Received: 13 October 2017 | Revised: 6 April 2018 | Accepted: 26 July 2018

DOI: 10.1002/sim.7946

SPECIAL ISSUE PAPER

WILEY Statistics in Medicine

Adaptive enrichment designs for confirmatory trials

Tze Leung Lai[1,2,3] | Philip W. Lavori[1,2] | Ka Wai Tsang[3]

[1]Department of Statistics, Stanford University, Stanford, California

[2]Department of Biomedical Data Science, Stanford University, Stanford, California

[3]School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

Correspondence
Tze Leung Lai, Stanford University, Stanford, CA 94305.
Email: lait@stanford.edu

Funding information
National Science Foundation, Grant/Award Number: DMS 1407828; National Institutes of Health, Grant/Award Number: 1P30 CA124435 and 1UL1 RR025744

After an overview of the Food and Drugs Administration's 2012 draft guidance on enrichment strategies for clinical trials to support drug/biologic approval, we describe subsequent advances in adaptive enrichment designs in this direction. We also provide a concrete application in the enrichment design of the Diffusion and Perfusion Imaging Evaluation for Understanding Stroke Evolution 3 trial comparing a new endovascular treatment with standard of care for ischemic stroke patients.

KEYWORDS

adaptive designs, clinical trials, heterogeneity

1 | INTRODUCTION

Because of developments in targeted therapies and precision medicine in the past two decades, there has been much recent interest in enrichment clinical trials that use biomarkers and other patient characteristics to enrich the study population so that the treatment effect can be detected in the selected subpopulation but would not have been statistically significant in the entire population. Although conventional RCT designs can be used for enrichment clinical trials through the inclusion-exclusion criteria for patient accrual if the patient characteristics for enrichment can be delineated at the planning stage on the basis of early-phase trials or related studies reported in the literature, this is often not the case even for confirmatory Phase III trials and adaptive designs that allow mid-course enrichment using data collected have recently been developed in the works of Simon and Simon[1] and Lai et al.[2] These adaptive designs can preserve the type I error probability, which is important for regulatory approval of the new therapy. Concerning regulatory issues, the Food and Drugs Administration (FDA) published in 2012 a draft guidance on enrichment strategies for clinical trials to support drug/biologic approval. In Section 2, we review these developments, describe the basic methodological advances in the design and analysis of the adaptive enrichment trials, and also discuss the underlying regulatory considerations. Section 3 provides a concrete case study in adaptive enrichment trial designs comparing a new treatment with standard care for ischemic stroke patients. Concluding remarks, together with a discussion of (i) other developments in adaptive enrichment designs and (ii) how adaptive designs can adapt not only to the information collected during the course of the trial but also to external information collected from other trials and scientific or technological advances, are given in Section 4.

2 | THE FDA 2012 DRAFT GUIDANCE AND SUBSEQUENT METHODOLOGICAL DEVELOPMENTS

2.1 | The FDA's Guidance for Industry and type I error control

The FDA's Guidance for Industry on "Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products" was published in December 2012. It groups enrichment strategies into three broad categories, ie, (i) strategies to decrease heterogeneity, (ii) prognostic enrichment strategies, and (iii) predictive enrichment strategies. Concerning (i), it says that approaches to increasing study power (ie, the probability of establishing a treatment effect if one is present) by decreasing nondrug related variability (heterogeneity) have already been widely used, but notes that "removing poor compliers identified after randomization is generally not acceptable" because compliance has been linked to outcome, whereas selecting patients likely to comply with treatment prior to randomization is an acceptable way to decrease heterogeneity. Prognostic enrichment in (ii) refers to choosing high-risk patients, ie, those more likely to experience the disease-related endpoint. Predictive enrichment in (iii) refers to choosing patients that are more likely to respond to the drug, based on their physiology or previous record of response to a treatment class, or disease characteristics that are related to the drug's mechanism. Section VI.C of FDA's December 2012 Guidance for Industry points out the importance of type I error rate control for regulatory approval of the new treatment and the challenges for enrichment designs, saying "Determining the required sample size that will provide reasonable power to test the different hypotheses while controlling type I error (usually including a prespecified order of testing or a multiple testing procedure allowing testing of both hypotheses) is challenging."

2.2 | Adaptive enrichment and type I error rate control

Section VI.D of FDA's guidance is entitled "Adaptive Enrichment" and begins with the following.

> Although an enrichment characteristic should almost always be specified before a study begins, certain adaptive designs can use enrichment strategies that identify predictive markers during the course of the study. Specifically, entry criteria or sample size can be modified for later stages of a trial if factors can be identified that increase event rate or treatment response (eg, discovery that the enrichment factor has a greater impact on response than anticipated or that the patients without the enrichment factor have a very low response or safety concern). Such changes will need appropriate type I error rate control to account for interim, unblinded analyses of the accumulating data and type I error allocation if there were analyses of multiple subgroups.

It emphasizes that "the issue of whether the statistical testing results obtained by such an adaptive enrichment strategy are reproducible needs to be addressed" and refers to FDA's 2010 Guidance for Industry on "Adaptive Design of Clinical Trials for Drugs and Biologics." It lists a few "potentially applicable" adaptive designs but points out that "there has been little practical experience with enriched study designs whose sample size changes after the start of the study, or where other changes in the design are preplanned to be based on accrued information during a trial." Among them are interim analyses, which "could reveal, either on an early endpoint (eg, imaging or PD biomarker or tumor response rate) or later endpoint (eg, progression-free survival) that the marker-negative population has a much lower response than the marker-positive group." It notes that "sample size planning in these designs can be difficult, because such designs are generally used when there is uncertainty about the prevalence of a marker, its predictiveness, and what sample size or entry criteria adjustments are contemplated."

Simon and Simon[1] proposed a class of adaptive enrichment designs, "which allow the eligibility criteria of a trial to be adaptively updated during the trial, restricting entry to patients likely to benefit from the new treatment." They show how the type I error probability can be preserved in the case of binary response variables, for which they use the number of successes on the new treatment plus the number of failures on the control as the test statistic S, using the framework of fixed sample size (FSS) enrichment design summarized in Table 1 as follows. Central to this framework is a function f that maps the covariate space (consisting of biomarkers and other predictors measured from each subject) to $\{0, 1\}$ defined by $f(x) = I_{\{p_T(x) > p_C(x)\}}$, where $p_T(x)$ (or $p_C(x)$) is the probability of response for a patient with covariate x, given treatment (or control). Under the null hypothesis that $p_T(x) = p_C(x)$ for all x, the test statistic S is Binomial$(n, 1/2)$, irrespective of how the enrollment criteria change during the course of the trial, and therefore the type I error probability can be preserved by applying the critical values of the conventional binomial test. Since $p_T(x)$ and $p_C(x)$ are actually unknown, $f(x)$ has to

TABLE 1 Simon and Simon's adaptive enrichment design for n patients

1. Randomize the first m_0 patients to treatment T and control C.
2. **for** $m = m_0 + 1, \ldots, n$ **do**
   (a) compute $\hat{f}_m$ based on previous $m - 1$ patients;
   (b) restrict entry into clinical trial to only patients with $\hat{f}_m(x) = 1$.
   end for

be substituted by its sequential estimates $\hat{f}_m$ during the course of the trial based on the responses and covariates of the previous $m - 1$ patients. Sections 3 and 6 in the work of Simon and Simon[1] discussed estimation of $f(x)$, with Section 3 focusing on univariate x (single biomarker setting, with "a discrete set of candidate cut points," which may represent quantiles of numerical assays) and section 4 describing possible extensions to continuous responses and survival data.

2.3 | Efficient adaptive enrichment tests that preserve type I error for FSS trials

It is widely recognized that the comparative efficacy of a new treatment can depend on certain characteristics of the patients that are difficult to prespecify at the design stage. 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. To begin with, suppose n patients are randomized to the new and control treatments and the responses are normally distributed, with common known variance σ^2, mean μ_j for the new treatment, and μ_{0j} for the control treatment if the patient falls in predefined subgroups Π_j for $j = 1, \ldots, J$, in which Π_J denotes the entire patient population. Let $x_+ = \max(x, 0)$ and p_j be the prevalence of patient subgroup Π_j; hence, 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 Kullback-Leiber (KL) information number for Π_j is $I_j = \frac{n}{4} p_j (\mu_j - \mu_{0j})_+^2 / \sigma^2$, 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 I_j over $1 \le j \le J$. 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 using the square root of a generalized likelihood ratio (GLR) statistic. Lai et al[2] considered adaptive testing of the multiple hypotheses $H_i : \mu_i \le \mu_{0i}$ using

$$\mathrm{GLR}_i = \{n_i n_{0i}/(n_i + n_{0i})\}^{1/2} (\hat{\mu}_i - \hat{\mu}_{0i})_+ / \sigma,$$

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. Note that GLR_i is the sample estimate of the KL information number I_i. The adaptive test of $\{H_j, 1 \le j \le J\}$ is carried out after all n patients have been randomized to the treatments and first tests H_J. If $\mathrm{GLR}_J \ge c_\alpha$, reject H_J and claim the new treatment to be superior to control. Otherwise, choose the patient subgroup $\hat{\mathrm{I}} \ne J$ with the largest GLR_i among all subgroups $i = 1, \ldots, J-1$, and reject $H_{\hat{\mathrm{I}}}$ if $\mathrm{GLR}_{\hat{\mathrm{I}}} \ge c_\alpha$. The threshold c_α is so chosen that $\alpha(\mathbf{0}) = \alpha$, where $\alpha(\boldsymbol{\theta})$ is defined below in (1). Thus, the adaptive test starts with testing for the entire population Π_J and then enriches the study population by choosing the patient subgroup $\hat{\mathrm{I}}$ with the largest estimated KL information if the test fails to establish superiority of new treatment for Π_J, allowing the new treatment to be claimed better than control for the patient subgroup $\hat{\mathrm{I}}$ if $H_{\hat{\mathrm{I}}}$ is rejected.

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

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

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. Since the null hypothesis is highly composite, a uniformly most powerful level-α test is not expected to exist, and the work of

Lai et al[2] established asymptotic efficiency of the test as $n \to \infty$, in the sense of attaining the asymptotically minimal rate of $n^{-1} \log \beta(\theta)$ at $\theta \notin \Theta_0$, or equivalently, at θ such that $\theta_j > 0$ for all $j \le J$, where

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

which is the type II error probability of wrongly rejecting the new treatment. By making use of the closed testing principle, it was shown in the work of Lai et al[2] that $\sup_{\theta \in \Theta_0} \alpha(\theta) = \alpha(\mathbf{0})$.

To compute $\alpha(\mathbf{0})$, Lai et al[2] used the approximations $n_i \approx n_{0i}$ (since study subjects are equally likely to receive the new treatment or control) and $n_i + n_{i0} \approx np_i$, thereby approximating the random variables n_i and n_{0i} by $np/2$ and $n_i n_{0i}/(n_i + n_{0i})$ by $np/4$. 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 $\theta = \mathbf{0}$, 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 $\Pi_1 \subset \cdots \subset \Pi_J$, for which $\text{Cov}(Z_i, Z_j) \approx \sqrt{p_i/p_j}$ for $i \le j$. Hence, the threshold c_α can be determined in this case by solving the equation

$$\alpha = \alpha(\mathbf{0}) = P_0(Z_J \ge c_\alpha) + \sum_{i=1}^{J-1} \int_{c_\alpha}^{\infty} P_0\left(Z_J < c_\alpha, Z_j < x \text{ for } j \notin \{i, J\} | Z_i = x\right) \phi(x) dx,$$

where ϕ is the density function of the standard normal Z_i, recalling that $\hat{I} = \arg\max_{1 \le i \le J-1} \text{GLR}_i$. Note that, unlike the work of Simon and Simon[1] that allows change of eligibility criteria after an initial sample of size m_0, the FSS trial in Lai et al[2] has a conventional RCT design and incorporates enrichment via adaptive testing of multiple hypothesis.

2.4 | Efficient group sequential enrichment designs

The effect size $\mu_J - \mu_{0J}$ chosen for 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" of Lai and Shih.[3(p81)] 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 midcourse sample size re-estimation; see chapter 8 of Bartroff and Lai.[4] Since adaptive enrichment allows treatment comparisons over smaller patient subgroups and there is usually little information about their effect size from previous studies at the beginning of the trial, a group sequential design that can both choose the patient subgroup and re-estimate the sample size is particularly attractive.

Bartroff and Lai[5] have developed a theory of efficient adaptive design for sample size re-estimation that involves three-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 Bartroff and Lai[5] and Jennison and Turnbull.[6] Lai et al[2] extended these approximations of the adaptive design to a three-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 midpoint of the first stage and final sample sizes. As in the case for FSS designs, the maximum sample size is the FSS for testing H_J, or some inflation thereof, determined by the power at some effect size δ for the entire population. As pointed out by Lai and Shih,[3] 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, ie, the subgroup $i \ne J$ that maximizes GLR_i, 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 continue to the next stage of the three-stage design and repeat the procedure. For $l = 1, 2, 3$, let $\hat{\mu}_i^l(\hat{\mu}_{i0}^l)$ denote the mean response of patients in the subgroup Π_i from the treatment(control) arm and $n_i^l(n_{0i}^l)$ denote the corresponding sample size at stage l. Table 2 summarizes the design and its implementation for $\Pi_1 \subset \cdots \subset \Pi_J$.

Note that, in Table 2, $\hat{I}$ and $\tilde{I}$ are random variables with values ranging from 1 to $J-1$. Moreover, the thresholds $b, \tilde{b} < 0$, and c are determined by solving the equations

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

TABLE 2 A three-stage adaptive enrichment design

1. Set $j = J$.
2. **for** $l = 1, 2$ **do**
   (a) compute the test statistics $Z_j^l = \{n_i^l n_{0i}^l/(n_i^l + n_{0i}^l)\}^{1/2}(\hat{\mu}_i^l - \hat{\mu}_{0i}^l)/\sigma$ and $\tilde{Z}_j^l = \{n_i^l n_{0i}^l/(n_i^l + n_{0i}^l)\}^{1/2}(\hat{\mu}_i^l - \hat{\mu}_{0i}^l - \delta)/\sigma$
   (b) if $Z_j^l > b$, STOP and reject H_j
   (c) if $\tilde{Z}_j^l \le \tilde{b}$ and $j < J$, STOP and claim futility
   (d) if $\tilde{Z}_j^l \le \tilde{b}$ and $j = J$, find $\hat{\mathrm{I}} = \arg\max_{i<J} Z_i^l$
      (i) if $Z_{\hat{\mathrm{I}}}^l > b$, STOP and reject $H_{\hat{\mathrm{I}}}$
      (ii) if $\tilde{Z}_{\hat{\mathrm{I}}}^l \le \tilde{b}$, STOP and claim futility
      (iii) else, NEXT and set $j = \hat{\mathrm{I}}$
   (e) else, NEXT
   end for
3. if the trial proceeds to stage 3, then compute Z_j^3 for $j = J$ or $\hat{\mathrm{I}}$
   (A) if $Z_j^3 \ge c$, STOP and reject H_j
   (B) else, then
      (i) if $j < J$, accept $\{H_i, 1 \le i \le J\}$
      (ii) if $j = J$, then compute Z_i^3 for $i < J$ and $\tilde{I} = \arg\max_{i<J} Z_i^3$. Reject $H_{\tilde{I}}$ if $Z_{\tilde{I}}^3 > c$, and accept $\{H_i, 1 \le i \le J\}$ otherwise

for the prescribed maximum type I error α, power $1 - \beta$ at the alternative δ, and the proportion $0 < \varepsilon < 1$ of type I (or type II) error spent at interim analyses. In 2, P_{1b}, P_{1c}, P_{2b}, and P_{2c} are defined by 3, 4, and 5 as follows using the decomposition of the upper bound $\alpha(\mathbf{0})$ of the type I error of this group sequential test into two parts. The first part is $P_{\mathbf{0}}$(Reject H_J) that is bounded by

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

The second part is $P_{\mathbf{0}}$(Accept H_J and reject $H_{\hat{\mathrm{I}}}$), which can be evaluated by $P_{2b} + P_{2c}$, where

$$P_{2b} = P_{\mathbf{0}}\left(\tilde{Z}_J^l \le \tilde{b} \text{ and } Z_{\hat{\mathrm{I}}}^{l'} \ge b \text{ for some } l \le l' < 3\right), \tag{4}$$

$$P_{2c} = P_{\mathbf{0}}\left(\tilde{Z}_J^l \le \tilde{b}, Z_{\hat{\mathrm{I}}}^{l'} < b \text{ for } l \le l' < 3, \text{ and } Z_{\hat{\mathrm{I}}}^3 \ge c\right). \tag{5}$$

The boundaries $b, \tilde{b}$, and c have been first introduced by Lai and Shih[3] for "group sequential GLR tests with modified Haybittle-Peto (MHP) boundaries" in their theory of asymptotically optimal group sequential tests subject to type I error and maximum sample size constraints. As pointed out by Lai and Shih,[3] although funding and administrative considerations often play a basic role in the choice of the maximum sample size M of a clinical trial, justification of this choice in the trial protocol is typically based on some prescribed power $1 - \beta$ at the implied alternative $\boldsymbol{\theta}(M)$. The MHP boundaries aim at attaining "nearly optimal power and expected sample size properties" subject to the type I error and maximum sample size constraints. Bartroff and Lai[5] used MHP boundaries in their three-stage (corresponding to group sequential with three groups) GLR tests, which Lai et al[2] modified to develop efficient group sequential enrichment designs. The $\tilde{Z}_j^l$ in Table 2 and (2), (4), and (5) is related to the type II error at the implied alternative δ for testing $H_j : \mu_j - \mu_{0j} \le 0$ versus $K_j : \mu_j - \mu_{0j} \ge \delta$. For the case, $\mathrm{Cov}(Z_J^l, Z_J^{l'}) \approx \sqrt{n_J^l/n_J^{l'}}$ for $l \le l'$. Hence, the probabilities P_{1b} and P_{1c} can be computed by the recursive numerical integration[4] using the joint normal density of $\{Z_J^l, 1 \le l \le 3\}$. Recursive numerical integration can be also used to compute the probability

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

and likewise for P_{2c}, using the joint normal distribution of $\{Z_i^{l'}, l \le l' \le 3\}$ conditional on Z_i^l, with l representing the stage when H_J is accepted, and the joint normal distribution of $\tilde{Z}_J^l \approx Z_J^l - \sqrt{n_J^l}\delta/(2\sigma)$ and $Z_j^l, j \notin \{i, J\}$, conditional on Z_i^l with $i \ne J$. We want to point out in this connection the typo in the formula for P_{2b} in the work of Lai et al,[2] in which $\int_{-\infty}^{b}$ should be $\int_{-\infty}^{\infty}$ as given here. This error has been corrected in the `ASSISTant` package,[7] distributed for R through the Comprehensive R Archive Network, http://cran.r-project.org/, for the implementation of the three-stage adaptive enrichment design.

2.5 | Adaptive randomization for enrichment designs

The enrichment designs in Tables 1 and 2 use equal randomization to treatment and control. Other randomization ratios have been used, eg, the 2 : 1 ratio used by the Interventional Management of Stroke (IMS) III trial described in the next section. Lai and Liao[8] and Lai et al[9] developed a theory of asymptotically optimal sampling ratios in 2012, as a frequentist alternative to Bayesian adaptive randomization (AR) schemes for enrichment designs introduced earlier and summarized by Berry et al[10] (which may fail to maintain the type I error probability as discussed in Section 2.3 in the work of Lai et al[11]). Although Lai and Liao[8] did not actually consider patient subgroups and focused on different treatment strategies (arms) for the new treatment, their work can provide an extension to the case where these strategies are customized for different subgroups of the population. Moreover, while Lai and Liao[8] and Lai et al[9] focused only on maintaining the type I error under the intersection null hypothesis $\cap_{1\le j\le J} H_j$, Lai et al[2] subsequently used the closed testing principle for multiple testing to show that $\alpha(\boldsymbol{\theta}) \le \alpha$ for all $\boldsymbol{\theta} \in \Theta_0$. Although Lai et al[2] used equal randomization, the results can be readily extended to AR.

Following up on the works of Lai et al[9] and Berry et al,[10] Lai et al[12] gave a review of adaptive enrichment designs, "briefly for clinical trials in new drug development and in more detail for comparativeness effectiveness trials involving approved treatments." They also used the ideas of Lai and Liao[8] and multiarm bandit theory, developed by Lai and Robbins[13] and Lai,[14] to introduce a new group sequential enrichment design, which uses AR and GLR statistics to "fulfill multiple objectives, which include (i) treating accrued patients with the best (yet unknown) available treatment, (ii) developing a treatment strategy for future patients, and (iii) demonstrating that the strategy developed indeed has better treatment effect than the historical mean effect of SOC plus a predetermined threshold." They note that, because of the need for informed consent, the clinical trial needs to use randomization in a double blind setting, and the "randomization probability $\pi_{jk}^{(l)}$, determined at the lth interim analysis, of assigning a patient in group j to treatment k cannot be too small to suggest obvious inferiority of the treatments being tried, that is $\pi_{jk}^{(l)} \ge \epsilon$ for some $0 < \epsilon < 1/K$." Using this constraint, they derive an AR scheme, called ϵ-greedy scheme in reinforcement learning, from multiarmed bandit theory. This randomization scheme is easy to implement, in contrast with the Thompson sampling scheme in the Bayesian approach that requires Markov chain Monte Carlo to implement and is also less efficient than ϵ-greedy sampling schemes. Further discussions will be given in Section 4.

3 | THE DEFUSE 3 TRIAL

3.1 | From IMS III trial to DEFUSE 3 trial for ischemic stroke patients

Broderick and Tomsick[15] gave an overview of the NIH-sponsored IMS trials, beginning with the IMS I trial that began in January 2001. Noting that stroke is the third leading cause of death, the leading cause of serious long-term disability in the US, and affects more than 700 000 individuals in the US annually, with ischemic stroke accounting for nearly 88% of all strokes, the aforementioned[15] pointed out that intravenous (IV) administration of tissue plasminogen activator (tPA) within 3 hours of the onset of ischemic stroke has become the standard of care that can be administered by most hospitals, after the publication of a National Institute of Neurological Disorders and Stroke (NINDS) trial demonstrating its effectiveness in 1995. The goal of IMS I, in which 80 subjects were treated at 17 centers, was to test if a combined IV and intra-arterial (IA) delivery of tPA was a feasible method for the reopening of blocked arteries in treating ischemic stroke patients. It was completed in early 2002 and showed besides the feasibility of the combined delivery that a higher percentage of IMS I patients with a similar safety profile achieved functional independence within 90 days than in the NINDS tPA stroke study, ie, 43% of patients had a modified Rankin score (mRS) of 0 to 2 at 90 days compared with 39% of NINDS patients. Because the median time to initiation of IV-tPA was found to be much longer (140 minutes) in IMS I compared with the NINDS trial (90 minutes), IMS II was planned to repeat the IMS I study with one important exception, ie, whenever possible, a more advanced EKOS microinfusion catheter would be used to deliver the tPA into the clot to gauge the efficacy of ultrasound technology in delivering the tPA into the blood clot. The rationale is as follows.

> To dissolve a blood clot, thrombolytic drugs must bind with plasminogen activation receptor sites, which are located in the tightly bound fibrin of a blood clot. Locally delivered, low-energy ultrasound helps temporarily loosen and separate the fibrin, which makes the clot more permeable and increases the availability of more plasminogen activation receptor sites. At the same time, the ultrasound helps drive the thrombolytic agents deep into the blood clot to accelerate the thrombolysis and ultimately dissolve the clot.

The results of IMS II, which involved 73 patients located in 13 different centers, "provided additional evidence that a combined IV/IA approach is a promising treatment for ischemic stroke patients when compared with IV administration of tPA alone and demonstrated for the first time that ultrasound-assisted drug delivery might lead to better results than treatment with a standard microcatheter." The positive results of IMS II led to the design and execution of IMS III, a randomized trial to further investigate the efficacy of a combined IV/IA therapy versus IV-tPA, randomizing patients in a 2:1 ratio within 3 hours after symptom onset. The primary outcome measure was mRS of 0 to 2 at 90% (scores range from 0 to 6, with higher scores indicating greater disability). The trial was planned to involve 900 patients treated at 50 centers and began in August 2006. It was "stopped early because of futility after 656 participants had undergone randomization" as the primary outcome measure "did not differ significantly according to treatment (40.8% with endovascular therapy and 38.7% with IV-tPA),"[16] in which "endovascular" refers to the IA delivery. Subsequent discussions[17] in the same journal issue of this and two other trials on endovascular treatment of acute ischemic stroke patients raised issues such as their lack of adjustments "for the use of sedation, anesthesia, or both" and their use of first-generation devices, whereas "recent randomized studies have clearly shown that stent retrievers were more efficacious in the Merci device," and the patient heterogenity in their studies (eg, "large and small distal-vessel occlusions" for the patients included).

Lansberg et al[18] described how the Diffusion and Perfusion Imaging Evaluation for Understanding Stroke Evolution (DEFUSE) 3 trial is designed to address these issues and used the IMS III data to simulate the operating characteristics of the enrichment trial design. They argue that "adaptive designs can allow modification of various aspects of a study, including the sample size, the treatment dose, the randomization algorithm, and the inclusion/exclusion criteria," and that "sample size reestimation is another adaptation that is commonly implemented in stroke trials," with the Insulin Resistance Intervention after Stroke (IRIS) trial being a recent example of a trial that increased its sample size because of lower than expected rates of recurrent stroke. They note that "trials with overly exclusive criteria are at the risk of slow enrollment and their results may lack generalizability" while "trials with overly inclusive enrollment criteria are at risk of being underpowered to demonstrate a treatment effect," due to dilution of the effect size. As in the case of heterogenous patients in IMS III, they use the Lai-Lavori-Liao design that we have summarized in Table 2 with some modifications due to funding constraint that will be described in the next section. The simulation studies of the design by Lansberg et al[18] were based on data from subjects enrolled in IMS III who had an arterial occlusive lesion of the internal carotid artery or middle cerebral artery on their baseline computed tomographic or magnetic resonance angiogram. Their results demonstrate that the DEFUSE 3 design can be a highly efficient method to test the effect of endovascular stroke treatment, yielding a substantial increase in power when compared with a fixed trial design when the treatment effect differs among subgroups in a predicted pattern, while having similar power compared with a fixed trial design in the case of relatively homogeneous treatment effect across subgroups.

Albers et al[19] provided further details and the background of the DEFUSE 3 trial, which involves the Stanford team of investigators in earlier DEFUSE studies and the IMS III investigators. The methods and results of the first DEFUSE study, which enrolled 74 patients between 2001 and 2005, were reported by of Albers et al[20] that also provided the scientific background:

> Early reperfusion of ischemic brain tissue in acute stroke patients can salvage hypoperfused tissue and improve neurological outcome. Currently, the only approved pharmacological therapy for stroke treatment is tPA administered intravenously within 3 hours of symptom onset. If the treatment window for effective reperfusion therapy can be expanded, considerably more stroke patients would be eligible for therapy. Unfortunately, controlled trials of tPA administered beyond 3 hours have not demonstrated significant benefits. This failure may have resulted from inclusion of patients who were unlikely to benefit from reperfusion therapy because they had minimal salvageable ischemic brain tissue or were at high risk for reperfusion-related complications. Recent observations suggest that new magnetic resonance imaging (MRI) techniques have the potential to identify patients who are optimal candidates for reperfusion therapies in extended time windows. A perfusion/diffusion mismatch has been proposed as a surrogate for the ischemic penumbra, and patients with a mismatch are hypothesized to be more likely to benefit from early reperfusion than patients with other MRI patterns. Therefore, to help clarify whether clinical trials of reperfusion therapies that select patients with specific baseline MRI profiles are likely to be more successful than conventional stroke trials, we tested the hypothesis that patients with predefined MRI profiles would demonstrate a differential clinical response after successful early reperfusion.

This study was followed in 2008 by a prospective cohort study DEFUSE 2 involving 138 patients, including 110 who had catheter angiography and among whom 104 had an MRI profile and 99 could be assessed for reperfusion. The DEFUSE 2 showed that target mismatch patients who had early reperfusion after endovascular stroke treatment had more favorable clinical outcomes, but found no association between reperfusion and favorable outcomes in patients without target mismatch, suggesting that a RCT of endovascular treatment for patients with the target mismatch profile "is warranted" and thus paving the way for the DEFUSE 3 trial; see Lansberg et al.[21] As pointed out by Albers et al,[19] the aim of DEFUSE 3 is to demonstrate that endovascular therapy plus IV-tPA can reduce the degree of disability 3 months past stroke over IV-tPA alone, "among patients with large vessel anterior circulation occlusion who have a favorable imaging profile on computed tomography perfusion or MRI." The primary endpoint is the distribution of mRS at day 90 while the secondary endpoint is the population of patients with mRS 0 to 2 at day 90. Moreover, "DEFUSE 3 will allow patient selection with both MRI and CT perfusion. Use of the latest generation FDA cleared thrombectomy devices, coupled with strict qualification and oversight criteria for the neurointerventionalists, should result in high rates of reperfusion."

3.2 | Adaptive enrichment design for DEFUSE 3

The DEFUSE 3 design originally used the group sequential enrichment design in Section 2.4, a nested sequence of $J = 6$ subsets of patients, defined by a combination of elapsed time from stroke to start of tPA and an imaging-based estimate of the size of the unsalvageable core region of the lesion. The sequence was defined by cumulating the cells in a two-way (3 volumes $\times$ 2 times) cross-tabulation as described in the work of Lai et al.[2(p195)] In the upper left cell, c_{11}, which consisted of the patients with a shorter time to treatment and smallest core volume, the investigators were most confident of a positive effect, while in the lower right cell c_{23} with the longer time and largest core area, there was less confidence in the effect. The six cumulated groups, $\Pi_1, \ldots, \Pi_6$ give rise to corresponding one-sided null hypotheses, $H_1, \ldots, H_6$ for the treatment effects in the cumulated groups. However, shortly before the final reviews of the protocol for funding were completed, four RCTs of endovascular reperfusion therapy administered to stroke patients within 6 hours after symptom onset demonstrated decisive clinical benefits.[22-24] As a result, the equipoise of the investigators shifted, making it necessary to adjust the intake criteria to exclude patients for whom the new therapy had been proven to work better than the standard treatment. The subset selection strategy became even more central to the design, since the primary question was no longer whether the treatment was effective at all, but for which patients should it be adopted as the new standard of care. Moreover, besides adapting the intake criteria to the new findings, another constraint was imposed by the NIH sponsor, which effectively limited the total randomization to 476 patients. Recall that in the original design, if H_J was accepted at an interim stage, the study would go on to recruit to the maximum sample size in the selected subgroup. The first interim analysis would be scheduled after the 200 patients, and the second interim analysis after an additional 140 patients. The limit on total randomization is an example of a constraint on adaptive design that should be discussed with sponsors and if necessary, included in formal design calculations.

DEFUSE 3 has an executive committee consisting of investigators from Stanford and University of Cincinnati who led the DEFUSE and IMS III trials, a data coordination unit at the Medical University of South Carolina, and an independent data and safety monitoring board (DSMB). Besides examining the unblinded efficacy results prepared by a designated statistician at the data coordination unit which also provides periodic summaries on enrollment, baseline characteristics of enrolled patients, protocol violations, timeliness and completeness of data entry by clinical centers, and safety data. During interim analyses, the DSMB also considers the unblinded safety data, comparing the safety of endovascular plus IV-tPA to that of IV-tPA alone, in terms of deaths, serious adverse events, and incidence of symptomatic intracranial hemorrhage.

3.3 | Early termination of DEFUSE 3

In June 2017, positive results of another trial DWI or CTP Assessment with Clinical Mismatch in the Triage of Wake-Up and Late Presenting Strokes undergoing Neurointervention with Trevo (DAWN), which involved patients and treatments similar to those of DEFUSE 3, were announced. Enrollment in the DEFUSE 3 trial was placed on hold; an early interim analysis of the 182 patients enrolled to date was requested by the sponsor (NIH); see the work of Albers et al[25] that said, "As a result of that interim analysis, the trial was halted because the prespecified efficacy boundary ($P < 0.0025$) had been exceeded." As reported by the aforementioned authors,[25] DEFUSE 3 "was conducted at 38 US centers and terminated early for efficacy after 182 patients had undergone randomization (92 to the endovascular therapy group and 90 to the

medical-therapy group)." For the primary and secondary efficacy endpoints, the results show significant superiority of endovascular plus medical therapies. Concerning subgroup analyses, "the power to assess the response to therapy in subgroup was limited owing to the lower than expected number of patients enrolled as a result of early termination of the trial" and "there was no hetergeneity of the treatment effect in any of the prespecified subgroups."

The DAWN trial "was a multicenter randomized trial with a Bayesian adaptive-enrichment design" and was "conducted by a steering committee, which was composed of independent academic investigators and statisticians, in collaboration with the sponsor, Stryker Neurovascular."[26] The DAWN investigators had conducted previous clinical studies, which involved patients with acute stroke and showed that endovascular thrombectomy had a clinical benefit when it was performed within 6 hours after the onset of stroke symptoms,[27-29] or that certain patients could still benefit from the reperfusion of occluded proximal anterior cerebral vessels performed more than 6 hours after the patient was last known to be well.[30] For details of the trial design, which involves an independent DSMB, see the work of Jovin et al.[31] At the first interim analysis that took place at 31 months after enrolling 200 patients, the trial was terminated because the interim analysis results "met the prespecified criterion for trial discontinuation," which was at least 95% posterior probability of the superiority of thrombectomy plus standard care over standard care alone, with respect to the mean score for disability on the utility weighted mRS at 90 days. "Because enrichment threshold had not been crossed, the (final) analysis included the full population of patients enrolled in the trial," which is similar to the work of Albers et al.[25]

4 | DISCUSSION

In their commentary on precision medicine in 2015, Collins and Varmus[32] argued that time was ripe for broad application of this concept because of recent development of large-scale biologic databases (such as the human genome sequence) powerful methods using proteomics, metabolomics, cellular arrays and even mobile health technology to characterize patients, and computational and statistical tools for analyzing the massive data. They pointed out the need for "more clinical trials with novel designs conducted in adult and pediatric patients and more reliable methods for preclinical testing." A seminal trial in this direction in oncology is the biomarker-integrated approaches of targeted therapy for lung cancer elimination (BATTLE) trial of personalized therapies for nonsmall cell lung cancer. As pointed out by Kim et al[33(p45-46)] concerning the biomarker classifiers, "the signaling pathways and targeted agents were selected on the basis of the highest scientific and clinical interest at the time (2005)" and include epidermal growth factor receptor mutation/copy number amplification, kirsten rat sarcoma gene/B-Raf protein (KRAS/BRAF) mutation, vascular endothelial growth factor/vascular endothelial growth factor/and receptor (VEGF/VGFR) expression, and retinoid X receptor/CyslinD1 expression, together with the recommended targeted agent for each. The BATTLE trial uses an AR scheme to select $K = 4$ treatments for $n = 255$ nonsmall cell lung cancer patients belonging to $J = 5$ biomarker classes, one of which contains patients whose biomarker scores are all negative. Let y_{mjk} denote the indicator variable of desease control, which is defined by progression-free survival at 8 weeks after treatment, of the mth patient in class j receiving treatment k. The AR scheme is based on a Bayesian probit model for $p_{jk} = P(y_{mjk} = 1) = P(\xi_{mjk} > 0)$, where ξ_{mjk} is assumed to be latent normal random variable with variance 1 and mean $\mu_{jk} \sim N(\phi_k, \sigma^2)$ such that $\phi_k \sim N(0, \tau^2)$. Large values of τ^2 in the hierarchical Bayesian model can be used to approximate a vague prior. The posterior mean $\gamma_{jk}^{(t)}$ of p_{jk} given all the observed indicator variables up to time t can be computed by Gibbs sampling. Letting $\hat{\gamma}_{jk}^{(t)} = \max(\gamma_{jk}^{(t)}, 0.1)$, the randomization proportion for a patient in the jth class to receive treatment of this scheme allows suspension of treatment k from randomization to a biomarker subgroup. The BATTLE design, which "allows researchers to avoid being locked into a single static protocol of the trial" that requires large sample sizes for multiple comparisons of several treatments across different biomarker classes, can "yield breakthroughs, but must be handled with care" to ensure that "the risk of reaching a false positive conclusion" is not inflated, as pointed out in an April 2010 editorial in *Nature Reviews in Medicine*, on such designs. The results of the BATTLE trial are reported by Kim et al.[33(pp46-48, p52)] Despite applying the Bayesian approach to AR, "standard statistical methods (used in the results section) included Fisher's exact test for contingency tables and log-rank test for survival data," together with standard confidence intervals based on normal approximations, without adjustments for Bayesian AR and possible treatment suspension, even though Zhou et al[34] have noted earlier that "one known ramification of the AR design is that it results in biased estimates due to dependent samples."

Bhatt and Mehta[35(pp. 71-72)] gave a review of recent developments in adaptive enrichment designs, together with a discussion of methodological, operational, and regulatory issues. Table S4 in the supplementary appendix in the closely related work of Mehta et al[36] listed several targeted therapeutic agents that had been approved in the US for specific subgroups of

patients. They say, "these examples (all from oncology trials[36]) showed the potential of predictive biomarkers to identify patients who are likely to benefit from targeted therapies and to thereby increase the success rate of confirmatory clinical trials." However, they also point out the following.

> At this time, regulatory agencies tend to review proposals for adaptive designs with greater scrutiny than they give to conventional designs. This situation is probably due to limited experience with such designs and serious concern that sponsors will submit poorly conceived designs that may not control the type I error and may actually be less efficient than conventional designs. The leakage of interim results could alter investigator behavior and lead to operational bias. Even if there is no leakage of interim results, the mere knowledge that there has been an adaptive change could potentially change the enrollment and characteristics of the patients after the interim analysis. It is critical to ensure that the sample size at the interim analysis is adequate for making the adaptive decision. If patients are enrolled too rapidly relative to the time needed to observe the primary end point, the planned enrollment might be completed before adequate information is available for an adaptive decision to be taken.

They note that the adaptive enrichment approach "will probably increase in other fields (beyond oncology) as validated biomarkers that predict response or lack of response to therapy emerge" and cite the work of Everett et al[37] on using cardiac troponin concentrations to "identify patients who would benefit from urgent revascularization for acute coronary syndromes" as an example. They note, however, the following difficulties with these studies "in which biomarkers have shown predictive capabilities (but which) were not designed for this purpose."

> Even in well-controlled phase 3 trials, the biomarker component of the analysis is often performed retrospectively or the trials restricted enrollment to the targeted subgroups from the start. However, the FDA 2012 Guidance recommends that, even in cases in which there is a strong biologic basis for a therapy to target a particular genetic marker, it is desirable to enroll patients in whom the marker is absent to show sensitivity in patients who have the marker and lack of sensitivity in patients who do not have the marker.

In Sections 2.4 and 2.5, we have described group sequential enrichment designs that have asymptotically efficient power and expected sample size properties while preserving the type I error probability. These designs, therefore, have addressed some of the issues in the preceding paragraph on Bhatt and Mehta's review. Section 3 illustrates the experience with these enrichment designs in the case of testing endovascular treatment for ischemic stroke patients. By reviewing how DEFUSE 3 evolved from previous DEFUSE and IMS trials, we show the evolutionary nature of precision medicine, thereby providing a concrete example of adaptive enrichment design for a confirmatory nononcology trial. Lai et al[9] have pointed out two challenges in designing clinical trials to test biomarker-based personalized therapies for their regulatory approval. One is "what to do for patients in whom the biomarkers predict resistance to all of the drugs tested." Depending on the specific enrichment mechanism used, the enrichment design also leaves open (i) the question of generalizability of the result and how the drug will work in a broader population and (ii) the question of how much data are needed, before or after regulatory approval, for the "nonselected" group. The December 2012 FDA Guidance for Industry says, "In general, then, FDA is prepared to approve drugs studied primarily or even solely in enriched populations and will seek to ensure truthful labeling that does not overstate either the likelihood of a response or the predictiveness of the enrichment factor. However, the extent of data that should be available on the nonenriched subgroup should always be considered. Postmarket commitments or requirements may be requested to better define the full extend of a drug's effect (including efficacy and safety studies and trials in a broader populations)." Another challenge is that the classifiers used in the personalized therapy "cannot be expected to be perfect straight from the box" and may be out of date by the time a validation study is completed, given the rapid pace of biomarker and biotechnology development. Both challenges argue for a flexible group sequential design that can adapt not only to endogenous information from the trial but also to exogenous information from advances in precision medicine. Early termination of DEFUSE 3 described in Section 3.3 provides a concrete example of this exogenous information and how it evolves over time because of technological advances. Recall from Section 3.1 that DEFUSE 2 did not find significant improvement of outcomes in the entire patient population, leading to the adaptive enrichment design of DEFUSE 3 in Section 3.2. However, both DAWN and DEFUSE 3 have terminated after the first interim analysis, showing favorable outcomes for endovascular stroke treatment (which has also improved over time together with perfusion imaging); see also Albers et al.[25(pp. 716-717)]

ACKNOWLEDGEMENTS

Lai's research was supported by National Science Foundation grant DMS 1407828 and National Institutes of Health grant 1P30 CA124435. Lavori's research was supported by National Institutes of Health grant 1UL1 RR025744. Tsang's research was supported by the President's Fund of the Chinese University of Hong Kong, Shenzhen.

ORCID

Ka Wai Tsang http://orcid.org/0000-0002-9383-4455

REFERENCES

1. Simon N, Simon R. Adaptive enrichment designs for clinical trials. *Biostatistics*. 2013;14(4):613-625.
2. Lai TL, Lavori PW, Liao OYW. Adaptive choice of patient subgroup for comparing two treatments. *Contemp Clin Trials*. 2014;39:191-200.
3. Lai TL, Shih MC. Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika*. 2004;91(3):507-528.
4. Bartroff J, Lai TL, Shih MC. *Sequential Experimentation in Clinical Trials*. New York, NY: Springer; 2013.
5. Bartroff J, Lai TL. Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Statist Med*. 2008;27(10):1593-1611.
6. Jennison C, Turnbull B. Adaptive and nonadaptive group sequential tests. *Biometrika*. 2006;93(1):1-21.
7. Lai TL, Lavori PW, Liao OYW, Narasimhan B, Tsang KW. ASSISTant: Adaptive Subgroup Selection in Group Sequential Trials. R package version 1.2-3. 2016. https://CRAN.R-project.org/package=ASSISTant
8. Lai TL, Liao OY. Efficient adaptive randomization and stopping rules in multi-arm clinical trials for testing a new treatment. *Seq Anal*. 2012;31:441-457.
9. Lai TL, Lavori PW, Shih MC, Sikic B. Clinical trial designs for testing biomarker-based personalized therapies. *Clin Trials*. 2012;9:141-154.
10. Berry SM, Carlin BP, Lee J, Muller P. *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC; 2010.
11. Lai TL, Lavori PW, Tsang KW. Adaptive design of confirmatory trials: advances and challenges. *Contemp Clin Trials*. 2015;45:93-102. Trials 10th Anniversary Special Issue 45 Part A.
12. Lai TL, Liao OYW, Kim DW. Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Contemp Clin Trials*. 2013;36:651-663.
13. Lai TL, Robbins H. Asymptotically efficient adaptive allocation rules. *Adv Appl Math*. 1985;6(1):4-22.
14. Lai TL. Adaptive treatment allocation and the multi-armed bandit problem. *Ann Stat*. 1987;15(3):1091-1114.
15. Broderick JP, Tomsick T. The IMS trials. *Endovascular Today*. 2006:23-26.
16. Broderick JP, Palesch YY, Demchuk AM, et al. Endovascular therapy after intravenous t-PA versus t-PA alone for stroke. *N Engl J Med*. 2013;368:893-903.
17. Ciccone A, Valvassori L, Nichelatti M, et al. Endovascular treatment for acute ischemic stroke. *N Engl J Med*. 2013;368:904-913.
18. Lansberg MG, Bhat NS, Yeatts SD, et al. Power of an adaptive trial design for endovascular stroke studies: simulations using IMS (Interventional Management of Stroke) III data. *Stroke*. 2016;47:2931-2937.
19. Albers GW, Lansberg MG, Kemp S, et al. A multicenter randomized controlled trial of endovascular therapy following imaging evaluation for ischemic stroke (DEFUSE 3). *Int J Stroke*. 2017;12:896-905.
20. Albers GW, Thijs VN, Wechsler L, et al. Magnetic resonance imaging profiles predict clinical response to early reperfusion: the diffusion and perfusion imaging evaluation for understanding stroke evolution (DEFUSE) study. *Ann Neurol*. 2006;60:508-517.
21. Lansberg MG, Straka M, Kemp S, et al. MRI profile and response to endovascular reperfusion after stroke (DEFUSE 2): a prospective cohort study. *Lancet Neurol*. 2012;11:860-867.
22. Berkhemer OA, Fransen PS, Beumer D, et al. A randomized trial of intraarterial treatment for acute ischemic stroke. *N Engl J Med*. 2015;372:11-20.
23. Campbell BC, Mitchell PJ, Kleinig TJ, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. *N Engl J Med*. 2015;372:1009-1018.
24. Goyal M, Demchuk AM, Menon BK, et al. Randomized assessment of rapid endovascular treatment of ischemic stroke. *N Engl J Med*. 2015;372:1019-1030.
25. Albers GW, Marks MP, Kemp S, et al. Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N Engl J Med*. 2018;378:708-718.
26. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N Engl J Med*. 2018;378:11-21.
27. Saver JL, Goyal M, Bonafe A, et al. Stent-retriever thrombectomy after intravenous t-PA vs. t-PA alone in stroke. *N Engl J Med*. 2015;372:2285-2295.
28. Saver JL, Goyal M, Lugt A, et al. Time to treatment with endovascular thrombectomy and outcomes from ischemic stroke: A meta-analysis. *JAMA*. 2016;316:1279-1289.

29. Jovin TG, Chamorro A, Cobo E, et al. Thrombectomy within 8 hours after symptom onset in ischemic stroke. *N Engl J Med*. 2015;372:2296-2306.

30. Jovin TG, Liebeskind DS, Gupta R, et al. Imaging-based endovascular therapy for acute ischemic stroke due to proximal intracranial anterior circulation occlusion treated beyond 8 hours from time last seen well. *Stroke*. 2011;42:2206-2211.

31. Jovin TG, Saver JL, Ribo M, et al. Diffusion-weighted imaging or computerized tomography perfusion assessment with clinical mismatch in the triage of wake up and late presenting strokes undergoing neurointervention with Trevo (DAWN) trial methods. *Int J Stroke*. 2017;12:641-652.

32. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med*. 2015;372(9):793-795.

33. Kim ES, Herbst RS, Wistuba II, et al. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov*. 2011;1:44-53.

34. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer–a step toward personalized medicine. *Clin Trials*. 2008;5:181-193.

35. Bhatt DL, Mehta C. Adaptive designs for clinical trials. *N Engl J Med*. 2016;375:65-74.

36. Mehta CR, Schäfer H, Daniel H, Irle S. Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statist Med*. 2014;33:4515-4531.

37. Everett BM, Brooks MM, Vlachos HE, Chaitman BR, Frye RL, Bhatt DL. Troponin and cardiac events in stable ischemic heart disease and diabetes. *N Engl J Med*. 2015;373:610-620.

How to cite this article: Lai TL, Lavori PW, Tsang KW. Adaptive enrichment designs for confirmatory trials. *Statistics in Medicine*. 2019;38:613–624. https://doi.org/10.1002/sim.7946