Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–10

DOI: 10.1177/0962280219859387
journals.sagepub.com/home/smm
SAGE

Adding dose modifications into Phase II and Phase II/III seamless trials

John Spivack,[1] Bin Cheng[2] and Bruce Levin[2]

Abstract

We present a technique for adding dose modifications into seamless Phase II and Phase II/III trials featuring dose selection at an interim analysis. The method is convenient to apply and can be used either in a fully prespecified, structured way or as a response to new considerations that emerge at interim. Strong control of the familywise error rate regarding false declarations of efficacy versus control is maintained. Two examples are given. One illustrates how the method could potentially "save" a trial performed in a Phase II context. The other is a seamless Phase II/III trial that uses an adaptive exploration strategy for an assumed nonmonotonic dose-response curve. It can result in greatly improved efficiency over a standard "promote the winner" rule.

Keywords

Adaptive design, dose addition, dose modification, familywise error rate, nonmonotonic dose response

1 Introduction

There are good reasons to believe that, in certain circumstances, it could be advantageous to allow the introduction of dose modifications into a study of a new drug or other therapy. In Phase II or in adaptive Phase III designs, a method using strategic dose addition might lead to more success in producing a final dose recommendation and/or require fewer resources than a standard "promote the winner" design. In spite of the attraction of this idea, there are limitations to existing dose-addition methods.

To consider introducing a dose modification, there would have to be sufficient scientific motivation to permit the newly added arm. There would have to be no assumption of monotonicity in the overall dose response, whether it is regarded in terms of a single endpoint, a combined endpoint, or an overall utility involving, for instance, a composite measure of efficacy, side effects, and logistical burden. In contrast, if such a strict monotonicity assumption is made, the maximum tolerated dose (MTD) is the only logical choice to investigate.

In terms of a design's statistical operating characteristics, such a dose addition method would have to be adequately powered under realistic sample sizes for detection of the assumed levels of effect. Economically and logistically, the presumed chance of success on the new arm would have to justify the additional outlay of patients and resources.

Although such a strategy has costs, especially in the case of Phase III trials, the cost of the failure of an entire research program due to a faulty dose selection and the cost to medicine of the missed opportunity to correctly identify a successful therapy can be much larger. Further, even if a trial concludes the existence of an effect significantly different than zero, these results may not be sufficient to persuade clinicians and patients that the particular treatment is worthwhile—another missed opportunity. Indeed, greater optimization may lead to more convincing evidence of benefit. The considerations involved in these decisions are discussed further in a comprehensive review article by Cohen et al.,[1] and the references therein.

Cohen et al.[1] note the desirability of dose addition, but identify shortcomings in existing methods and their applications. A notable proposal in the Phase II/III context was made by Chang and Wang,[2] though their approach is complex, involving three stages. They give an application to unimodal dose-response curves,

[1]Department of Environmental Medicine and Public Health, Mount Sinai Medical Center, New York, NY, USA
[2]Department of Biostatistics, Columbia University, New York, NY, USA

Corresponding author:
John Spivack, Mount Sinai Health System, 1425 Madison Avenue, New York, NY 10019, USA.
Email: john.spivack@mountsinai.org

but do not provide a more complete characterization of dose-response curves for which the design would be advantageous. Improvements in sample size over more standard ''promote the winner'' designs of approximately 10–20% are stated. For dose-response curves, which are not necessarily unimodal, Spivack et al.[3] provide a more detailed characterization of the problem in terms of doses arranged in a ''Limb–Leaf System'' and dose-response curves having ''locatable effects'' in their proposal for a two-stage ''Limb–Leaf'' design. The improvement over standard designs in terms of risk-adjusted expected sample size (RAESS) can exceed 50%. However, this method uses complex formulas for the combination of evidence across the stages and is based on a generic strategy of applying combination rules to stagewise *p*-values. Consequently, although statistical power is guaranteed based on ''unfavorable'' configurations, in pathological cases, power may fall below the intended design constraints, and the confirmation of one dose's effect may depend in a complicated way on the observed effects at unrelated doses.

We emphasize that in an adaptive exploration (AdEx) experiment, the rules for selection and addition of dose modifications should allow some flexibility. The dose with best observed first-stage effect, for instance, may not be the best candidate for promotion into the second stage if concerns emerge about its toxicity or tolerability. Similar issues may influence the decision to add an additional arm, regarded as a modification to a promoted arm, into the second stage of an experiment or not to do so. However, any allowed flexibility must not undermine the strict control of the familywise error rate (FWER).

An example of a suitable setting for application of these designs is given in *The Trial of High Dose Coenzyme Q10 in ALS* (QALS Trial), published by Kaufmann et al.[4] The intention was to investigate high doses of Coenzyme Q10 (CoQ10) as a possible therapy for amyotrophic lateral sclerosis (ALS); it was not assumed that the dose-response curve would be monotonic, and potential toxicity was a major concern. The study was designed in two stages, a selection stage followed by a non-superiority test. 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 of selection bias and type 1 error.

The outcome of the first stage of the trial was that the higher dose did somewhat better than the lower dose on the outcome measure (the ALS functional rating scale, revised (ALSFRSr)) and had high tolerability. After continuation to the second stage, the test statistic associated with this higher dose was nominally sufficient to avoid a declaration of statistically significant non-superiority. Nonetheless, investigators did not consider the evidence promising enough to give it full endorsement. Further details are given in Kaufmann et al.[4]

Notwithstanding the importance of certain key design features in the QALS trial, it is easy to imagine that in this or a similar study, it might be useful 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 add doses both above and below it, this freedom might have been attractive to investigators. It is perhaps possible that an efficacious dose might have been among those added, and this could have earned full endorsement. We would like to make such options available to investigators by design in areas such as ALS research. In such difficult diseases, it is relatively more worthwhile to put in the additional effort and expense to explore a candidate therapy with greater care than it would be if there were a large number of other candidate therapies already in the development pipeline.

Other potential applications where a more detailed examination of the dose-response curve could be advantageous include trials of multi-component combination therapies and trials of complex treatment regimes in personalized medicine. In such cases, a full strategy of AdEx, adding dose modifications in response to certain criteria, could be used. This is particularly true of applications such as those discussed in the Drugs for Neglected Diseases[5] initiative, where effective therapies may already exist, but the need to optimize or nearly optimize drug combinations and deployment strategies as soon as possible and on limited budgets can be critical.

2 Characterization of the procedure

In order to focus on the essentials of the problem, we begin with the case where the outcomes associated with dose d follow the distribution $N(\mu_d, 1)$, where $\delta_d = \mu_d - \mu_0$ is the effect relative to control.

Let the individual doses in the experiment be organized as $\{L_1, l_{1,1}, \ldots, l_{1,m_1}; \ldots; L_K, l_{K,1}, \ldots, l_{K,m_K}\}$, where $\{L_1, \ldots, L_K\}$ and $K \geq 1$ are considered as the main doses, and for each such L_k, the doses $\{l_{k,1}, \ldots, l_{k,m_k}\}$ are associated dose modifications, thought of as small to moderate scale alterations of dose L_k. While the exact formulations of $\{L_1, \ldots, L_K\}$ must be prespecified, the exact formulations of $\{l_{k,1}, \ldots, l_{k,m_k}\}$ may be made at an interim point before their first actual use. Denote the respective effects of this collection relative to control by

$\{\delta_{L_1}, \delta_{l_{1,1}}, \ldots, \delta_{l_{1m_1}}, ; \ldots ; \delta_{L_K}, \delta_{l_{K,1}}, \ldots, \delta_{l_{Km_K}}\}$. Given design constants n_{1L}, n_{2L}, n_{2La}, n_{2la}, c_1, c_2, and α_1, we consider a selection and promotion design of the following form.

Procedure:

(1) First stage. Use sample size n_{1L} for each main dose and control. Calculate $\overline{Y}_{1,1}, \ldots, \bar{Y}_{K,1}$, the first-stage average observed effects on each main dose relative to control. Let $\bar{Y}_{L^*,1}$ be the maximum, corresponding to dose selection L^*.
(2) Interim analysis with halt, selection, or selection and addition decision.
   (a) If $\bar{Y}_{L^*,1} < c_1$, then stop;
   (b) If $c_1 \le \bar{Y}_{L^*,1} \le c_2$, then promote L^* with sample size n_{2La} for L^* and control and add the associated modifications, with sample size n_{2la} on each;
   (c) If $\bar{Y}_{L^*,1} > c_2$, promote L^* only, with second stage sample size n_{2L} for L^* and control.
(3) Final analysis. At the end of the second stage, perform the final analysis as follows.

In case (b): calculate $\bar{Y}_{L^*,2a}$, the second stage average observed effect on the selected main dose relative to control.

(i) If $\bar{Y}_{L^*,\,overall} = \frac{n_{1L}\,\bar{Y}_{L^*,1} + \sqrt{\frac{n_{2L}}{n_{2La}}}\,n_{2La}\,\bar{Y}_{L^*,2a}}{n_{1L}+n_{2L}} \le y_{cutoff}$, then stop without rejecting any hypothesis;
Note: In this case, where modifications are added and n_{2La} is used as the modified sample size for the promoted main dose and control, a rescaling by a factor of $\sqrt{n_{2L}/n_{2La}}$ is applied to the observed effect of the second-stage main dose. The rescaling enforces that there will be the same null distribution (mean and variance) of the overall normally distributed test statistic as in case (c). If no such sample size modification is made, $n_{2L} = n_{2La}$ and the scaling factor is 1.
(ii) If $\bar{Y}_{L^*,\,overall} > y_{cutoff}$, then reject H_{L^*} to conclude the presence of an effect on L^*. Test the effect of each added modification at level α_1 from the second-stage data (using the standard test for a difference of normal means) in step-down sequence according to their indices. Select the final recommendation among the confirmed leaves using their treatment effect estimates.

In case (c): calculate $\bar{Y}_{L^*,2}$, the second-stage average observed effect on the selected main dose relative to control.

If $\bar{Y}_{L^*,\,overall} = \frac{n_{1L}\,\bar{Y}_{L^*,1} + n_{2L}\,\bar{Y}_{L^*,2}}{n_{1L}+n_{2L}} > y_{cutoff}$, then reject H_{L^*}, otherwise fail to reject.

Since constants c_1 and c_2 are on the same scale as observed effect sizes, they may be chosen at the initial design stage based on clinical judgment of what observed effect size would justify continuation and what observed effect size would justify exclusive focus on the promising limb, respectively. Constants y_{cutoff} and α_1 can be found by the computational methods presented, for instance, by Genz and Bretz[6]; however, Monte Carlo methods may be a good shortcut. Simulations with 10^6 iterations run on current computers in minutes and provide high enough accuracy for well-regulated error control to the third decimal place. If higher accuracy is required, however, the specialized computational methods may be preferred.

Theorem 1. Let n_{1L}, n_{2L}, n_{2La}, n_{2la}, c_1, and c_2 be given. Let y_{cutoff} be calculated such that $P_{NULL}(\bar{Y}_{L^*,1} > c_1, \bar{Y}_{L^*,\,overall} > y_{cutoff}) = \alpha$ using the above procedure and selection strategy where "NULL" denotes the global null hypothesis. Let $\alpha_1 = \alpha - P_{H^*}(\bar{Y}_{L^*,1} > c_1,\ \bar{Y}_{L^*,\,overall} > y_{cutoff})$ under the above selection strategy where H^* is the modification of the global null configuration with effect of $-\infty$ on L_1 and effects 0 on $L_2, \ldots, L_K$. Then the above procedure strongly controls the FWER at level α with respect to all hypotheses under consideration.

Theorem 2. Strong control of the FWER above remains true even when different selection rules for L^* (other than that which chooses the dose with maximum first-stage performance) and different thresholds for addition of modifications than c_2 are used, so long as c_1 and y_{cutoff} are unchanged.

The proofs of Theorems 1 and 2 are provided in the online Appendix. Results from simulation studies into the actual level of FWER control produced under selected adaption strategies are also summarized. The design and proofs of these two theorems can also be extended to other data types using standard arguments given, for instance, by Jennison and Turnbull.[7] In the situation of normal data with unknown variance, Student's t statistics are used. The null distributions of all needed statistics are well defined and can be approximated by simulation (after the rescaling of case (b), one could use the lower original degrees of freedom to be mildly conservative). In an asymptotic

situation, possibly with estimated nuisance parameters, the method based on the distribution of the score function under a local alternative sequence can be used. The locations of decision points are then based on estimated information, not total sample size. A similar result is presented with proof by Spivack et al.[3]

3 Phase II application based on the QALS trial

This section presents a fictitious example where interim results of a Phase II study suggest a change of plan with the introduction of a modification of the apparently best performing first-stage dose. The example is constructed from the QALS trial of Kaufmann et al.[4] investigating CoQ10 as a potential treatment for ALS. We use the same dose levels and sample sizes per group as the original study. Similar effect sizes and error rates are chosen, but for simplicity here, we present this example in terms of a conventional test for superiority rather than the test of "futility" (nonsuperiority) used in the original QALS trial. Versions in terms of "futility" or noninferiority hypotheses can be constructed along similar lines.

The outcome is a patient's change in ALSFRSr score over a nine-month period, an approximately normally distributed random variable with assumed standard deviation of 9 units and assumed mean of –9 units for the placebo arm. We note that according to the rating scale, the lower the score, the *worse* it is for the patient.

Let the initial design be as follows. Stage 1 consists of three doses, placebo (Dose 0), Dose 1 of 1800 mg of CoQ10, and Dose 2 of 2700 mg, administered as in Kaufmann et al.[4] Thirty-five patients will be randomized to each dose. At interim, mean changes in score on each arm will be computed, and the effect with respect to Placebo estimated.

In the notation of the design $\overline{X}_{0,1} \sim N\left(-9, \frac{81}{35}\right)$, $\overline{X}_{1,1} \sim N\left(\mu_1, \frac{81}{35}\right)$, $\overline{X}_{2,1} \sim N\left(\mu_2, \frac{81}{35}\right)$, $\overline{Y}_{1,1} \sim N\left(\delta_1 = \mu_1 + 9, \frac{162}{35}\right)$, and $\overline{Y}_{2,1} \sim N\left(\delta_2 = \mu_2 + 9, \frac{162}{35}\right)$. We use $\bar{Y}_{L^*,1}$ to denote the maximum observed effect in the first stage, corresponding to dose selection $L^* \in \{1,2\}$.

If no mean change in score exceeds that of the control by one unit, $\bar{Y}_{L^*,1} < 1$, the study will terminate by design. Otherwise, the arm with best observed first-stage performance will be selected for promotion into the second stage along with control. In the second stage, 40 patients will be randomized to each. The final analysis compares the difference in overall sample means between the selected treatment arm and control to the cutoff $y_{cutoff} = 2.127$, which is selected such that under the hypothesis of no effect relative to control on either dose, the experiment will have $\alpha = .1$, one sided, as in the original trial. Under the assumption, for instance, that the true effects of Doses 1 and 2 relative to control are 0 and 4.5 units, respectively, the design has power .9 to select and confirm the effect on Dose 2.

Suppose that the experiment unfolds as follows. The first-stage sample means of Doses 0, 1, and 2 are –9.96, –7.27, and –6.81, respectively, such that $\overline{Y}_{1,1} = 2.69$ and $\overline{Y}_{2,1} = 3.15$. Dose 2 is selected for study in the second stage. To both Dose 2 and Dose 0 are randomized an additional 40 patients. However, given that the signal observed on Dose 2 seems lukewarm and the tolerability of CoQ10 appears to be excellent, the investigators decide that they would also like to consider a modification of Dose 2 (Dose 2a) at the level of 3000 mg. Therefore, 75 additional patients are included in stage 2 on Dose 2a, such that the total evidence provided by the trial with respect to Dose 2a will be the same as that on the original treatment arms. Clinical consideration, based on the tolerability of CoQ10, the fact that Dose 2a is a close modification of Dose 2, and that it does not exceed any previously established MTD level allows the modification to be introduced under appropriate safety monitoring. The significance level for testing Dose 2a is found according to Theorem 1, and calculated, for instance, by Monte Carlo simulation, as $\alpha_1 = .037$.

Let the observed second-stage sample means on Dose 0, Dose 2, and Dose 2a be –8.04, –5.89, and –4.84, respectively, such that such that $\overline{Y}_{2,2} = 2.15$ and $\overline{Y}_{2a,2} = 3.20$. In the final analysis, we calculate $\bar{Y}_{L^*,\,overall} = 2.62$, which exceeds the original y_{cutoff}. An effect on Dose 2 is confirmed. Next, the effect on Dose 2a is tested at $\alpha_1 = .037$, one sided, using the second-stage data. The resulting two-sample Z-test statistic, 1.82, is significant.

Let us assume that the effect of Dose 2 is nominally significant, but additional considerations revealed by sensitivity analyses, such as those described in Kaufmann et al.,[4] led investigators to withhold full endorsement of Dose 2. However, the evidence in favor of the dose modification 2a in terms of its observed effect size is stronger; this allows a positive endorsement for CoQ10 (at Dose 2a) that would not have otherwise been achieved without exploring the dose modification.

4 Seamless Phase II/III application based on a Limb-Leaf design

Here, we illustrate the application of the procedure of Section 2 to a Phase II/III study. The Limb–Leaf approach of Spivack et al.[3] is intended to better explore the dose response in cases where it is considered likely to be

nonmonotonic. Candidate doses are prespecified and classified as limbs and leaves, with leaves interpreted as finer scale modifications of their limbs.

Here, we implement this approach using the more practical procedure of Section 2, rather than the somewhat opaque *p*-value combination rules of Spivack et al.[3] We call this new implementation the AdEx strategy.

In a Limb–Leaf approach, main doses $\{L_1, \ldots, L_K\}$ are termed limb doses, and for each such L_k, the small-scale dose modifications $\{l_{k,1}, \ldots, l_{k,m_k}\}$ are termed its associated leaf doses. These could correspond, for instance, to increments or decrements of dose level, changes in administration method, or the relative proportions of components in a combination therapy. A *locatable effect* exists relative to the vector $\Delta = (\Delta_1, \Delta_2, \Delta_3)$, with $\Delta_1 < \Delta_2 < \Delta_3$ and Δ_3 considered as the smallest desired level of effect, if the following two conditions hold:

(1) The effects on each limb, $\delta_{L_k}, k = 1, \ldots, K$, are either less than or equal to Δ_1, or greater than or equal to Δ_2, with at least one k such that $\delta_{L_k} \geq \Delta_2$.
(2) For any limb L_k such that $\delta_{L_k} \geq \Delta_2$, each of $\delta_{L_k}, \delta_{l_{k,1}}, \ldots, \delta_{l_{k,m_k}}$ is either greater than or equal to Δ_3 or less than or equal to Δ_2, with at least one of these effects greater than or equal to Δ_3.

To illustrate these definitions in the setting of the example presented in Section 3: if the underlying effects on Doses 1, 2, and 2a with respect to placebo were assumed to be 2.0, 3.0, and 3.5 units, respectively, the doses under investigation could be considered as $\{L_1 = \text{Dose 1}, L_2 = \text{Dose 2}, l_{2,1} = \text{Dose 2a}\}$. A locatable effect would then be assumed to exist with respect to the vector $\Delta = (2.0, 3.0, 3.5)$ on Dose 2a. If the assumed effects on Doses 1, 2, and 2a were instead 2.0, 3.5, and 3.0, a locatable effect would exist with respect to the same vector on Dose 2.

The hypotheses to be tested are, for each dose d, H_0: $\delta_d \leq 0$, where δ_d represents the effect of dose d relative to control, and FWER control at prespecified α is enforced. Spivack et al.[3] show that for any given dose response, a locatable effect exists with respect to some Δ, and characterize the Δ values over which such an AdEx strategy may be appropriate.

The experiment follows the procedure of Section 2, with the component before the interim analysis identified with a Phase II stage, the component after the interim analysis with a Phase III stage, and the overall study considered as an adaptive, seamless Phase II/III trial.

Power is enforced in the Limb–Leaf approach by means of unfavorable configurations. Specifically, given vector $\Delta = (\Delta_1, \Delta_2, \Delta_3)$ characterizing a locatable effect, we identify $\mathrm{U_{limb}}$ as a vector of effects satisfying the conditions: $\delta_{L_k} = \delta_{l_{k,1}} = \cdots = \delta_{l_{k,m_k}} = \Delta_1$ for $k \neq k^*$, $\delta_{L_{k^*}} = \Delta_3$, and $\delta_{l_{k^*,1}} = \cdots = \delta_{l_{k^*,m_{k^*}}} = \Delta_2$ for some k^*. We define $\mathrm{U_{leaf}}$ as the vector of effects satisfying $\delta_{L_k} = \delta_{l_{k,1}} = \cdots = \delta_{l_{k,m_k}} = \Delta_1$ for $k \neq k^*$, $\delta_{L_{k^*}} = \Delta_2$, $\delta_{l_{k^*,1}} = \Delta_3$, and $\delta_{l_{k^*,2}} = \cdots = \delta_{l_{k^*,m_{k^*}}} = \Delta_2$, for some k^*.

The optimization of design constants is carried out to minimize the RAESS, a form of Bayes risk. For a given set of limb and leaf doses and a specified $\Delta = (\Delta_1, \Delta_2, \Delta_3)$, the vector of constants n_{1L}, n_{2L}, n_{2La}, n_{2la}, c_1, and c_2 has an associated value $\mathrm{RAESS} = (1\text{-}p_{U_{limb}}\text{-}p_{U_{leaf}})\, E_{NULL}(N) + p_{U_{limb}}\, E_{U_{limb}}(N) + p_{U_{leaf}}\, E_{U_{leaf}}(N)$, where N is the total sample size, NULL represents the global null hypothesis, U_{limb} and U_{leaf} are the limb-effect and leaf-effect unfavorable configurations with respect to Δ, and $p_{U_{limb}}$ and $p_{U_{leaf}}$ are assumed prior probability values. Minimization over feasible parameters meeting the power constraints can be done by grid search or more advanced methods such as simulated annealing by Černý.[8]

For the purpose of comparison, we use a simple promote-the-winner rule based on that of Thall et al.[9] (TSE design). There are 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 $\delta_1, \ldots, \delta_I$. We assume that at either stage, the outcomes at a given dose d_j are independent and identically distributed as normal random variables with mean μ_j and variance $\sigma^2 = 1$, $j = 0, 1, \ldots, I$. 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, $T_{1,i} = \left(\bar{X}_{1,i} - \bar{X}_{1,0}\right)/\sqrt{2\sigma^2}$, $\bar{X}_{1,0}$ is the sample mean for control d_0, and $\bar{X}_{1,i}$ is the sample mean for dose $d_i, i = 1, \ldots, I$ at stage 1. If $T_1 > y_1$, then continue by selecting 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.

Table 1. Design constants for proposed AdEx designs.

$\Delta = (\Delta_1, \Delta_2, \Delta_3)$	n_{1L}	n_{2L}	n_{2La}	n_{2la}	c_1	c_2	y_{cutoff}	α_1	$\text{RAESS}_{\text{AdEx}}$
Two doses (one limb and one leaf per limb)									
(1/8,5/8,1)	27.32	42.88	46.38	48.69	0.21	1.95	0.31	0.025	105.98
(1/8,6/8,1)	32.19	27.22	36.79	63.78	0.24	1.56	0.34	0.025	108.67
(1/8,7/8,1)	47.39	123.15	156.50	242.65	0.36	1.56	0.08	0.025	221.42
(3/8,5/8,1)	29.66	81.20	40.35	48.29	0.21	1.80	0.23	0.025	105.78
(3/8,6/8,1)	30.12	87.13	34.41	76.24	0.24	1.73	0.22	0.025	108.32
(3/8,7/8,1)	44.40	136.72	199.75	205.23	0.36	1.75	0.10	0.025	230.11
Three doses (one limb and two leaves per limb)									
(1/8,5/8,1)	35.68	83.98	68.73	81.16	0.25	1.97	0.22	0.025	150.30
(1/8,6/8,1)	36.70	79.47	54.62	114.54	0.33	2.27	0.19	0.025	161.21
(1/8,7/8,1)	49.50	284.19	271.21	330.48	0.36	2.25	0.02	0.025	371.37
(3/8,5/8,1)	38.80	57.80	67.96	80.98	0.26	1.64	0.25	0.025	149.36
(3/8,6/8,1)	36.74	63.27	51.30	114.30	0.32	2.06	0.21	0.025	160.19
(3/8,7/8,1)	44.93	308.82	249.73	354.22	0.37	1.61	0.03	0.025	367.85
Four doses (two limbs and one leaf per limb)									
(1/8,5/8,1)	33.34	74.69	46.75	60.01	0.23	2.43	0.27	.010	163.40
(1/8,6/8,1)	31.99	63.20	40.02	91.84	0.29	3.39	0.28	.010	157.36
(1/8,7/8,1)	46.08	99.48	173.69	260.22	0.42	3.21	0.10	.010	276.13
(3/8,5/8,1)	61.67	64.32	47.38	71.25	0.23	2.22	0.25	.011	239.76
(3/8,6/8,1)	44.04	54.25	46.81	97.16	0.33	2.19	0.27	.011	185.93
(3/8,7/8,1)	50.42	106.93	204.50	225.42	0.40	1.73	0.10	.011	295.87
Six doses (two limbs and two leaves per limb)									
(1/8,5/8,1)	45.66	99.33	84.55	103.65	0.31	1.92	0.21	.011	245.50
(1/8,6/8,1)	37.53	83.47	67.84	117.61	0.35	2.04	0.22	.011	217.95
(1/8,7/8,1)	48.29	273.47	243.89	373.42	0.42	2.99	0.03	.011	422.85
(3/8,5/8,1)	71.57	85.72	94.34	105.46	0.25	1.73	0.21	.011	330.47
(3/8,6/8,1)	47.05	88.26	66.70	131.68	0.35	2.80	0.20	.011	243.29

RAESS: risk-adjusted expected sample size; AdEx: adaptive exploration.

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

$$T_2 = \frac{n_1}{n_1+n_2}\frac{\left(\bar{X}_{1,i}-\bar{X}_{1,0}\right)}{\sqrt{2\sigma^2}} + \frac{n_2}{n_1+n_2}\frac{\left(\bar{X}_{2,i}-\bar{X}_{2,0}\right)}{\sqrt{2\sigma^2}}$$

If $T_2 > y_2$, then reject $H_{0,i^*} : \delta_{i^*} \leq 0$ and conclude that $\delta_{i^*} > 0$. If $T_2 \leq y_2$, then do not reject H_{0,i^*}.

This 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 subject to overall (one sided) type 1 error rate α and maintaining a target power under an unfavorable dose-response configuration. Since the TSE design does not recognize a distinction between limbs and leaves, both U_{limb} and U_{leaf} may be expressed as U_{TSE}, which we set to be identical to U_{limb} for convenience.

Comparisons between the AdEx- and the TSE-based design are made in terms of RAESS for (a) one limb and (b) two limb and cases, with one or two leaves per limb, over ranges of the vector $\Delta = (\Delta_1, \Delta_2, \Delta_3)$. Results are presented in Tables 1 and 2. While RAESS is the only directly comparable measure of performance, the other design constants, especially those of the first- and second-stage sample sizes, are roughly analogous. We include them in order to provide additional insight into the qualitative behavior of each competitor.

In the comparisons, we enforce that our AdEx achieves 90% power to detect a locatable effect for given Δ by requiring power of 90% in both U_{limb} and U_{leaf} configurations, with overall (one sided) $\alpha = .025$. Power of 90% is required of the TSE design under U_{TSE} with overall (one sided) $\alpha = .025$. Design constants are optimized using $\text{RAESS} = (1\text{-}p_{U_{limb}}\text{-}p_{U_{leaf}})\ \ E_{NULL}(N) + p_{U_{limb}}\ \ E_{U_{limb}}(N) + p_{U_{leaf}}\ \ E_{U_{leaf}}(N)$, with the assumed values $p_{U_{limb}} = p_{U_{leaf}} = .1$; for the TSE design, this reduces to $\text{RAESS} = (1\text{-}p_{U_{TSE}})\ \ E_{NULL}(N) + p_{U_{TSE}}\ \ E_{U_{TSE}}(N)$,

Table 2. Design constants for competing promote-the-winner designs (TSE designs).

$\Delta = (\Delta_1, \Delta_2, \Delta_3)$	n_1	n_2	y_1	y_2	$\text{RAESS}_{\text{TSE}}$
Two doses (one limb and one leaf per limb)					
(1/8,5/8,1)	26.43	13.11	0.42	0.47	86.72
(1/8,6/8,1)	52.58	3.20	0.04	0.41	162.85
(1/8,7/8,1)	207.19	3.70	0.04	0.21	625.90
(3/8,5/8,1)	25.49	16.03	0.37	0.46	86.62
(3/8,6/8,1)	52.81	3.00	0.04	0.41	164.45
(3/8,7/8,1)	209.49	14.91	0.21	0.18	635.09
Three doses (one limb and two leaves per limb)					
(1/8,5/8,1)	37.03	4.06	0.42	0.43	164.02
(1/8,6/8,1)	79.97	4.26	0.25	0.31	322.47
(1/8,7/8,1)	317.12	7.58	0.06	0.16	1276.39
(3/8,5/8,1)	36.18	2.35	0.02	0.46	164.18
(3/8,6/8,1)	80.48	3.95	0.02	0.31	329.73
(3/8,7/8,1)	319.80	237.23	0.15	0.06	1396.61
Four doses (two limbs and one leaf per limb)					
(1/8,5/8,1)	24.96	19.30	0.34	0.42	141.44
(1/8,6/8,1)	53.00	2.88	0.03	0.40	269.55
(1/8,7/8,1)	209.48	116.67	0.20	0.10	1105.08
(3/8,5/8,1)	28.18	10.65	0.42	0.45	147.86
(3/8,6/8,1)	52.90	3.44	0.37	0.39	266.38
(3/8,7/8,1)	207.99	39.95	0.16	0.18	1064.00
Six doses (two limbs and two leaves per limb)					
(1/8,5/8,1)	36.96	7.12	0.30	0.48	265.20
(1/8,6/8,1)	79.51	3.11	0.34	0.35	558.17
(1/8,7/8,1)	317.33	371.08	0.16	0.07	2412.15
(3/8,5/8,1)	37.59	3.34	0.33	0.50	265.84
(3/8,6/8,1)	78.94	6.32	0.18	0.35	558.82

RAESS: risk-adjusted expected sample size; TSE: Thall et al.[9]

with $p_{U_{TSE}} = .2$. The RAESS of the optimized AdEx design for each comparison is denoted in Table 1 by $\text{RAESS}_{\text{AdEx}}$, and that of the optimized TSE design is denoted in Table 2 by $\text{RAESS}_{\text{TSE}}$.

We note that the one-limb cases have an appealing form. At interim, the experimenter has the option to add leaves for further testing in simple step-down sequence at (one sided) $\alpha_1 = .025$. The parameter Δ_1 is unused, since no other limb is allowed, such that results depend only on Δ_2 and Δ_3. Once value Δ_2 reaches 75% of the desired effect, corresponding to an adequate but not especially accurate choice for the location of the initial limb, the RAESS of the AdEx design is 33% better than that of the TSE design. Once the initial limb selection improves to 87.5% of the desired effect, the improvement in RAESS exceeds 60%. However, that relative performance degrades at lower values of Δ_2, corresponding to poorer selections of the limb dose and poorer ability of the AdEx design to identify the region of promising activity for further exploration.

Similar patterns are seen in the two limb cases. With Δ_2 equal to 75% or more of the desired effect, the AdEx design has a major advantage in terms of RAESS. Additional simulations show that this advantage continues to grow as Δ_2 approaches Δ_3. Since there is now more than one limb under consideration, the value of Δ_1 affects results in a foreseeable way: increasing Δ_1 makes the region of desired activity harder to discern and degrades the performance of the AdEx design. Additional simulations show that as Δ_1 approaches Δ_2, corresponding to an increasing difficulty in identifying the region of desired activity, the required sample size of the AdEx design increases without bound.

From these and other simulation studies, we conclude that the AdEx approach may be advantageous in many situations. If the assumed Δ vector allows sufficient discrimination of the neighborhood containing the desired effect, specified for instance by $\Delta_2 - \Delta_1 \geq .25\ \Delta_3$, and sufficient discrimination within the correct neighborhood, specified for instance by $\Delta_3 - \Delta_2 \geq .125\ \Delta_3$, then over reasonable choices for the numbers of limbs and leaves, the AdEx design is superior to a classic promote-the-winner design. Furthermore, the increased efficiency of the AdEx approach may allow exploration of certain problems for which a more standard design would be infeasible due to impractically large costs.

We may heuristically expect the previous conclusions favorable to the AdEx design to hold against other comparators, such as those using modeling strategies to guide dose selection or those organized as sequences of independent studies with one study for dose finding and a second for confirmation of the selected dose. Specifically, the demands of correct selection in a design that does not make specific use of the Limb–Leaf structure for stagewise exploration require a large first-stage sample size in order to identify the correct candidate dose and distinguish it from competitors that are nearby in magnitude and dose level. There is little additional information to be gained in making such a selection from the observed effects of other doses far away from the region of promising activity—given these observed effects elsewhere, all the various possibilities within the region of promising activity still remain almost equally plausible. Thus, we expect gains from more sophisticated two-stage schemes to be modest. These limitations also pertain to sequential schemes, with the additional consideration that the second study's inference would not be allowed to borrow strength from the first stage.

It might also be of interest to match the method proposed here against other possible AdEx strategies. Detailed study is outside the scope of the present paper; however, the simulation tables presented in Spivack et al.[3] indicate that performance of the AdEx method is qualitatively comparable to that of the previous Limb–Leaf method, while its analysis is far simpler as it does not rely on complex formulas for combination of stagewise p-values. It also appears not to suffer a corresponding loss of power in certain pathological cases where p-value combination rules behave especially poorly. One such scenario detrimental to the original Limb–Leaf implementation, but not to the AdEx strategy, occurs in attempting to confirm an effective leaf when the performance relative to control of an unrelated limb is large and negative.

Further simulations show good robustness of this AdEx design to perturbations in the components of Δ from its assumed value and to relaxing the condition that the order of leaves be correctly chosen a priori. This is not an issue for one-leaf cases. However, in the two-leaf cases above, we modified U_{leaf} to reflect an incorrect ordering by switching the effects on the first and second leaves to create U'_{leaf} satisfying: $\delta_{L_k} = \delta_{l_{k,1}} = \cdots = \delta_{l_{km_k}} = \Delta_1$ for $k \neq k^*$, $\delta_{L_{k^*}} = \Delta_2$, $\delta_{l_{k,1}} = \Delta_2$, and $\delta_{l_{k^*,2}} = \Delta_3$, for some k^*. The impact on power from this change was negligible (data not shown), which we attribute to the fact that selection of the correct leaf using the observed effects seems to be the determining factor for power.

To summarize, the use of a prespecified AdEx strategy, based on the procedure introduced in Section 2, seems to be an option worth considering for a Phase II/III trial where an automatic recommendation of the MTD is not anticipated, and it is assumed that the dose response may require further exploration. The potential for improvements in efficiency using such a design may make it possible to attack certain problems whose costs would otherwise be prohibitive.

5 Discussion

The option we present for including dose modifications into a standard selection and promotion design, whether motivated by emerging results or as part of a prespecified AdEx strategy, is convenient to implement. Further, it does not entail major alterations to the analyses with which practitioners are familiar. An option for AdEx is not mandatory: if it is not utilized, the design reverts to the standard form of selection followed by promotion with no changes to cutoff values or other multiplicity penalty.

We expect the added flexibility to appeal to researchers and to be a convenient addition to many study protocols. Furthermore, the improved efficiency of AdEx enabled by this method in certain situations may allow attack on problems that may be prohibitively costly otherwise.

Several concerns need to be addressed, however, when implementing the addition of a dose modification. It would be advisable as far as possible to prespecify the potential dose modifications and the intended adaptation rule. This would also allow the most complete prior discussion of their safety concerns, cost effectiveness, and potential medical value. Nonetheless, it is statistically valid in terms of FWER control for the modifications to be less than fully specified until their actual use in the second stage. In practice, this could be essential: if, for instance, a modification were prespecified but information external to the trial led to a reconsideration of its toxicity, a decision to lower such a dose, alter its administration, or make other changes for reasons of safety might be unavoidable.

It is necessary that the trial accumulate sufficient patient experience on any added dose such that the overall conclusions with respect to that dose would be regarded as clinically meaningful. If the size of the arm added to study a dose modification is too small, not only may statistical power suffer, but even a statistically significant conclusion of efficacy may not be seen as clinically convincing.

It is also important from the point of view of oversight that the notion of a dose modification, as a close relative of its underlying main dose, not be abused. The MTD may not be exceeded by any newly introduced modification. Increased safety monitoring may be a prudent addition to the study protocol when the MTD is approached or there are other grounds for concern. Ethical conduct, clinical judgment, and good oversight are obviously necessary.

In practice, we expect from one to three main doses, and two or fewer modifications per main dose to be used in applications. Throughout our simulations, the significance level α_1 applied to modifications was greater than the significance level associated with a Bonferroni correction, which varies as $\frac{\alpha}{K+1}$. Nonetheless, if the number of main doses exceeds five, performance would be expected to suffer.

The use of a step-down sequence for testing the dose modifications is defensible. However, any procedure that uses FWER control at level α_1 among the modifications is also valid. In a setting where there was little basis for an a priori ordering of the modifications, a different method such as a Bonferroni or Holm step-down correction at overall level α_1 could clearly be used.

Future work is motivated. Group sequential continuation is obviously valid and can be included with small additional effort. Further clarification of the appropriate context for use of dose modifications is important. The roles of preliminary clinical and pre-clinical studies, pharmacokinetics and pharmacodynamics, and expert judgment in their choice and their use deserve further study. Existing methods for point and interval estimation such as those based on bias-adjusted estimators,[10] test inversion, and median-unbiased point estimators are applicable,[7,11–13] but the specific implementations of these methods and comparisons of their performances are worth consideration.

Acknowledgments

We wish to thank Professor Brian Everitt and two anonymous referees for thoughtful comments and suggestions, which improved the manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

John Spivack https://orcid.org/0000-0002-5347-1696
Bin Cheng https://orcid.org/0000-0001-8295-2467

Supplementary material

Supplementary material is available for this article online.

References

1. Cohen DR, Todd S, Gregor WM, et al. Adding a treatment arm to an ongoing clinical trial: A review of methodology and practice. *Trials* 2015; **16**: 179.
2. Chang M and Wang J. The add-arm design for unimodal response curve with unknown mode. *J Biopharm Stat* 2015; **25**: 1039–1064.
3. Spivack J, Cheng B and Levin B. Limb-leaf designs with adaptive exploration of the dose response curve. *Contemp Clin Trials* 2018; **64**: 210–218.
4. Kaufmann P, et al. Phase II trial of CoQ10 for ALS finds insufficient evidence to justify phase III. *Ann Neurol* 2009; **66**(2): 235–244.
5. Responding to neglected patients' needs through innovation. Neglected diseases initiative, https://www.dndi.org/wp-content/uploads/2018/08/DNDi_AR_2017.pdf (2017, accessed 14 December 2018).
6. Genz A and Bretz F. *Computation of multivariate normal and T probabilities*. *Lecture notes in statistics*. Vol. 195. Heidelberg, Germany: Springer-Verlag, 2009.

7. Jennison C and Turnbull B. *Group sequential methods with applications to clinical trials*. New York, NY: CRC Press, 2000.
8. Černý V. A thermodynamical approach to the travelling salesman problem: An efficient simulation algorithm. *J Optim Theory Appl* 1985; **45**: 41–51.
9. Thall PF, Simon R and Ellenberg SS. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; **75**: 303–310.
10. Stallard N and Todd S. Point estimates and confidence regions for sequential trials involving selection. *J Stat Plan Inference* 2005; **135**: 402–419.
11. Bretz F, Schmidli H, König F, et al. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biom J* 2006; **48**: 623–634.
12. Jennison C and Turnbull B. Adaptive seamless designs: Selection and prospective testing of hypotheses. *J Biopharm Stat* 2007; **17**: 1135–1161.
13. Liu Q, Proschan M and Pledger G. A unifed theory of two-stage adaptive designs. *J Am Stat Assoc* 2002; **97**: 1034–1041.