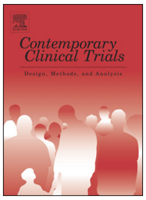
Contents lists available at ScienceDirect

Contemporary Clinical Trials

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



Adaptive choice of patient subgroup for comparing two treatments

Tze Leung Lai [a], Philip W. Lavori [b], Olivia Yueh-Wen Liao [c,*]

[a] Department of Statistics, Stanford University, Stanford, CA, USA
[b] Department of Health Research Policy, Stanford University, Stanford, CA, USA
[c] Onyx Pharmaceuticals, South San Francisco, CA, USA

ARTICLE INFO

Article history:
Received 24 June 2014
Received in revised form 30 August 2014
Accepted 1 September 2014
Available online 7 September 2014

Keywords:
Adaptive selection
Generalized likelihood ratio statistics
Group sequential design
Kullback–Leibler information
Multiple testing
Normalized Wilcoxon statistic

ABSTRACT

This paper is motivated by a randomized controlled trial to compare an endovascular procedure with conventional medical treatment for stroke patients, in which the endovascular procedure may be effective only in a subgroup of patients. Since the subgroup is not known at the design stage but can be learned statistically from the data collected during the course of the trial, we develop a novel group sequential design that incorporates adaptive choice of the patient subgroup among several possibilities which include the entire patient population as a choice. We define the type I and type II errors of a test in this design and show how a prescribed type I error can be maintained by using the closed testing principle in multiple testing. We also show how asymptotically optimal tests can be constructed by using generalized likelihood ratio statistics for parametric problems and analogous standardized or Studentized statistics for nonparametric tests such as Wilcoxon's rank sum test commonly used for treatment comparison in stroke patients.

1. Introduction

It is widely recognized that the comparative efficacy of a new treatment can depend on certain characteristics of the patients that are difficult to pre-specify at the design stage. In a randomized trial, ignoring characteristics that account for patient heterogeneity in response may yield a false negative result. On the other hand, narrowly defining the patient characteristics for inclusion and exclusion limits the proven usefulness of the treatment to a small patient subpopulation. A trial may also encounter difficulties in patient accrual when relatively few patients satisfy the stringent inclusion/exclusion criteria. Adaptive (data-dependent) choice of the patient subgroup to compare the new and control treatments is a natural compromise between ignoring patient heterogeneity and using stringent inclusion/exclusion criteria in the trial design and analysis.

In this paper we introduce some methods for adaptively choosing the patient subgroup and testing the efficacy of the new treatment on the chosen subgroup, with prescribed type I error probability for the data-dependent choice. Our work was motivated by a clinical trial suggested by our Stanford Neurology colleagues. The trial aims to compare standard medical therapy with the endovascular procedure to remove the clot in ischemic stroke. There are two baseline characteristics that define the subgroups of interest: One of them, an MRI-based measure referred to as DWI (diffusion weighted imaging) measures the size of the central infarct that is not considered salvageable, relative to the surrounding penumbra that has been affected by the loss of perfusion but is considered salvageable tissue if the area can be reperfused by recanalization of the artery that is blocked by the clot. The other is time since stroke in hours, which may moderate treatment effects by the way that the salvageable tissue is gradually degraded by chronic lack of perfusion. The clinical outcome is change from

[*] Corresponding author at: Onyx Pharmaceuticals, 249 E. Grand Ave., South San Francisco, CA 94080, USA. Tel.: +1 6502667857.
E-mail address: oliao@onyx.com (O.Y.-W. Liao).

http://dx.doi.org/10.1016/j.cct.2014.09.001
1551-7144/© 2014 Published by Elsevier Inc.

baseline, within three months, of an ordered categorical score (Rankin) measuring the change in impairment, that ranges from full recovery to complete disability or death. The clinical hypothesis is that in some subgroup of the patient population, defined by a region in the DWI × Time space, the null hypothesis of no treatment difference in the 3-month modified Rankin score is false. The investigators hope to reject the null for the largest possible subgroup for which it is false, although they also recognize that the difference may not be significant over the entire patient population.

Further details of the adaptive choice of the treatment subgroup are given in Section 3. The test statistic commonly used for two-sample comparison of the Rankin scores is Wilcoxon's rank sum. To avoid obscuring the main ideas underlying our methodology by technical arguments involving nonparametric statistics, we first consider in Section 2 the parametric case of normally distributed treatment outcomes with common known variance and mean μ_j (or μ_{0j}) for patient subgroup j and the new (or control) treatment. We develop the basic theory for adaptive choice of patient subgroup to compare the two treatments in this setting, first for designs with fixed sample size and then for a three-stage group sequential design that is used to approximate an adaptive design with mid-course determination of the patient subgroup and sample size re-estimation. Section 4 gives further discussion, extensions and concluding remarks.

2. Basic theory in prototypical normal setting

2.1. Asymptotic theory via Kullback–Leibler information and prevalence

Suppose n patients are randomized to the new and control treatments and the responses are normally distributed, with mean μ_j for the new treatment and μ_{0j} for the control treatment if the patient falls in pre-defined subgroups Π_j for $j = 1, \ldots, J$. We assume that the responses have common known variance σ^2. Extensions to unknown and possibly unequal variances will be discussed in Section 4. Let Π_J denote the entire patient population on which a traditional randomized controlled trial (RCT) comparing the two treatments focuses. The Kullback–Leibler (KL) information number for Π_J is $\mathcal{I}_J = \frac{n}{4}\left(\mu_J - \mu_{0J}\right)_+^2/\sigma^2$, where x_+ denotes max(x, 0), noting that the comparison involves testing the one-sided null hypothesis $\mu_J \le \mu_{0J}$. The KL information number, or relative entropy, quantifies the amount of information in the sample to distinguish the treatment mean μ_J from the control mean μ_{0J}, and plays an important role in the asymptotic theory of efficient parametric tests; see [1,2]. In particular, if $\mathcal{I}_J$ is not large enough, then the RCT may not have sufficient power, i.e., probability of showing a significant mean difference between the treatment and the control.

Let p_j be the prevalence of patient subgroup Π_j. Therefore np_j is the expected number of subjects in Π_j for a trial with a total sample size n that randomizes patients to the two treatments. The KL information number for Π_j is $\mathcal{I}_j = \frac{n}{4}p_j \left(\mu_j - \mu_{0j}\right)_+^2/\sigma^2$, which is the product of the prevalence p_j, the KL information $(\mu_j - \mu_{0j})_+^2/(2\sigma^2)$ from a pair of patients in Π_j receiving the new treatment and control, respectively, and the expected number $n/2$ of such pairs. Not only does this show the trade-off between the prevalence of the patient subgroup and the magnitude of the difference $\mu_j - \mu_{0j}$ in choosing the patient subgroup to compare the two treatments, but it also suggests that an asymptotically optimal choice of subgroup is the maximizer of $\mathcal{I}_j$ over $1 \le j \le J$. We can alternatively regard a sample size of m from Π_j as an effective sample size of m/p_j from the entire population.

2.2. An efficient approach for fixed sample size trials

Since there is typically little information from previous studies about the subgroup effect size $\mu_j - \mu_{0j}$ for $j \ne J$, we begin with a standard RCT to compare the new treatment with the control over the entire population. Determination of the sample size n is based on power at an assumed effect size $\mu_J - \mu_{0J}$ for testing $H_J : \mu_J \le \mu_{0J}$, recalling that μ_J and μ_{0J} are the mean responses of the new and control treatments, respectively, over the entire patient population. The main innovation of the proposed trial design is that it allows adaptive choice of the patient subgroup $\hat{I}$ in the event H_J is not rejected, to continue testing $H_i : \mu_i \le \mu_{0i}$ with $i = \hat{I}$, and can claim the new treatment to be better than control for the patient subgroup $\hat{I}$ if $H_{\hat{I}}$ is rejected. We can interpret $\hat{I}$ as an estimate of $I = \text{argmax}_{i \ne J}\mathcal{I}_i$. Letting $\theta_j = \mu_j - \mu_{0j}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$, the probability of a false claim is the type I error

$$\alpha(\boldsymbol{\theta}) = \begin{cases} P_{\boldsymbol{\theta}}\left(\text{Reject } H_J\right) + P_{\boldsymbol{\theta}}\left(\theta_{\hat{I}} \le 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}\right) & \text{if } \theta_J \le 0 \\ P_{\boldsymbol{\theta}}\left(\theta_{\hat{I}} \le 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}\right) & \text{if } \theta_J > 0, \end{cases} \tag{1}$$

for $\boldsymbol{\theta} \in \Theta_0$, where Θ_0 consists of all "null" parameter vectors $\boldsymbol{\theta}$ such that $\theta_j \le 0$ for some $j \le J$. Note that there is no false claim if H_J is rejected when $\theta_J > 0$, because in that case we would not go on to test a subpopulation. We want to maintain a type I error constraint $\alpha(\boldsymbol{\theta}) \le \alpha$, and will describe a procedure that uses the closed testing principle in the theory of multiple testing [3,4] to satisfy this constraint.

We also want the procedure to be efficient in the sense of maximizing the power at $\boldsymbol{\theta} \notin \Theta_0$. Since the null hypothesis is highly composite, a uniformly most powerful level-α test is not expected to exist. Instead we try to attain asymptotic efficiency as $n \to \infty$. The KL information numbers in the preceding subsection fit nicely with the concept of Bahadur efficiency [5]. In Appendix A we give a precise statement of the asymptotic efficiency result of the proposed procedure together with its derivation. Note that the complement of Θ_0 is $\{\boldsymbol{\theta} : \theta_j > 0 \text{ for all } j \le J\}$ and therefore the type II error at $\boldsymbol{\theta} \notin \Theta_0$ is

$$\beta(\boldsymbol{\theta}) = \sum_{i=1}^{J-1} P_{\boldsymbol{\theta}}\left(\hat{I} = i, \text{ accept } H_J \text{ and } H_i\right). \tag{2}$$

which is the probability of falsely ending up with no positive claim for the new treatment. We focus here on the description of the procedure and its implementation to maintain a prescribed type I error constraint on level α.

The trial randomly assigns n patients to the experimental treatment and the control. We reject H_J if

$$\frac{n_i n_{0i}}{n_i + n_{0i}} (\hat{\mu}_i - \hat{\mu}_{0i})_+^2 / \sigma^2 \geq c_\alpha \tag{3}$$

for $i = J$, where $\hat{\mu}_i$ ($\hat{\mu}_{0i}$) is the mean response of patients in Π_i from the treatment (control) arm and $n_i(n_{0i})$ is the corresponding sample size. Otherwise we choose the patient subgroup $\hat{I} \neq J$ with the largest value of the generalized likelihood ratio statistic GLR$_i$, which is the left-hand side of Eq. (3), among all subgroups $i = 1, \ldots, J-1$, and reject $H_{\hat{I}}$ if GLR$_{\hat{I}} \geq c_\alpha$. The threshold c_α is chosen such that $\alpha(\boldsymbol{\theta}) \leq \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$, and we next describe its computation.

It will be shown in Appendix A that

$$\sup_{\boldsymbol{\theta} \in \Theta_0} \alpha(\boldsymbol{\theta}) = \alpha(\mathbf{0}), \tag{4}$$

by making use of the closed testing principle [3,4]. To compute $\alpha(\mathbf{0})$, we first use the approximations $n_i \approx n_{0i}$ (since study subjects are equally likely to receive the new treatment or control) and $n_i + n_{0i} \approx np_i$, thereby approximating the random variables n_i and n_{0i} by the constant $np_i/2$ and $n_i n_{0i}/(n_i + n_{0i})$ by $np_i/4$. The error probability $\alpha(\mathbf{0})$ can then be computed as a sum of integrands, over certain sets, of the multivariate normal density of $(Z_1, \ldots, Z_J)$ under $\boldsymbol{\theta} = \mathbf{0}$, where $Z_j = \sqrt{np_i}(\hat{\mu}_i - \hat{\mu}_{0i})/2\sigma$. The covariance matrix of this multivariate normal distribution is particularly simple in the case $\Pi_1 \subset \cdots \subset \Pi_J$, which we assume in the sequel because of the motivating clinical trial described in the second paragraph of Section 1. In this nested case, $\mathrm{Cov}\left(Z_i, Z_j\right) \approx \sqrt{p_i/p_j}$ for $i \leq j$. Therefore, in this case the threshold c_α can be determined by solving the equation

$$\alpha = \alpha(\mathbf{0}) = P_\mathbf{0}\left(Z_J \geq c_\alpha\right) + \sum_{i=1}^{J-1} \int_{c_\alpha}^{\infty} P_\mathbf{0}\left(Z_J < c_\alpha,\ Z_j < x \text{ for } j \notin \{i, J\} | Z_i = x\right) \phi_i(x) dx, \tag{5}$$

where $\phi_i(x)$ is the density function of the standard normal Z_i.

2.3. A group sequential design with adaptive patient subgroup selection

The effect size $\mu_J - \mu_{0J}$ underlying the sample size calculation in a RCT is typically based on some related studies and also on constraints on funding and study duration, which leads to the notion of "implied alternative" in [6, p. 81]. The observed effect size may differ substantially from the assumed effect size during the course of the trial. This has led to adaptive designs with mid-course sample size re-estimation; see Chapter 8 of [7]. Since our design also allows treatment comparisons over smaller patient subgroups and there is usually little information at the beginning of the trial from previous studies about the effect sizes in the subgroups, an adaptive design that can both choose the patient subgroup and re-estimate the sample size is particularly attractive.

Bartroff and Lai [8,9] have developed a theory of efficient adaptive design for sample size re-estimation that involves 3-stage GLR tests. At the first interim analysis, the sample size for the second stage is estimated. If the GLR test rejects the null hypothesis or stops early for futility at the second interim analysis, the trial stops. Otherwise the trial continues to the third stage which corresponds to the maximum sample size of the trial. These adaptive designs can be approximated by standard group sequential designs that do not estimate the sample size for the second stage; see [8,10,11]. We next extend these approximations of the adaptive design to a 3-stage group sequential design in which the last stage corresponds to the maximum sample size and the sample size up to the second stage is near the mid-point of the first-stage and final sample sizes.

As in the preceding section for fixed sample size designs, the maximum sample size is the fixed sample size for testing H_J, or some inflation thereof, determined by the power at some effect size δ for the entire population. As pointed out in [6], this maximum sample size has order $8\delta^2|\log\alpha|/\sigma^2$. At an interim analysis, we first test H_J and can then discontinue testing H_J early for efficacy or futility. If early stopping for efficacy occurs, we terminate the trial and claim that the new treatment is better than the control on average over the entire population. If stopping occurs for futility of testing H_J, then we accept H_J and continue the trial with the most promising patient subgroup, that is, the subgroup $i \neq J$ that maximizes GLR$_i$ defined by Eq. (3), but with n_i and n_{0i} replaced by the corresponding sample sizes at the time of interim analysis. If the test for H_J does not stop, then we continue to the next stage of the 3-stage design and repeat this procedure.

To be more specific, similarly to the square root of the left-hand side of Eq. (3), let

$$Z_i^\ell = \sqrt{\frac{n_i^\ell n_{0i}^\ell}{n_i^\ell + n_{0i}^\ell}} \left(\hat{\mu}_i^\ell - \hat{\mu}_{0i}^\ell\right) / \sigma \tag{6}$$

denote the test statistic for patient subgroup Π_i at the ℓth interim analysis. At this analysis, the trial rejects H_J, terminates, and claims efficacy of the new treatment if

$$Z_J^\ell \geq b. \tag{7}$$

It accepts H_J and proceeds to the subgroup selection if

$$\widetilde{Z}_J^\ell \leq \widetilde{b} \tag{8}$$

where $\widetilde{Z}_i^\ell$ is the test statistic of the alternative $K_i : \mu_i \geq \mu_{0i} + \delta$, that is,

$$\widetilde{Z}_i^\ell = \sqrt{\frac{n_i^\ell n_{0i}^\ell}{n_i^\ell + n_{0i}^\ell}} \left(\hat{\mu}_i^\ell - \hat{\mu}_{0i}^\ell - \delta\right) / \sigma. \tag{9}$$

Once H_J is accepted at stage ℓ and a subgroup $\hat{I}$, with the largest value of Z_i^ℓ for $i \neq J$, is chosen, the future enrollment of the trial will include patients of this subgroup only, while the maximum total sample size N remains the same. Similar to testing H_J, the trial rejects $H_{\hat{I}}$ at stage ℓ' and terminates with an efficacy claim of the new treatment in this subgroup if

$$Z_{\hat{I}}^{\ell'} \geq b. \tag{10}$$

It may also terminate for futility if

$$\widetilde{Z}_{\hat{I}}^{\ell'} \le \widetilde{b}. \tag{11}$$

If neither Eq. (7) nor Eq. (8) ever occurs, the trial proceeds to the final stage, in which case H_J is rejected if

$$Z_J^3 \ge c \tag{12}$$

and is accepted otherwise. Fig. 1 gives a schematic summary of the sequence of decisions in this group sequential trial.

Note that in the equations above, $\hat{I}$ is a random variable with values ranging from 1 to $J-1$. Moreover, the upper bound $\alpha(\mathbf{0})$ of the type I error, which will be explained in Appendix A, of this group sequential design can be decomposed into two parts similar to Eq. (1). The first part, $P_{\mathbf{0}}(\text{Reject } H_J)$, is bounded by

$$P_{\mathbf{0}}\left(Z_J^\ell \ge b \text{ for } \ell = 1 \text{ or } 2\right) + P_{\mathbf{0}}\left(Z_J^\ell < b \text{ for } \ell = 1, 2, \text{ and } Z_J^3 \ge c\right) \equiv P_{1b} + P_{1c}.$$

The second part, $P_{\mathbf{0}}(\text{Accept } H_J \text{ and reject } H_{\hat{I}})$, is bounded by

$$P_{\mathbf{0}}\left(\widetilde{Z}_J^\ell \le \widetilde{b} \text{ and } Z_{\hat{I}}^{\ell'} \ge b \text{ for some } \ell \le \ell' < 3\right) + P_{\mathbf{0}}\left(\widetilde{Z}_J^\ell \le \widetilde{b},\ Z_{\hat{I}}^{\ell'} < b \text{ for } \ell \le \ell' < 3, \text{ and } Z_{\hat{I}}^3 \ge c\right) \equiv P_{2b} + P_{2c}.$$

As shown in Appendix A, $\mathrm{Cov}\left(Z_J^\ell, Z_J^{\ell'}\right) \approx \sqrt{N_\ell / N_{\ell'}}$ for $\ell \le \ell'$, where N_ℓ is the total number of patients enrolled up to the ℓth interim analysis. Hence, the probabilities P_{1b} and P_{1c} can be computed by the recursive numerical integration [7, p. 86] using the joint normal density of $\{Z_J^\ell, 1 \le \ell \le 3\}$.

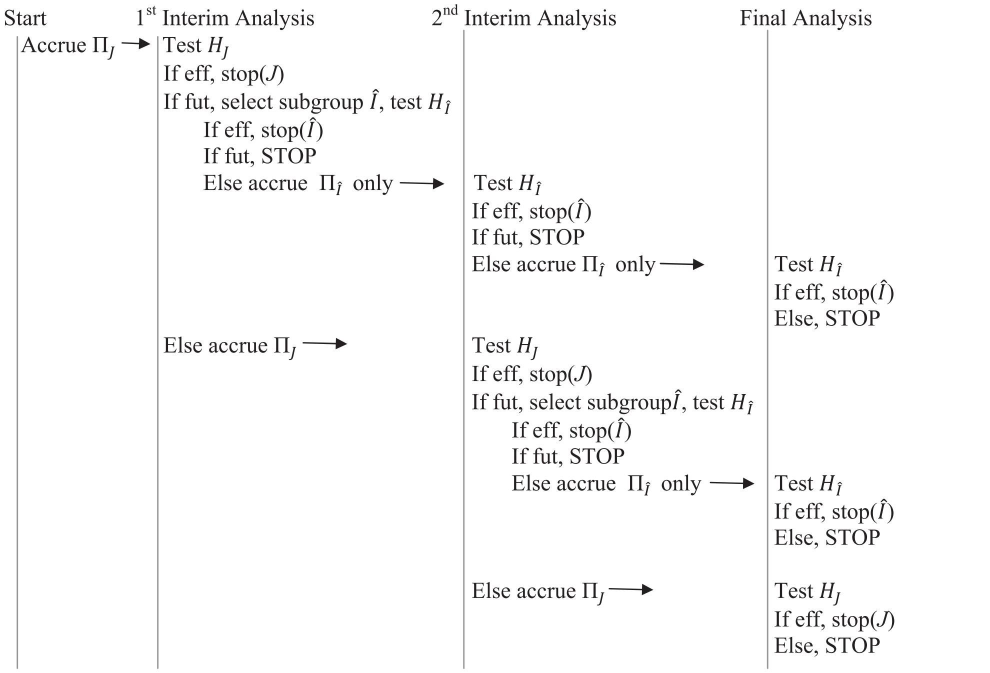
Recursive numerical integration can also be used to compute the probability

$$P_{2b} = \sum_{\ell=1}^{2} \int_{-\infty}^{b} \sum_{i=1}^{J-1} P_{\mathbf{0}}\left(\widetilde{Z}_J^\ell \le \widetilde{b},\ Z_j^\ell < x \text{ for } j \notin \{i, J\} | Z_i^\ell = x\right) \times P_{\mathbf{0}}\left(Z_i^{\ell'} \ge b \text{ for some } \ell' \ge \ell | Z_i^\ell = x\right) \phi_i(x) dx$$

(and similarly for P_{2c}). We use here the joint normal distribution of $\left\{Z_i^{\ell'},\ \ell \le \ell' \le 3\right\}$ conditional on Z_i^ℓ, with ℓ representing the stage when H_J is accepted, and the joint normal distribution of $\widetilde{Z}_J^\ell \approx Z_J^\ell - \sqrt{N_\ell}\delta/(2\sigma)$ and $Z_j^\ell, j \notin \{i, J\}$, conditional on Z_i^ℓ with $i \ne J$. The thresholds $b,\ \widetilde{b} < 0$ and c are determined by solving the equations

$$P_{\theta_J=\delta}\left(\widetilde{Z}_J^\ell \le \widetilde{b} \text{ for } \ell = 1 \text{ or } 2\right) = \epsilon\beta, \tag{13}$$
$$P_{1b} + P_{2b} = \epsilon\alpha,\quad P_{1c} + P_{2c} = (1-\epsilon)\alpha,$$

for the prescribed maximum type I error α, power $1-\beta$ at the alternative δ, and the proportion $0 < \epsilon < 1$ of type I (or type II) error spent at interim analyses.



Π_J=entire population; $\Pi_{\hat{I}}$ =selected subgroup; eff=efficacy boundary crossed; fut=futility boundary crossed; stop($\cdot$) = stop trial for efficacy (in the tested group - J or $\hat{I}$); STOP = stop trial and claim futility

Fig. 1. Flow chart of the 3-stage group sequential design. Events that line up vertically occur at the same analysis.

3. Adaptive design of a randomized controlled trial to test endovascular therapy

3.1. Endovascular versus standard therapy following imaging evaluation for ischemic stroke

The clinical trial mentioned in the second paragraph of Section 1 is a randomized trial for patients experiencing acute ischemic stroke (IS) due to a proximal anterior circulation large vessel occlusion who are treated within 12 h of IS onset. Patients will undergo MRI studies after enrollment but before randomization. The only FDA approved drug for the treatment of acute IS is IV-tPA, yet only 2–5% of IS patients receive it since most patients are outside the treatment window of the drug. The standard medical treatment to be compared with endovascular therapy may include IV-tPA administered up to 4.5 h after symptom onset. The study involves multiple centers. At each center, a multi-disciplinary team of stroke neurologists and endovascular therapists who treat patients with acute IS will be recruited as investigators. Patients will be screened according to inclusion and exclusion criteria in the protocol, and an informed consent form will be submitted to the local IRB for each center before enrollment begins at the center. However, in certain cases the acute stroke patient may not be asked to give consent. If neurologists of the stroke care team find that the patient's condition precludes them from giving informed consent, it will be obtained from the patient's legally recognized representative as governed by local laws and stipulated by the local IRB. Patients will be monitored for symptomatic intracranial hemorrhage and other endovascular complications during hospital admission. Clinical follow-up of the patient will be performed at 30 and 90 days and will include the modified Rankin score and standard Case Report Forms. A Data and Safety Monitoring Board (DSMB) will be appointed for the study, comprising of representatives from multiple disciplines including neurology and biostatistics, and independent of the investigational sites.

The maximum sample size calculation is based on a non-adaptive trial to detect projected difference between endovascular treatment and standard medical management for these patients, using Wilcoxon's rank sum of the Rankin scores as the test statistic to test the null hypothesis $P(Y \le X) \le 1/2$ of no improvement in the Rankin score X by the endovascular treatment compared to the score Y from the standard medical treatment arm in the entire population. We follow the approach in [12, p. 419] to use a working model for a fixed-sample-size nonparametric two-sample test at the design stage of an adaptive or group sequential trial, in which the actual pattern of the treatment effects can be adapted to during the course of the trial. As pointed out in [12], the basic idea is to have power at the working model specifying the alternative close to that of the fixed-sample-size test that is asymptotically most powerful at the specified alternative. It is shown in [12] that the design also has good power properties at other alternatives and can also reduce the expected sample size under the type I error probability and maximum sample size constraints. Using this approach, the maximum sample size is planned to be $N = 500$. The pre-specified treatment effect size, significance level, and power used for this calculation will be described in the second paragraph of Subsection 3.2. Interim analysis is performed with the results of approximately 300 and 400 patients randomized to the endovascular and standard treatments. As described in Section 1, the data from the interim analyses can be used to determine early stopping for futility or efficacy, or to identify a patient subgroup based on their baseline covariate in the DWI × Time space.

To keep the dimensionality of the subset selection manageable, we partition the covariate space into six categories, which we refer to as 'low' and 'high' DWI, and 'short', 'medium' and 'long' Time since stroke. We label these categories by c_{ij} that corresponds to the cell defined by DWI level i and Time level j ($i = 1, 2; j = 1, 2, 3$), where the cell with the shortest time and lowest DWI is c_{11}. The primary objective of the study is to test the null hypothesis H_J of no improvement of endovascular therapy over the standard medical treatment in the entire patient population $\Pi_J = \cup_{i,j} c_{ij}$, with $J = 6$. However, in the case that H_J is accepted, the prior experience of the investigators and the biological rationale behind the endovascular procedure suggest that the search at the interim analysis should begin with $\Pi_1 = c_{11}$ (for which endovascular treatment is believed to be the most efficacious) and greedily agglomerate *neighboring* cells into $J - 1 = 5$ subgroups $\Pi_2 = \Pi_1 \cup c_{12}$, $\Pi_3 = \Pi_2 \cup c_{21}$, $\Pi_4 = \Pi_3 \cup c_{22}$, $\Pi_5 = \Pi_4 \cup c_{31}$.

Let W_i^ℓ be the sum of the ranks of the Rankin scores from the endovascular treatment arm in the combined sample of Π_i (consisting of both the endovascular and standard treatment arms) at the ℓth interim analysis. The standardized Wilcoxon statistic is

$$Z_i^\ell = \left\{ W_i^\ell - n_i^\ell \left(n_i^\ell + n_{0i}^\ell + 1 \right)/2 \right\} \Big/ \left\{ n_i^\ell n_{0i}^\ell \left(n_i^\ell + n_{0i}^\ell + 1 \right)/12 \right\}^{1/2}. \tag{14}$$

In Appendix B, we derive an exact formula for $\mathrm{Cov}\left(Z_i^\ell, Z_i^{\ell'}\right)$, $i \le j \le J$, $1 \le \ell \le \ell' \le 3$, and use it to derive the limiting joint normal distributions of the Z_i^ℓ, $i \le j \le J$ and $1 \le \ell \le \ell' \le 3$, under $P(Y \le X = 1/2)$ and under contiguous alternatives $P(Y \le X = 1/2 + \theta)$. The limiting joint normal distributions show that the procedures in Subsections 2.2 and 2.3 for standardized sample means (6) of normally distributed observations can be applied to the standardized Wilcoxon statistics (14). In particular, in place of the statistic (9) for testing $\mu_i \ge \mu_{0i} + \delta$ to allow early stopping for futility, we now define

$$\widetilde{Z}_i^\ell = \left\{ W_i^\ell - n_i^\ell \left(n_i^\ell + n_{0i}^\ell + 1 \right)/2 - n_i^\ell n_{0i}^\ell \theta \right\} \Big/ \left\{ n_i^\ell n_{0i}^\ell \left(n_i^\ell + n_{0i}^\ell + 1 \right)/12 \right\}^{1/2} \tag{15}$$

for testing $K_i : P(Y \le X | \Pi_i) = 1/2 + \theta$. With Z_i^ℓ and $\widetilde{Z}_i^\ell$ defined by Eqs. (14) and (15), we can apply the fixed sample size design in Subsection 2.2 and the group sequential design in Subsection 2.3 with $\sigma = 1$. The thresholds c_α in Eq. (3) and $b, \widetilde{b}, c$ in Eqs. (7)–(12) remain the same because of the aforementioned limiting multivariate normal distribution for the standardized Wilcoxon statistics.

3.2. Operating characteristics of adaptive clinical trial designs

To demonstrate the advantages of the proposed adaptive choice of patient subgroup, we compare the following RCT

Table 1
Operating characteristics in the case $q_{ij} = 1/6$ for all i, j.

		Design A				Design B	Design C
	Scenario	p $(p_1, \ldots p_6)$	N	P_e	P_f	p $(p_1, \ldots p_6)$	p
S0	$\mu = (0, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	5 (1.1, 0.7, 0.4, 0.4, 0.2, 2)	424	2.2	54.6	4.8 (0.9, 1, 0.9, 0.4, 0.2, 1.3)	4.8
S1	$\mu = (0.3, 0.3, 0.3, 0.3, 0.3, 0.3)$ $v = (1, 1, 1, 1, 1, 1)$	89 (0.9, 0.9, 0.6, 0.6, 1, 85)	371	73.6	0.6	89.8 (0.6, 0.6, 0.8, 1, 0.9, 86)	94.5
S2	$\mu = (0.3, 0.3, 0.3, 0.3, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	80.5 (3.8, 4.4, 6.7, 13.6, 2.3, 49.7)	432	42.2	1	77.7 (2.1, 3.2, 5.1, 14.3, 2.8, 50.2)	69.9
S3	$\mu = (0.3, 0.3, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	65.7 (14.7, 28.5, 6, 2.2, 0.7, 13.6)	456	30.9	3.6	49 (6.5, 20, 5.3, 2.2, 1, 14.1)	28.8
S4	$\mu = (0, 0, 0.3, 0.3, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	26.5 (0.6, 0.2, 1.4, 7.8, 1.6, 15)	451	11.6	21.8	25.9 (0.3, 0.2, 1, 7.7, 2.5, 14.3)	28.9
S5	$\mu = (0.6, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	86.8 (60.7, 9, 2.6, 1, 0.4, 13.1)	439	46.6	0.9	72.8 (49.9, 5.9, 2.1, 0.9, 0.3, 13.7)	28.1
S6	$\mu = (0.5, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	78 (55.4, 8.6, 2.2, 1.4, 0.6, 10.2)	443	44	2.3	57.4 (36.9, 5.9, 2.5, 1.3, 0.5, 10.4)	22.3
S7	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	92.5 (7.4, 13.4, 20, 3, 0.6, 48.1)	426	48	0.2	88.7 (5.4, 11.7, 18.3, 2.5, 0.8, 50)	68.9
S8	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (2, 1.5, 1, 0.5, 0.5, 0.5)$	83 (5.8, 7.4, 24.9, 3.5, 0.9, 40.5)	438	40.4	1.1	77.1 (3.6, 5, 23, 3.2, 0.9, 41.4)	61.3
S9	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (0.5, 1, 1.5, 2, 2, 2)$	72.2 (15.2, 16.2, 13.7, 4.1, 1.3, 21.7)	451	33.8	2.2	61.1 (8.9, 11.9, 13, 3.6, 1.7, 21.9)	39.8
S10	$\mu = (0, 0, 0, 0.3, 0.4, 0.5)$ $v = (1, 1, 1, 1, 1, 1)$	47.9 (0.2, 0, 0, 0.1, 0, 47.6)	423	35	12.6	50.1 (0.1, 0.1, 0, 0, 0.2, 49.7)	69.2

designs involving $N = 500$ patients in two simulation studies that are reported in Tables 1 and 2.

Design A: The group sequential design described in Subsection 2.3 with interim and final analyses performed at 300, 400 and 500 patients.
Design B: The fixed sample size design described in Subsection 2.2, with 500 patients.
Design C: A fixed sample size RCT that compares the endovascular and standard treatments in 500 patients and does not consider subpopulation selection.

The simulation studies, each covering ten scenarios for the alternative hypothesis besides the null hypothesis denoted by S0, assume that the (normalized) Rankin score of patients in the standard treatment arm follows a standard normal distribution regardless of DWI × Time status, except in scenario S8 where the variance is 0.5, and in scenario S9 where the variance is 2. The Rankin scores of patients in the endovascular treatment arm are assumed to be normally distributed with mean μ_{ij} and variance v_{ij} when the DWI × Time status falls in the cell c_{ij}, which has a prevalence q_{ij}. The μ_{ij} is given in a vector μ consisting of μ_{11}, μ_{12}, μ_{21}, μ_{22}, μ_{31} and μ_{32} (in that order) and

Table 2
Operating characteristics in the case $q = (0.2, 0.1, 0.3, 0.1, 0.1, 0.2)$.

		Design A				Design B	Design C
	Scenario	p $(p_1, \ldots p_6)$	N	P_e	P_f	p $(p_1, \ldots p_6)$	p
S0	$\mu = (0, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	5.2 (1, 0.8, 0.5, 0.3, 0.2, 2.4)	420	2.9	55.6	4.8 (1, 0.8, 0.7, 0.5, 0.4, 1.3)	4.8
S1	$\mu = (0.3, 0.3, 0.3, 0.3, 0.3, 0.3)$ $v = (1, 1, 1, 1, 1, 1)$	89 (0.9, 0.7, 0.9, 0.8, 1, 84.8)	370	73.6	0.5	89.7 (0.5, 0.6, 0.6, 0.6, 0.9, 86.6)	94.5
S2	$\mu = (0.3, 0.3, 0.3, 0.3, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	83 (3.5, 3.8, 7, 11, 3, 54.6)	425	45.6	0.6	79.8 (1.8, 3.1, 6.3, 10, 3, 55.5)	72.9
S3	$\mu = (0.3, 0.3, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	64.5 (17.4, 30.7, 2.8, 1.2, 1, 11.3)	454	31.9	4.5	46.1 (7.9, 20.9,2.9, 1.5, 0.8, 12.2)	25.6
S4	$\mu = (0, 0, 0.3, 0.3, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	33 (0.2, 0.1, 2.7, 7.3, 3.1, 19.6)	451	15	17.5	35.1 (0.3, 0.1, 2.6, 8.4, 3.4, 20.2)	36.5
S5	$\mu = (0.6, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	90.9 (56, 12.7, 1.6, 0.6, 0.5, 19.6)	438	46.2	0.4	81.1 (50, 9.4, 1.4, 0.8, 0.6, 18.9)	35.9
S6	$\mu = (0.5, 0, 0, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	82.9 (51.5, 14.1, 2.1, 1, 0.5, 13.7)	441	44.9	1.9	67.3 (40.2, 9.4, 2, 1, 0.6, 14)	28.4
S7	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (1, 1, 1, 1, 1, 1)$	94.1 (5.7, 8.6, 13.9, 3.4, 0.8, 61.6)	413	53.9	0.2	92.3 (4.6, 8.3, 12.8, 2.5, 1, 63)	79
S8	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (2, 1.5, 1, 0.5, 0.5, 0.5)$	87.9 (3.2, 3.7, 17.7, 3.3, 1, 58.9)	418	49.9	0.8	84.9 (2.2, 2.9, 17.4, 3.1, 1.2, 58.1)	75.5
S9	$\mu = (0.5, 0.4, 0.3, 0, 0, 0)$ $v = (0.5, 1, 1.5, 2, 2, 2)$	74.3 (16.1, 14.4, 10.7, 4, 1.8, 27.4)	447	35.9	2.2	65.3 (9.1, 11.2, 10.1, 3.9, 2, 29)	47.2
S10	$\mu = (0, 0, 0, 0.3, 0.4, 0.5)$ $v = (1, 1, 1, 1, 1, 1)$	38 (0.3, 0.1, 0, 0, 0, 37.5)	430	26.8	18.2	37.5 (0.3, 0.2, 0, 0, 0, 37)	57.2

the v_{ij} is given in a similar vector for each scenario in the tables. The q_{ij} is given in a similar vector shown in the heading of each table, which reports the average sample size N of the trial, the probability p_i (reported in %) of rejecting the null hypothesis H_i, and the overall probability $p = P(\text{reject } H_i \text{ for some } i) = \sum_{i=1}^{6} p_i$. The probabilities p_e and p_f of early stopping at any interim analyses by Design A for efficacy and futility, respectively, are also reported in %. Each result is based on 5000 simulations. Table 1 considers equal prevalence among individual cell c_{ij}. In this study, Design A uses $\widetilde{b} = -1.46$, $b = 2.39$, and $c = 2.31$ corresponding to $\alpha = 0.05$, $\beta = 0.2$ and $\epsilon = 0.5$, while Design B uses $c_\alpha = 2.16$. To examine the impact of the prevalence, the patient subgroups considered in the second simulation study have different prevalence rates, with most patients falling in cell c_{21}. In this simulation study, which is reported in Table 2, Design A uses $\widetilde{b} = -1.46$, $b = 2.37$, and $c = 2.28$, and Design B uses $c_\alpha = 2.14$. Design A assumes the effect size $\theta = 0.064$ to be the projected treatment effect used in the maximum sample size calculation at the design stage as mentioned in Subsection 3.1.

All three designs have type 1 error controlled at 5% and early stopping for futility can reduce the expected sample size by about 15–20% (scenario S0 in the tables). When the treatment effect is uniform over the patient subgroups (scenario S1), the power of Design A is slightly smaller than that of Design C (and nearly the same as that of Design B), but there is a considerable (>20%) reduction in expected sample size. In scenario S1, both Design A and Design B seldom choose a smaller subpopulation. When the actual distribution of the effect sizes agrees with the assumptions underlying the ordering of the subgroups (scenarios S2, S3, S5, S6, S7, S8, and S9), Design A has the highest power, and also results in a sample size that is on average at least 10% less than those of the other two designs. Although Design B provides a chance to test treatment effects in a selected subpopulation after H_{0J} is accepted, its power for testing the selected subpopulation is not as high as that of Design A. When the actual subgroup effects are contrary to the underlying assumptions (scenarios S4 and S10), Designs A and B exhibit a clear loss in power in comparison with Design C.

4. Discussion

This paper proposes novel efficient methods for the design and analysis of RCTs that allow adaptive choice of patient subgroups for confirmatory testing of a new treatment against a control. These patient subgroups can be defined by biomarkers, brain imaging, or other risk factors measured at baseline. Rosenblum et al. [13] have pointed out the usefulness of such methods in connection with testing the efficacy of antiretroviral medications for HIV positive patients. In particular, they note suggestive evidence from two recently completed RCTs of maraviroc that the treatment benefit may differ depending on the suppressive effect, as measured by the phenotypic severity score (PSS) at baseline, of the patient's background therapy. Specifically, the estimated average treatment benefit among patients with PSS less than 3 was found to be larger than among those with PSS 3 or more. The approach used by [13] is to conduct multiple testing of the average treatment effect for the overall population and for each of the two PSS-defined subpopulations, leading to the hypotheses H_j, $1 \le j \le 3$, with Π_3 being the overall population and Π_1 and Π_2 being disjoint subpopulations. Strong control of FWER imposes many constraints on the parametric space when the responses are assumed to be normally distributed, and [13] addresses the problem of constructing multiple testing procedures that satisfy these constraints and optimize power at a given set of alternatives by transforming a fine discretization of the multiple testing problem into a sparse linear program that has over a million variables and constraints and can be solved by recent advances in large, sparse linear programming. In contrast, our approach addresses testing treatment difference in an adaptively chosen patient subgroup, by partitioning the parameter space and defining the corresponding type I and type II errors. Instead of exact optimality, we resort to asymptotic efficiency via Kullback–Leibler information numbers, and use recent advances in group sequential and adaptive designs to further enhance adaptation and efficiency. Multiple testing methodology is used in our approach only as a means to an end, rather than as the end (in the sense of controlling FWER) itself.

The basic theory underlying our approach is developed in Section 2 for normally distributed responses with common known variance σ. Since σ is typically unknown in practice, an obvious modification is to replace σ in Subsection 2.2 or 2.3 by the sample estimate $\hat{\sigma}$, or $\hat{\sigma}^\ell$ at the ℓth interim analysis. Actually the methodology in Section 2 can be readily extended from the one-parameter normal family (with $\theta_j = \mu_j - \mu_{0j}$ as the unknown parameter for the jth patient subgroup) to multiparameter exponential families (thereby allowing unknown variances for different patient subgroups and treatments even for normally distributed responses). The GLR statistics which provide a basic tool for our parametric approach in Section 2 can be readily extended to multiparameter families; see Subsection 3.4 of [6] which shows how the GLR statistics can be used to develop efficient group sequential tests in multiparameter exponential families.

Section 3 extends our approach from parametric to nonparametric settings. We have focused on the Wilcoxon statistic because it is commonly used for two-sample comparison of the Rankin scores. Clearly the approach works much more generally for other nonparametric test statistics that have Chernoff–Savage-type representations; see [14]. In fact, as shown in Subsections 3.3 and 3.4 of [12], the group sequential testing theory in [6] based on GLR statistics can be extended to nonparametric or semiparametric statistics with fully observable or censored survival outcomes. This also applies to the group sequential tests with adaptive patient subgroup choice developed herein.

Although the group sequential designs in Section 2.3 and in Tables 1 and 2 have substantial savings in expected sample size, the savings may be compromised by the longer duration to recruit patients with narrower eligibility criteria. If the study duration turns out to be longer than that of a fixed sample size design that allows adaptive choice of patient subgroup in the final analysis (Design B in the tables), then Design B may be preferred to the group sequential Design A despite the sample size savings. Simon and Simon [15] have proposed a class of adaptive enrichment designs, quite different and considerably more restrictive than that proposed herein, which allow the eligibility criteria to be adaptively updated at interim analyses of a trial. They also point out that “the more one restricts the eligibility criteria, the longer patient accrual will take” and that their adaptive enrichment design "is most powerful relative to the standard non-adaptive approach when only a small subset of patients benefit, however, this is exactly when the accrual rate is most decreased." Note that our approach does not seek to find the patient subgroup for which the new treatment differs the most from the control. Instead, it seeks to find the patient subgroup that has the largest Kullback–Leibler information number $\mathcal{I}_j$ (which involves both the prevalence and the mean treatment difference) for $j \neq J$, after accepting the null hypothesis $H_J: \mu_J \leq \mu_{0J}$ for the entire population Π_J.

Yeatts et al. [16] recently reported a randomized clinical trial comparing IV-tPA followed by endovascular treatment with IV-tPA alone in IS patients. The trial intended to enroll 900 subjects, but was terminated after 656 subjects were randomized, based on a pre-specified criterion for futility stopping at the fourth interim analysis. Challenges facing the DSMB of the trial are described in [16]. One such challenge was the heterogeneity of the study population and the possibility of significant improvements of the endovascular approach in some patient subgroups, defined by severity of the stroke. However, these subgroups were not pre-specified and no statistically significant treatment difference between the two post-hoc severity-based subgroups was observed. It is pointed out in [16, p. 1412] that if it became clear during the study that the assumption of equal treatment difference over the two subgroups was invalid, the trial protocol might be amended to account for this unexpected difference. The new trial design proposed herein, therefore, provides an efficient method to address this difficulty.

Acknowledgments

Lai's research was supported by the NSF grant DMS-1106535 and the NIH grant 1 P30 CA124435-01. Lavori acknowledges the support by the Clinical and Translational Science Award UL1 RR025744 to the Stanford Center for Clinical and Translational Education and Research (Spectrum) from the National Center for Research Resources, National Institutes of Health to Stanford University.

Appendix A. Bahadur efficiency, closed testing principle, and group sequential extensions

A.1. Bahadur efficiency of proposed procedure in Subsection 2.2

Recall that $\theta_j = \mu_j - \mu_{0j}$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$. For $\boldsymbol{\theta} \notin \Theta_0$, $\theta_j > 0$ for all $1 \leq j \leq J$, and define $j(\boldsymbol{\theta}) = \ \text{argmax}_{1 \leq j \leq J} \sqrt{p_j}\theta_j$, i.e., $j(\boldsymbol{\theta})$ corresponds to the patient subgroup j with the largest KL information number $\mathcal{I}_j$, which depends on $\boldsymbol{\theta}$ and which will be denoted by $\mathcal{I}_j(\boldsymbol{\theta})$. Note that $\mathcal{I}_j(\boldsymbol{\theta})$ is linear in n and that the type II error $\beta(\boldsymbol{\theta})$ defined in Eq. (2) satisfies

$$\begin{aligned}\lim_{n\to\infty} n^{-1} \log \beta(\boldsymbol{\theta}) &= -\lim_{n\to\infty} n^{-1} \log \mathcal{I}_{j(\boldsymbol{\theta})}(\boldsymbol{\theta}) \\ &= \max_{1 \leq j \leq J} p_j \left(\mu_j - \mu_{0j}\right)^2 / \left(4\sigma^2\right),\end{aligned} \tag{16}$$

in view of the asymptotic behavior of Gaussian tail probabilities. In fact, $\mathcal{I}_{j(\boldsymbol{\theta})}(\boldsymbol{\theta})$ is also the KL divergence $I_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, where $\boldsymbol{\theta}_0 \in \Theta_0$ is the minimizer of

$$Q(\boldsymbol{\lambda}|\boldsymbol{\theta}) = (\boldsymbol{\theta} - \boldsymbol{\lambda}) \operatorname{diag}\left(p_1, \ldots, p_J\right) (\boldsymbol{\theta} - \boldsymbol{\lambda}' / \left(2\sigma^2\right) \tag{17}$$

over $\boldsymbol{\lambda} \in \Theta_0$; the expected value under $\boldsymbol{\theta}$ of the logarithm of the likelihood ratio of $\boldsymbol{\theta}$ to $\boldsymbol{\lambda}$ based on a sample of size $n/2$ is $\frac{n}{2} Q(\boldsymbol{\lambda}|\boldsymbol{\theta})$. Bahadur [5,17] has shown that $\exp\{-(1 + o(1) I_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0))\}$ is the asymptotically minimal value (as $n \to \infty$) for the type II error, at an alternative $\boldsymbol{\theta} \notin \Theta_0$, of any level-$\alpha$ test of $H: \boldsymbol{\theta} \in \Theta_0$ based on a sample of size $n/2$. A level-α test whose type II error attains this asymptotically minimal rate at every alternative $\boldsymbol{\theta}$ is said to be *Bahadur efficient*. This way of comparing tests via large deviation approximations applies to general parametric families. Even in cases where large deviation approximations to type II errors are difficult to derive, Bahadur [5,17] suggests to fix the type II error at an alternative $\boldsymbol{\theta}$ and to try deriving large deviation approximations for the type I error instead; specifically, for $\sup_{\boldsymbol{\lambda} \in \Theta_0} \alpha(\boldsymbol{\lambda})$. He also suggests the closely related idea of deriving large deviation approximations for the p-value of the test at $\boldsymbol{\theta} \notin \Theta_0$. These large deviation approximations lead to a limiting quantity $\inf_{\boldsymbol{\lambda} \in \Theta_0} q(\boldsymbol{\lambda}|\boldsymbol{\theta})$, which is called the Bahadur exact slope and is maximized where $q = Q$ given by Eq. (17).

A.2. Closed testing principle and proof of Eq. (4)

To prove Eq. (4), recall that $H_j : \theta_j \leq 0$ and $\Theta_0 = \{\boldsymbol{\theta} : H_j \text{ is true for some } j\}$. For $\boldsymbol{\theta} \in \Theta_0$, let $J(\boldsymbol{\theta})$ denote the set of the true hypotheses. The family-wise error rate (FWER) is defined as $P_{\boldsymbol{\theta}}\{\text{Reject } H_j \text{ for some } j \in J(\boldsymbol{\theta})\}$, and one can use the closed testing principle (CTP) to control FWER [3,4]. This principle assumes that there is a level-α test of every intersection hypothesis $\cap_{i \in I} H_i$, $I \subset \{1, \ldots, J\}$, and uses it to test the intersection hypothesis if and only if all intersection hypotheses $\cap_{i \in I^*} H_i$ with $I \overset{\subset}{\neq} II^*$ are tested and rejected. Since a type I error is committed if and only if $\cap_{i \in J(\boldsymbol{\theta})} H_i$ is tested and rejected under CTP, it follows that FWER $\leq a$.

We now use a variant of CTP to prove Eq. (4). First note that the procedure in Subsection 2.2 is tantamount to (a) first testing H_J by using GLR$_J$ and (b) using $\max_{1 \leq i \leq J-1}$GLR$_i$ to test the intersection null $\cap_{i \neq J} H_i$ if H_J is accepted; the critical value remains the same for (a) and (b). If we use CTP and apply $\max_{i \in I}$GLR$_i$ to test the intersection hypothesis $\cap_{i \in I} H_i$ with $I \subset \{1, \ldots, J-1\}$, this gives the FWER for testing H_i, $1 \leq i \leq J-1$ after accepting H_J. Note that

$$\begin{aligned}&P\left(\text{GLR}_J \leq c_\alpha, \ \max_{1 \leq i \leq J-1} \text{GLR}_i \geq c_\alpha | \cap_{i=1}^J H_i\right) \\ &\geq P\left(\text{GLR}_J \leq c_\alpha, \ \max_{i \in I} \text{GLR}_i \geq c_\alpha \vee \max_{i \notin I \cup \{J\}} \text{GLR}_i | H_J \cap \left(\bigcap_{i \in I} H_i\right)\right),\end{aligned} \tag{18}$$

and that the square root of the GLR statistics are jointly normal random variables, with means ≤ 0 except for $\sqrt{\text{GLR}_i}$ for $i \notin I$ in the second probability, under the respective intersection hypotheses. It then follows that FWER $\leq a(\mathbf{0})$, which is the type I error of testing the global null hypothesis $\cap_{i=1}^J H_i$ under the parameter value $\mathbf{0}$. This shows that $\alpha(\boldsymbol{\theta}) \leq \alpha(\mathbf{0})$ for $\boldsymbol{\theta} \in \Theta_0$.

A.3. Extension of Eq. (4) to group sequential design

CTP is also applicable to group sequential design [18,19]. In particular, for the 3-stage design in Subsection 2.3, subgroup selection immediately follows acceptance of H_J at an interim analysis and then proceeds with the selected subgroup, using the maximum of GLR_i over $i \neq J$ as the selection criterion at the interim analysis. Although there seems to be additional complication because the 3-stage design allows early stopping for futility, this can be easily incorporated into the argument to show that an analog of Eq. (18) still holds for the group sequential design; see a similar argument in Appendix B of [20] for a closely related problem.

Appendix B. Joint distribution of standardized Wilcoxon statistics over subgroups and time

B.1. Formula for null covariances of standardized Wilcoxon statistics

Consider the standardized Wilcoxon statistics Z_i^ℓ given by Eq. (14) for patient subgroup i at the ℓth interim analysis, under the null hypothesis of no treatment difference between the new and the standard treatments. We now derive an explicit formula for $\mathrm{Cov}\left(Z_i^\ell, Z_j^{\ell'}\right)$, with $1 \le i \le j \le J$ and $1 \le \ell \le \ell'$. To simplify the notation, let W denote the Wilcoxon statistic based on a sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ and let W' denote that based on a larger sample $X_1, \ldots, X_{m'}, \; Y_1, \ldots, Y_{n'}$ with $m \le m'$ and $n \le n'$. The null hypothesis of no treatment difference translates into the same distribution of X_i and Y_j. Since the X_i and Y_j are i.i.d., a standard combinatorial argument gives $E(W) = m(m+n+1)/2$ and $\mathrm{Var}(W) = mn(m+n+1)/12$; see [21, p. 70 and pp. 332–333]. The same argument can be used to show that

$$\mathrm{Cov}(W, W') = \frac{mn}{4} + \frac{mn(n'-1)}{12} + \frac{mn(m'-1)}{12} = \frac{mn(m'+n'+1)}{12}. \tag{19}$$

Let $Z = \{W - m(m+n+1)/2\}/\{mn(m+n+1)/12\}^{1/2}$ be the standardized Wilcoxon statistic and define Z' similarly with W', m', n' in place of W, m, n. Then it follows from Eq. (19) that

$$\begin{aligned}\mathrm{Cov}(Z, Z') &= \{mn(m'+n'+1)\} \\ &\quad / \left[\{mn(m+n+1)\}^{1/2}\{m'n'(m'+n'+1)/12\}^{1/2}\right] \qquad 20 \\ &= \left(\frac{mn}{m+n+1}\right)^{1/2} \bigg/ \left(\frac{m'n'}{m'+n'+1}\right)^{1/2}.\end{aligned}$$

The preceding formula is the same as that in the case $Z = (\overline{X}_m - \overline{Y}_n)/(m^{-1} + n^{-1})^{1/2}$ for the sample means $\overline{X}_m$ and $\overline{Y}_n$ of i.i.d. observations $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ that have common variance 1. Here $(m^{-1} + n^{-1})^{-1} = mn/(m+n)$ is the "information clock" of $\overline{X}_m - \overline{Y}_n$, as first observed by Robbins and Siegmund [22].

B.2. Limiting null distribution of $(Z_i^\ell : 1 \le i \le J,\ 1 \le \ell \le L)$

A key lemma of [22] is that for the difference of sample means $\overline{X}_m - \overline{Y}_n$ of (standard) normal observations,

$$\left(\frac{mn}{m+n}\right)\left(\overline{X}_m - \overline{Y}_n\right) \text{ has the same joint distribution as } B\left(\frac{mn}{m+n}\right), \tag{21}$$

where $B(t)$, $t \ge 0$ is Brownian motion. Hence the joint distribution has independent increments in the information time $mn/(m+n)$, and this independent increments property is used in the recursive integration algorithms in Subsection 2.3. For the standardized Wilcoxon statistics, Eq. (20) shows the same covariance structure as that for the standardized sample means $\left(\overline{X}_m - \overline{Y}_n\right)/\left(m^{-1} + n^{-1}\right)^{1/2}$. In fact, Eq. (21) still holds asymptotically as $\min(m, n) \to \infty$:

$$\frac{W_{m,n} - m(m+n+1)/2}{\{mn(m+n+1)/12\}^{1/2}} \approx \frac{B(mn/(m+n))}{\{mn/(m+n)\}^{1/2}}. \tag{22}$$

In Eq. (22), we denote Wilcoxon's rank sum by $W_{m,n}$ to emphasize the dependence on the sample sizes, and use $\approx$ to denote having the same limiting joint distribution. As in Eq. (21), we consider the joint distribution in Eq. (22) over a prescribed set of information times of the form $m_k n_k/(m_k + n_k)$, $k = 1, \ldots, K$. To prove Eq. (22), we can use the Chernoff–Savage representation [23] of rank statistics in terms of sample means of i.i.d. random variables, plus an asymptotically negligible remainder term, and its refinement in [14] that gives stronger results on the remainder term.

For contiguous alternatives that satisfy $\sqrt{N}\theta$ converging to some finite constant μ, where N is the maximum sample size, the standardized Wilcoxon statistics Z_i^ℓ in Eq. (14) still have the same limiting covariance structure as in the null case but its mean is of the order

$$\left(\frac{n_i^\ell n_{0i}^\ell}{n_i^\ell + n_{0i}^\ell}\right)^{1/2} \theta \sim \frac{1}{2}\left(p_i N_\ell\right)^{1/2}\theta,$$

recalling that patients are randomized to treatment and control and that the total number of patients in subgroup i enrolled up to the ℓth interim analysis is of the order $p_i N_\ell$. Hence the limiting joint distribution of $(Z_i^\ell : 1 \le i \le J,\ 1 \le \ell \le L)$ is that of a multivariate normal as in the null case but with means $\frac{1}{2}\left(p_i N_\ell/N\right)^{1/2}\mu$.

References

[1] Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Ann Math Stat 1952;23(4):493–507.

[2] Chernoff H. Sequential analysis and optimal design. Society for Industrial and Applied Mathematics; 1972.

[3] Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 1976;63(3):655–60.

[4] Hochberg Y, Tamhane AC. Multiple comparison procedures. New York: John Wiley & Sons, Inc.; 1987.

[5] Bahadur RR. Stochastic comparison of tests. Ann Math Stat 1960;31(2):276–95.

[6] Lai TL, Shih MC. Power, sample size and adaptation considerations in the design of group sequential clinical trials. Biometrika 2004;91(3):507–28.

[7] Bartroff J, Lai TL, Shih MC. Sequential experimentation in clinical trials. New York: Springer; 2013.

[8] Bartroff J, Lai TL. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. Stat Med 2008;27(10):1593–611.

[9] Bartroff J, Lai TL. Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. Seq Anal 2008;27(3):254–76.

[10] Jennison C, Turnbull B. Mid-course sample size modification in clinical trials based on the observed treatment effect. Stat Med 2003;22(6):971–93.

[11] Jennison C, Turnbull B. Adaptive and nonadaptive group sequential trials. Biometrika 2006;93(1):1–21.

[12] He P, Lai TL, Liao OY. Futility stopping in clinical trials. Stat Interface 2012;5(4):415–23.

[13] Rosenblum M, Liu H, Yen EH. Optimal tests of treatment effects for the overall population and two subpopulations in randomized trials, using sparse linear programming. J Am Stat Assoc 2014. http://dx.doi.org/10.1080/01621459.2013.879063 [in press].

[14] Chernoff H, Savage R. On Chernoff–Savage statistics and sequential rank tests. Ann Stat 1975;3(4):825–45.

[15] Simon N, Simon R. Adaptive enrichment designs for clinical trials. Biostatistics 2013;14(4):613–25.

[16] Yeatts S, Martin R, Coffey C, Lyden P, Foster L, Woolson R, et al. Challenges of decision making regarding futility in a randomized trial: the Interventional Management of Stroke III experience. Stroke 2014;45(5):1408–14.

[17] Bahadur RR. Rates of convergence of estimates and test statistics. Ann Math Stat 1967;38(2):303–24.

[18] Jennison C, Turnbull B. Confirmatory Phase II/III clinical trials with hypotheses selection at interim: opportunities and limitations. Biom J 2006;48(4):650–5.

[19] Tang D, Geller N. Closed testing procedure for group sequential trials with multiple endpoints. Biometrics 1999;55(4):1188–92.

[20] Wang Y, Lan KG, Li G, Ouyang SP. A group sequential procedure for interim treatment selection. Stat Biopharm Res 2011;3(1):1–13.

[21] Lehmann E. Nonparametrics: statistical methods based on ranks. San Francisco: Holden-Day, Inc.; 1975.

[22] Robbins H, Siegmund D. Sequential tests involving two populations. J Am Stat Assoc 1974;69(1):132–9.

[23] Chernoff H, Savage R. Asymptotic normality and efficiency of certain nonparametric test statistics. Ann Math Stat 1958;29(4):972–94.