theorem.

Proof of Theorem 2.

1. For normally distributed outcomes with unknown σ^2, let $Z_{1,L_k} = (\overline{X}_{1,L_k} - \overline{X}_{1,L_0}) \Big/ \sqrt{\frac{2\hat{\sigma}^2_{1,L_k}}{n_{1L}}}$ denote the first stage test statistic of the effect of dose L_k, $k = 1,\ldots,K$, where $\hat{\sigma}^2_{1,L_k}$ is the pooled estimate of σ^2 in the control arm and arm L_k from stage 1. Instead of being normal with known variance, each Z_{1,L_k} now follows a t-distribution with $2(n_{1L} - 1)$ df under its associated null hypothesis of no treatment effect.
   Similarly, let $Z_{2,L_{k^*}} = (\overline{X}_{2,L_{k^*}} - \overline{X}_{2,L_0}) \Big/ \sqrt{\frac{2\hat{\sigma}^2_{2,L_{k^*}}}{n_{2L}}}$ or $(\overline{X}_{2,L_{k^*}} - \overline{X}_{2,L_0}) \Big/ \sqrt{\frac{2\hat{\sigma}^2_{2,L_{k^*}}}{n'_{2L}}}$ as required denote the second stage test statistic for the effect of dose L_{k^*}, where $\hat{\sigma}^2_{2,L_k}$ is the pooled estimate of σ^2 in the control arm and arm L_{k^*} in stage 2. $Z_{2,L_{k^*}}$ follows a t-distribution with $2(n_{2L} - 1)$ df under its associated null hypothesis. We may also define $Z_{2,l_{k^*,l}} = (\overline{X}_{2,l_{k^*,l}} - \overline{X}_{2,L_0}) \Big/ \sqrt{\hat{\sigma}^2_{2,l_k}(\frac{1}{n_{2l}} + \frac{1}{n_{2L}})}$, $l = 1, ...,m_{k^*}$, where $\hat{\sigma}^2_{2,l_k}$ is the pooled estimate of σ^2 in the control arm and arm $l_{k^*,l}$ in stage 2, such that under the associated null hypothesis it follows a t-distribution with $(n_{2L} + n_{2l} - 2)$ df.
   For prespecified values of n_{1L}, n_{2L}, $n_{2L}'$and n_{2l}, the functions F_K, $G_{m_{k^*}}$, $G_{Limb+m_{k^*}}$ are then well defined under each adaptation; they can be calculated or approximated by Monte-Carlo simulation from the reference case of unit variance since the parameter σ^2 has been removed by invariance. Subject only to these changes in interpretation, the design of Section 3.2.2 applies and the proof of Theorem 1 applies to guarantee

control of the FWER at level α.

The Satterthwaite approximation may be used as usual to accommodate unequal variances.

2. A general method to draw inferences about parameters of interest in large samples is based on the calculation of the score statistic and observed Fisher information. The method is valid for general endpoints under regularity conditions satisfied, for instance, when the response distribution belongs to an exponential family; nuisance parameters are accommodated through use of consistent estimators. Details are given by Stallard and Todd [16]. Theoretical justification is given by Whitehead [21].

   Let θ_d be a measure of the superiority of treatment d relative to control with $\theta_d > 0$ indicating superiority, and let $S_{i,d}$ and $V_{i,d}$ denote the efficient score and observed Fisher Information for θ_d at stages $i = 1,2$, (evaluated at $\theta_d = 0$). It may be shown that the asymptotic joint distribution of the $Z_{1,L_1} = \frac{S_{1,L_1}}{\sqrt{V_{1,L_1}}}, \ldots, Z_{1,L_K} = \frac{S_{1,L_K}}{\sqrt{V_{1,L_K}}}$ is multivariate normal with mean proportional to the vector of assumed effect sizes and fixed covariance matrix. Similarly, given any adaptation decision applied in the second stage, $Z_{2,L_{k^*}} = \frac{S_{2,L_{k^*}}}{\sqrt{V_{2,L_k^*}}}, Z_{2,l_{k^*,1}} = \frac{S_{2,l_{k^*,1}}}{\sqrt{V_{2,l_{k^*,1}}}}, \ldots, Z_{2,l_{k^*,m_{k^*}}} = \frac{S_{2,l_{k^*,m_{k^*}}}}{\sqrt{V_{2,l_{k^*,m_{k^*}}}}}$ have an asymptotic distribution with mean proportional to the vector of assumed effect sizes and fixed covariance matrix. Stagewise effect estimates may be calculated as $\hat{\theta}_{1,L_{k^*}} = \frac{Z_{1,L_{k^*}}}{V_{1,L_k^*}}, \hat{\theta}_{2,L_{k^*}} = \frac{Z_{2,L_{k^*}}}{V_{2,L_k^*}}$, and $\hat{\theta}_{2,l_{k^*,l}} = \frac{Z_{2,l_{k^*,l}}}{V_{2,l_{k^*,l}}}, l = 1, \ldots, m_{k^*}$.

   It follows that the plan of a Limb-Leaf Design of Section 3.2.2 can be implemented with the above new meanings for the symbols $\hat{\theta}_{1,L_{k^*}}, \hat{\theta}_{2,L_{k^*}}$, and $\hat{\theta}_{2,l_{k^*,l}}, l = 1, \ldots, m_{k^*}$, as well as $Z_{1,L_1}, \ldots, Z_{1,L_K}$ and $Z_{2,L_{k^*}}, Z_{2,l_{k^*,1}}, \ldots, Z_{2,l_{k^*,m_{K^*}}}$. The functions $F_K, G_{m_{k^*}}, G_{Limb+m_{k^*}}$ are then well defined. Subject only to these changes in interpretation, the design of Section 3.2.2 applies and the proof of Theorem 1 shows that the FWER is controlled approximately at level α in large samples.

References

[1] P. Bauer, M. Keiser, Combining different phases in the development of medical treatments within a single trial, Stat. Med. 18 (1999) 1833–1848.

[2] P. Bauer, K. Köhne, Evaluation of experiments with adaptive interim analyses, Biometrics 50 (1994) 1029–1041.

[3] M.F. Beal, A.E. Lang, A.C. Ludolph, Neurodegenerative Diseases: Neurobiology, Pathogenesis and Therapeutics, Cambridge University Press, Cambridge, England, 2005.

[4] F. Bretz, H. Schmidli, F. König, A. Racine, W. Maurer, Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts, Biom. J. 48 (2006) 623–634.

[5] J.P. Hubert, J.C. Delumeau, J. Glowinski, J. Prémont, A. Doble, Antagonism by riluzole of entry of calcium evoked by NMDA and veratridine in rate cultured granule cells: evidence for a dual mechanism of action, Br. J. Pharmacol. 113 (1994) 261–267.

[6] C. Jennison, B. Turnbull, Group Sequential Methods with Applications to Clinical Trials, CRC Press, New York, New York, 2000.

[7] C. Jennison, B. Turnbull, Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations, Biom. J. 48 (2006) 650–655.

[8] C. Jennison, B. Turnbull, Adaptive seamless designs: selection and prospective testing of hypotheses, J. Biopharm. Stat. 17 (2007) 1135–1161.

[9] Kaufmann et al, Phase II trial of CoQ10 for ALS finds insufficient evidence to justify phase III, Ann. Neurol. 66 (2) (2009) 235–244.

[10] B. Levin, The futility study - progress over the last decade, Contemp. Clin. Trials 45 (A) (2015) 69–75.

[11] Q. Liu, M.A. Proschan, G.W. Pledger, A unified theory of Two-Stage adaptive designs, J. Am. Stat. Assoc. 97 (2002) 1034–1041.

[12] A.C. Ludolph, S. Jesse, Review: evidence-based drug treatment in amyotrophic lateral sclerosis and upcoming clinical trials, Ther. Adv. Neurol. Disord. 2 (2009) 319–326.

[13] R. Marcus, E. Peritz, K.R. Gabriel, On closed testing procedures with special reference to ordered analysis of variance, Biometrika 63 (1976).

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

[15] K.M. Noh, J.Y. Hwang, H.C. Shin, J.Y. Koh, A novel neuroprotective mechanism of riluzole: direct inhibition of protein kinase C, Neurobiol. Disord. 7 (2000) 375–383.

[16] N. Stallard, S. Todd, Sequential designs for phase III clinical trials incorporating treatment selection, Stat. Med. 22 (2003) 689–703.

[17] N. Stallard, S. Todd, Point estimates and confidence regions for sequential trials involving selection, J. Stat. Plan. Infer. 135 (2005) 402–419.

[18] A. Tamhane, Y. Hochberg, C. Dunnett, Multiple test procedures for dose finding, Biometrics 52 (1996) 21–37.

[19] P.F. Thall, R. Simon, S.S. Ellenberg, Two-stage selection and testing designs for comparative clinical trials, Biometrika 75 (1988) 303–310.

[20] S. Todd, N. Stallard, A new clinical design combing phases 2 and 3: sequential designs with treatment selection and a change of endpoint, Drug Inf. J. 39 (2005) 109–118.

[21] J. Whitehead, The Design and Analysis of Sequential Clinical Trials, Revised Second Edition, Wiley, England, Chichester, 1997.