Annals of Mathematical Sciences and Applications
Volume 4, Number 2, 413–438, 2019

Statistical science in information technology and precision medicine

Tze Leung Lai*†, Anna Choi, and Ka Wai Tsang

Broadly speaking, statistical science is the science that deals with the collection and analysis of data. Although its methodological developments are rooted in mathematical sciences, particularly probability and optimization, it has transcended mathematical sciences as a discipline by growing in multidisciplinary platforms. Examples in the United States and Canada are biostatistics and epidemiology, biomedical data science/informatics, statistics in information engineering/data science, financial engineering and statistics, statistics and actuarial science. In China, mathematics and statistics are first-class disciplines at Peking University and Nankai University, whereas statistics but not mathematics is listed as a first-class discipline at Renmin University, Xiamen University and Shanghai University of Finance and Economics. Herein we first give an overview of statistical science in online experimentation, web-based personalized marketing, recommender systems, and precision medicine and health. We then develop new statistical methods to address two challenging problems in online experimentation and precision medicine.

Keywords and phrases: A/B testing, adaptive subgroup selection, enrichment designs, familywise error rate in multiple testing, group sequential methods, hybrid resampling, online experiments.

1. Introduction

We begin by describing the background of this paper, which is related to the Forum on Mathematical Sciences at the Frontiers of Science and Technology on the opening day of 8th International Congress of Chinese Mathematicians (June 9–13, 2019). The forum featured lectures on discrete geometric analysis for materials research and mathematicians in the development of science and technology from a European perspective, by Motoko Kotani and Jean-Pierre Bourguignon, respectively, followed by panel discussions on Statistics

*Corresponding author.
†T.L. Lai is supported by NSF DMS-1811818.

and Artificial Intelligence. The first author attended the forum, in which he was a panelist and which led him to reflect on his work with other two coauthors in online experimentation, A/B testing and recommender systems – topics that can be considered as some of the frontiers in information technology (IT). This section gives an overview of this and related work not only at the IT frontiers but also at the frontiers of biomedicine and healthcare. It also provides the background for the new methods and results in Section 2 on group sequential multiple testing in online experimentation and precision medicine. Further discussion and concluding remarks are given in Section 3.

1.1. Online experimentation, sequential analysis, multiple testing

Online controlled experiments conducted on internet traffic are important for data-driven decision making at many IT companies, including Amazon, eBay, Facebook, Google, Microsoft, MSN.com, Optimizely, Yahoo and Zynga. A/B testing is arguably the most common type of online experiments. A small fraction of users are randomly selected to experience a new treatment (e.g., changing the color of the logo on a web browser) within a predetermined period of time. During the same period, another small fraction of users are also randomly selected to serve as the control group that is not exposed to the new treatment. Hundreds of metrics will then be computed and compared between the treatment and control groups. The metrics are related to the experimentation objectives, e.g., Amazon's e-ecommerce-revenue per user, revenue per session, etc.; MSN's content – number of visitors, clicks per visitor, clicks per session, etc.; Bing's search – searches per session, sessions per user, etc.; Microsoft's support – length of session, session success, etc. The scope of new treatments to be tested includes changes in text/content/design, colors and fonts for user interface, options for user support, email and other customer contact channels, apps, algorithms and underlying codebase, etc.

For a particular metric, say clicks per user, letting $X_1, \ldots, X_{n_t}$ be the number of clicks for the users in the treatment group and $\bar{X}$ be their sample mean, $Y_1, \ldots, Y_{n_c}$ be the number of clicks for the users in the control group with sample mean $\bar{Y}$, $\bar{X} - \bar{Y}$ is approximately normal with mean 0 under the null hypothesis of no mean treatment difference between the new and old web designs, and $\mathrm{Var}(\bar{X} - \bar{Y})$ can also be consistently estimated. Because hundreds of such tests are performed for the hundreds of metrics, multiplicity adjustments need to be made to guarantee that the type I error of multiple hypothesis testing, either in the form of familywise error rate

(FWER) or the weaker false discovery rate [11, pp. 158–159], does not exceed some prescribed level. In Section 2, after giving an overview of group sequential multiple testing and adaptive design of clinical trials, we further develop some recent advances in these areas for A/B testing in online experiments and for the development and testing of biomarker-guided treatment strategies in confirmatory clinical trials. Further discussion, additional applications and concluding remarks are given in Section 3.

1.2. Classical and contextual multi-armed bandits, precision medicine and recommender systems

The K-armed bandit problem, introduced by Robbins [41] for the case $K = 2$, is prototypical in the area of stochastic adaptive control that addresses the dilemma between "exploration" (to generate information about the unknown system parameters needed for efficient system control) and "exploitation" (to set the system inputs that attempt to maximize the expected rewards from the outputs). Robbins considered the problem of which of K populations to sample from sequentially in order to maximize the expected sum $E(\sum_{i=1}^{N} Y_i)$. If the population with the largest mean were known, then obviously one should sample from it to receive expected reward $N\mu^*$, where $\mu^* = \max_{1\le k\le K} \mu_k$ and μ_k is the mean of the kth population, which is assumed to be finite. By using the law of large numbers and infrequent forced sampling from all populations, he showed that $\lim_{N\to\infty} N^{-1}E(\sum_{i=1}^{N} Y_i) = \mu^*$ could still be attained. Important advances were subsequently made by Bellman [8], Chernoff and Ray [12], Gittins [21], Whittle [48], before Robbins revisited the problem with Lai in [34], which we review below.

Let $\mathcal{F}_t$ be the σ-algebra of events up to time t. An allocation rule $\phi = (\phi_1, \ldots, \phi_N)$ is said to be "adaptive" if its choice of the population ϕ_t to sample from at time t is $\mathcal{F}_{t-1}$-measurable, i.e., $\{\phi_t = k\} \in \mathcal{F}_{t-1}$ for $k = 1, \ldots, K$. Suppose Y_t has density function f_{θ_k} (with respect to some measure m) when $\phi_t = k$, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$. Then

$$E_{\boldsymbol{\theta}}\Big(\sum_{t=1}^{N} Y_t\Big) = \sum_{t=1}^{N}\sum_{k=1}^{K} E_{\boldsymbol{\theta}}\{E_{\boldsymbol{\theta}}(Y_t I_{\{\phi_t=k\}}|\mathcal{F}_{t-1})\} = \sum_{k=1}^{K} \mu(\theta_k) E_{\boldsymbol{\theta}} T_N(k), \tag{1.1}$$

where $\mu(\theta) = \int_{-\infty}^{\infty} y f_\theta(y) dm(y)$ and $T_N(k) = \sum_{t=1}^{N} I_{\{\phi_t=k\}}$ is the total sample size from population k. Hence maximizing the expected sum $E_{\boldsymbol{\theta}}(\sum_{t=1}^{N} Y_t)$

is equivalent to minimizing the regret, or shorfall from $N\mu^*(\boldsymbol{\theta})$:

(1.2)
$$R_N(\boldsymbol{\theta}) = N\mu^*(\boldsymbol{\theta}) - E_{\boldsymbol{\theta}}(\sum_{t=1}^{N} Y_t) = \sum_{k:\mu(\theta_k)<\mu^*(\boldsymbol{\theta})} \{\mu^*(\boldsymbol{\theta}) - \mu(\theta_k)\} E_{\boldsymbol{\theta}} T_N(k),$$

in which the second equality follows from (1.1) and shows that the regret is a weighted sum of expected sample sizes from inferior populations. Making use of this representation in terms of expected sample sizes, Lai and Robbins [34] derive an the asymptotic lower bound, as $N \to \infty$, for the regret $R_N(\boldsymbol{\theta})$ of uniformly good adaptive allocation rules:

$$R_N(\boldsymbol{\theta}) \geq (1 + o(1)) \sum_{k:\mu(\theta_k)<\mu^*(\boldsymbol{\theta})} \frac{\mu(\theta^*) - \mu(\theta_k)}{I(\theta_k, \theta^*)} \log N, \tag{1.3}$$

where $\theta^* = \theta_{K(\boldsymbol{\theta})}$ and $K(\boldsymbol{\theta}) = \arg\max_{1\leq k\leq K} \mu(\theta_k)$, $I(\boldsymbol{\theta}, \lambda) = E_\theta\{\log(f_\theta(Y)/f_\lambda(Y))\}$ is the Kullback-Leibler information number, and an adaptive allocation rule is called "uniformly good" if $R_N(\boldsymbol{\theta}) = o(N^a)$ for all $a > 0$ and $\boldsymbol{\theta}$.

Putting independent priors G_j on θ_j, the infinite-horizon problem of maximizing

$$E(\sum_{t=1}^{\infty} \beta^{t-1} Y_t) = \int \cdots \int E_{\boldsymbol{\theta}}(\sum_{t=1}^{\infty} \beta^{t-1} Y_t) dG_1(\theta_1) \ldots dG_K(\theta_K), \tag{1.4}$$

with $0 < \beta < 1$ can be analyzed by Markovian dynamic programming (MDP) introduced by Bellman [8], who also introduced the term "multi-armed bandit" for an imagined slot machine (which typically "robs" the player's money) with K levers, Gittins [21] and Whittle [48] used MDP to show that the Bayes rule that maximizes (1.4) in the index rule ϕ^* that samples at stage $n+1$ from the population k_n^* will be the largest dynamic allocation index $M(G_{k_n^*|T_n(k)})$ over the posterior distributions $G_{k|T_n(k)}$. The dynamic allocation index, subsequently called *Gittins index*, of a distribution G is the infimum of solutions M of the equation

$$\sup_{\tau\geq 0} E\{\sum_{t=0}^{\tau-1} \beta^t E[\mu(\theta)|Y_1, \ldots, Y_t] + M \sum_{i=\tau}^{\infty} \beta^t\} = M \sum_{t=0}^{\infty} \beta^t, \tag{1.5}$$

where $\sup_{\tau \geq 0}$ is the supremum over stopping times τ. As pointed out by Whittle [48], the right-hand side of (1.5) can be interpreted as retiring immediately to collect retirement rewards M now and over subsequent periods with discount factor β per period, which the left-hand side of (1.5) refers to finding the optimal stopping time for retirement, hence (1.5) can be interpreted as the value of M for which one is indifferent to retiring now or optimally hereafter. Although $M(G)$ may be difficult to compute, the index rule represents a major advance as it reduces a K-dimensional stochastic control problem to K optimal stopping problems. For the finite-horizon Bayes rule, Chernoff and Ray [12] considered the one-armed bandit problem of choosing between sampling from a population with 0 reward and another population with $N(0,1)$ rewards, in which θ has a normal prior distribution. The Bayes procedure in this case samples from Π_1 until $T^* = \min\{n \leq N : \sum_{i=1}^{n} Y_i + a_{n,N} \leq 0\}$. Chernoff and Ray used Brownian motion and an associated free boundary problem for the PDE to determine asymptotic expansions for $a_{n,N} \approx h(n/N)$.

Chang and Lai [10] developed asymptotic expansions for the Gittins index M_c define by (1.5) as $\beta = e^{-c} \to 1$ and found them to agree with $a_{n,N}$ as $n/N \to 0$. Noting that Chernoff and Ray's rule is tantamount to sampling from the $N(0,1)$ population or the zero-reward population according to whether $U_{n,N} > 0$ or $U_{n,N} \leq 0$, where $U_{n,N} = \bar{Y}_n + n^{-1}a_{n,N}$ is an upper confidence bound for θ, Lai [29] proposed a general upper confidence bound (UCB) rule for the finite-horizon problem in the exponential family of densities $f_\theta(y) = \exp\{\theta y - \Psi(\theta)\}$ with respect to some dominating measure: Sample at stage $n+1$ from the population that has the largest $U_{k,T(k)}$, where $U_{j,t} = \inf\{\lambda \geq \hat{\theta}_{j,t} : 2tI(\hat{\theta}_{j,t}, \lambda) \geq h^2(t/N)\}$, $\hat{\theta}_{j,t}$ is the maximum likelihood estimate and $I(\theta, \lambda)$ is the Kullback-Leibler information number. $U_{j,t}$ is therefore an upper confidence bound for θ obtained by inverting generalized likelihood ratio (GLR) tests in the exponential family. Lai [29] has shown that the UCB rule attains the assumption lower bound (1.3) for the regret $R_N(\theta)$ of uniformly good rules at every $\boldsymbol{\theta}$, and that it also attains asymptotically the Bayes regret as $N \to \infty$. The Bayes regret is of the order $C(\log N)^2$ when the prior distribution for $\boldsymbol{\theta}$ has a positive continuous density over $\theta_k \in (\theta_k^* - \delta, \theta_k^* + \delta)$ for $1 \leq k \leq K$, where $\theta_k^* = \max_{j \neq k} \theta_j$ [29].

New applications and advances in IT and biomedicine in the new millenium have led to the development of *contextual multi-armed bandits*, also called bandits with side information or covariates, while the classical multi-armed bandits reviewed above are often referred to as "context-free" bandits. Personalized marketing (e.g., Amazon) uses web sites to track a customer's purchasing records and thereby to market products that are individualized

for the customer. Recommender systems select items such as movies (e.g., Netflix) and news (e.g., Yahoo) for users based on the users' and items' features (covariates). Li et al. [37] model the click probability of a news article as a function, estimated by machine learning methods, of the user's and article's features. They apply a UCB-type policy targeted towards maximizing the click probability, but provide no theoretical analysis or simulation study of the performance of the policy. Tang et al. [44] describe web-based personalization to show online ads for each user, with the goal of maximizing "its effectiveness, measured in terms of click-through rate or total revenue." They formulate the optimization problem as a contextual multi-armed bandit problem with a page request of each user as side (covariate) information and layouts of advertisements available for the requested page as arms. In Section 2.2 we describe applications of contextual multi-armed bandits to personalized treatment strategies (also called precision medicine) for stroke or cancer patients.

Whereas classical K-armed bandits reviewed above aim at choosing ϕ_i sequentially so that $E_{\boldsymbol{\theta}}(\sum_{i=1}^N Y_i)$ is as close as possible to $N \max_{1\le k\le K} \mu_k$, contextual bandits basically replace $N\mu_k$ by $\sum_{i=1}^N \mu_k(\boldsymbol{x_i})$, where $\boldsymbol{x_i}$ is the covariate of the ith subject, noting that analogous to (1.1),

$$E_{\boldsymbol{\theta}}(Y_i) = \sum_{k=1}^K E_{\boldsymbol{\theta}}\{E_{\boldsymbol{\theta}}(Y_i I_{\{\phi_i=k\}}|\boldsymbol{x_i}, \mathcal{F}_{t-1})\} = \sum_{k=1}^K E_{\boldsymbol{\theta}}(\mu_k(\boldsymbol{x_i}) I_{\{\phi_i=k\}}). \tag{1.6}$$

Assuming $\boldsymbol{x_i}$ to be i.i.d. with distribution G, we can define $g^*(x) = \arg\max_{1\le k\le K} \mu(\theta_k, x)$, $\theta^*(x) = \theta_{j^*(x)}$ and the regret

$$\begin{aligned} R_N(\boldsymbol{\theta}, B) &= N \int_B \mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) dG(\boldsymbol{x}) - \sum_{i=1}^N \sum_{k=1}^K \int_B \mu(\theta_k, \boldsymbol{x}) E_{\boldsymbol{\theta}}(I_{\{\phi_i=k\}}) dG(\boldsymbol{x}) \\ &= \sum_{k=1}^K \int_B \{\mu(\theta^*(\boldsymbol{x}), \boldsymbol{x}) - \mu(\theta_k, \boldsymbol{x})\} E_{\boldsymbol{\theta}} T_N(k, \boldsymbol{x}) dG(\boldsymbol{x}) \end{aligned} \tag{1.7}$$

for Borel subsets B of $supp(G)$, where $T_N(k, B) = \sum_{i=1}^N I_{\{\phi_i=k, \boldsymbol{x_i}\in B\}}$, noting that the measure $E_{\boldsymbol{\theta}} T_N(k, \cdot)$ is absolutely continuous with respect to G, hence $E_{\boldsymbol{\theta}} T_N(k, \boldsymbol{x})$ in (1.7) is its Radom-Nikodym derivative with respect to G. The UCB rule (index policy) in classical bandit theory basically samples from an inferior arm k until the sample size satisfies the asymptotic lower bound for $E_{\boldsymbol{\theta}} T_N(k)$. For contextual bandits, an arm that is inferior at $\boldsymbol{x}$ may be the best at $\boldsymbol{x'}$. Therefore the uncertainty in the sample mean reward at

$\boldsymbol{x}_t$ does not need to be immediately reduced, and adaptive randomization (rather than UCB rule) can yield an asymptotically optimal policy, as will be illustrated in Sections 2.2 and 2.3.

2. Group sequential multiple testing in biomedicine and IT

We begin with group sequential design of comparative clinical trials in Section 2.1 that also describes valid p-values and confidence intervals for statistical analysis of the data. Because of the lack of information on both the magnitude and the sampling variability of the treatment effect of a new treatment at the design stage, there has been increasing interest from the biopharmaceutical industry in adaptive designs that can adapt to the information collected during the course of the trial, as reviewed in Section 2.1. The past decade witnessed major developments in innovative designs of confirmatory clinical trials, and adaptive designs represent the most active area of these developments. Section 2.2 gives an overview of adaptive enrichment designs for subgroup selection in precision medicine and introduces new methods for the development and testing of biomarker-guided treatment strategies in confirmatory clinical trials. Section 2.3 further develops similar ideas in the context of A/B testing in online experiments.

2.1. Design and analysis of group sequential comparative trials

As pointed in [6, p. 77], in standard designs of clinical trials comparing a new treatment with a control (which is a standard treatment or placebo), the sample size is determined by the power at a given alternative, but it is often difficult to specify a realistic alternative in practice because of lack of information on the magnitude of the treatment effect difference before actual clinical trial data are collected. On the other hand, many trials have Data and Safety Monitoring Committees (DSMCs) who conduct periodic reviews of the trial, particularly with respect to incidence of treatment-related adverse events, hence one can use the trial data at interim analyses to estimate the effect size. This is the idea underlying group sequential trials in the late 1970s, and one such trial was the Beta-blocker Heart Attack Trial (BHAT) that was terminated in October 1981, prior to its prescheduled end in June 1982 [6, pp. 3–4]. BHAT, which was a multicenter, double-blind, randomized placebo-controlled trial to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction (MI), drew immediate attention to the benefits of sequential methods not because it reduced the number of patients but because it shortened a 4-year

study by 8 months, with positive results for a long-awaited treatment for MI patients. The success story of BHAT paved the way for major advances in the development of group sequential methods in clinical trials and for the widespread adoption of group sequential design. Sections 3.5 and 4.2 of [6] describe the theory developed by Lai and Shih [35] for nearly optimal group sequential tests in exponential families to provide a definitive method amidst the plethora of group sequential stopping boundaries that were proposed in the two decades after BHAT, as reviewed in [6, Chapter 6].

Lai and Shih's theory is based on (a) asymptotic lower bounds for the sample sizes of group sequential tests that satisfy prescribed type I and type II error probability bounds, and (b) group sequential GLR tests with modified Haybittle-Peto boundaries that can be shown to attain these bounds. Noting that the efficiency of a group sequential test depends not only on the choice of the stopping rule but also on the test statistics, Lai and Shih use generalized likelihood ratio statistics that have been shown in earlier works of Lai to have asymptotically optimal properties for sequential testing in one-parameter exponential families [6, Section 3.7] and can be readily extended to multiparameter exponential families for which the type I and type II errors are evaluated at $u(\boldsymbol{\theta}) = u_0$ and $u(\boldsymbol{\theta}) = u_1$, respectively, where $u : \Theta \to \mathbb{R}$ is a continuously differentiable function on the natural parameter space Θ such that Kullback-Leibler information number $I(\boldsymbol{\gamma}, \boldsymbol{\theta})$ is increasing in $|u(\boldsymbol{\theta}) - u(\boldsymbol{\gamma})|$ for every $\boldsymbol{\gamma}$ [6, Section 4.2.4]. An important consideration in this approach is the choice of the alternative θ_1 (in the one-parameter case, or u_1 in the multiparameter exponential families). To test $H_0 : \theta \le \theta_0$, suppose the significance level is α and no more than M observations are to be taken because of funding and administrative constraints on the trial. The FSS (fixed sample size) test that rejects H_0 if $S_M \ge c_\alpha$ has maximal power at any alternative $\theta > \theta_0$. Although funding and administrative considerations often play an important role in the choice of M, justification of this choice in clinical trial protocols is typically based on some prescribed power $1 - \beta$ at an alternative $\theta(M)$ "implied" by M. The implied alternative is defined by that M and can be derived from the prescribed power $1 - \beta$ at $\theta(M)$. It is used to construct the futility boundary in the modified Haybittle-Peto group sequential test [6, pp. 81–85].

Using Lai and Shih's theory of modified Haybittle-Peto group sequential tests, Bartroff and Lai [3, 4] developed a new approach to adaptive design of clinical trials. In standard clinical trial designs, the sample size is determined by the power at a given alternative, but in practice, it is often difficult for investigators to specify a realistic alternative at which sample size determination can be based. Although a standard method to

address this difficulty is to carry out a preliminary pilot study, the results from a small pilot study may be difficult to interpret and apply, as pointed out by Wittes and Brittain [47], who proposed to treat the first stage of a two-stage clinical trial as an internal pilot from which the overall sample size can be re-estimated. The specific problem considered by [47] as an example of internal pilots actually dated back to Stein's two-stage procedure [43], introduced in 1945 for testing hypothesis $H_0 : \mu_X = \mu_Y$ versus the two-sided alternative $\mu_X \neq \mu_Y$ for the means of two independent normal distributions with common, unknown variance σ^2. In its first stage, Stein's procedure samples n_0 observations from each of the two normal distributions and computes the usual unbiased estimate s_0^2 of σ^2. The second stage samples $n_1 = n_0 \vee [\left(t_{2n_0-2,\alpha/2} + t_{2n_0-2,\beta}\right)^2 2s_0^2/\delta^2]$ observations from each population, where α is the prescribed type I error probability, $t_{\nu,\alpha}$ denotes the upper α-quantile of the t-distribution with ν degrees of freedom, and $1 - \beta$ is the prescribed power at the alternatives satisfying $|\mu_X - \mu_Y| = \delta$. The null hypothesis $H_0 : \mu_X = \mu_Y$ is then rejected if $|\bar{X}_{n_1} - \bar{Y}_{n_1}| > t_{2n_0-2,\alpha/2}\sqrt{2s_0^2/n_1}$. Modifications of the two-stage procedure were provided by [47] and [9, 16, 22, 24], which represent the "first generation" of adaptive designs. The second generation of adaptive designs adopts a more aggressive viewpoint of re-estimating the sample size from the estimate of δ (instead of the nuisance parameter σ) based on the first-stage data; see [20, 40].

Assuming normally distributed outcomes with known variances, Jennison and Turnbull [26] introduced adaptive group sequential tests that choose the jth group size and stopping boundary on the basis of the cumulative sample size n_{j-1} and the sample sum $S_{n_{j-1}}$ over the first $j - 1$ groups, and that are optimal in the sense of minimizing a weighted average of the expected sample sizes over a collection of parameter values, subject to prescribed error probabilities at the null and a given alternative hypothesis. They showed how the corresponding optimization problem can be solved numerically by using backward induction algorithms. They also showed in [27] that standard (non-adaptive) group sequential tests with the first stage chosen appropriately are nearly as efficient as their optimal adaptive tests. Earlier they showed in [10] that the adaptive tests proposed in the preceding paragraph performed poorly in terms of expected sample size and power in comparison with the group sequential tests. Tsiatis and Mehta [46] independently came to the same conclusion, attributing this inefficiency to the use of the non-sufficient "weighted" statistic. Bartroff and Lai's new approach to adaptive designs [3, 4], developed in the general framework of multiparameter exponential families, uses efficient generalized likelihood ratio statistics

in this framework and adds a third stage to adjust for the sampling variability of the first-stage parameter estimates that determine the second-stage sample size. The possibility of adding a third stage to improve two-stage designs dated back to Lorden [38]. Whereas Lorden used crude upper bounds for the type I error probability that are too conservative for practical applications, Bartroff and Lai overcame this difficulty by using new methods to compute the type I error probability, and also extended the three-stage test to multiparameter and multi-armed settings, thus greatly broadening the scope of these efficient adaptive designs. Hybrid resampling plays an important role in the statistical analysis of the data generated by these adaptive designs, especially for primary and secondary analysis.

Rosner and Tsiatis [42] developed exact confidence intervals for the mean of a normal distribution with known variance following a group sequential test. Subsequently, Chuang and Lai [13, 14] noted that even though $\sqrt{n}(\bar{X}_n - \mu)$ is a pivot in the case of $X_i \sim N(\mu, 1)$, $\sqrt{T}(\bar{X}_T - \mu)$ is highly non-pivotal for a group sequential stopping time, hence the need for the *exact method* of [42], which they generalized as follows. If $\mathcal{F} = \{F_\theta : \theta \in \Theta\}$ is indexed by a real-valued parameter θ, an exact equal-tailed confidence region can always be found by using the well-known duality between hypothesis tests and confidence regions. Suppose one would like to test the null hypothesis that θ is equal to θ_0. Let $R(\mathbf{X}, \theta_0)$ be some real-valued test statistic. Let $u_\alpha(\theta_0)$ be the α-quantile of the distribution of $R(\mathbf{X}, \theta_0)$ under the distribution F_{θ_0}. The null hypothesis is accepted if $u_\alpha(\theta_0) < R(\mathbf{X}, \theta_0) < u_{1-\alpha}(\theta_0)$. An exact equal-tailed confidence region with coverage probability $1 - 2\alpha$ consists of all θ_0 not rejected by the test and is therefore given by $\{\theta : u_\alpha(\theta) < R(\mathbf{X}, \theta) < u_{1-\alpha}(\theta)\}$. The exact method, however, applies only when there are no nuisance parameters and this assumption is rarely satisfied in practice. To address this difficulty, Chuang and Lai [13, 14] introduced a *hybrid resampling method* that "hybridizes" the exact method with Efron's [18, 19] bootstrap method to construct confidence intervals. The bootstrap method replaces the quantiles $u_\alpha(\theta)$ and $u_{1-\alpha}(\theta)$ by by the approximate quantiles u^*_α and $u^*_{1-\alpha}$ obtained in the following manner. Based on $\mathbf{X}$, construct an estimate $\hat{F}$ of $F \in \mathcal{F}$. The quantile u^*_α is defined to be α-quantile of the distribution of $R(\mathbf{X}^*, \hat{\theta})$ with $\mathbf{X}^*$ generated from $\hat{F}$ and $\hat{\theta} = \theta(\hat{F})$, yielding the confidence region $\{\theta : u^*_\alpha < R(\mathbf{X}, \theta) < u^*_{1-\alpha}\}$ with approximate coverage probability $1 - 2\alpha$. For group sequential designs, the bootstrap method breaks down because of the absence of an approximate pivot, as shown in [13]. The hybrid confidence region is based on reducing the family of distributions $\mathcal{F}$ to another family of distributions $\{\hat{F}_\theta : \theta \in \Theta\}$,

which is used as the "resampling family" and in which θ is the unknown parameter of interest. Let $\hat{u}_\alpha(\theta)$ be the α-quantile of the sampling distribution of $R(\mathbf{X}, \theta)$ under the assumption that $\mathbf{X}$ has distribution $\hat{F}_\theta$. The hybrid confidence region results from applying the exact method to $\{\hat{F}_\theta : \theta \in \Theta\}$ and is given by

$$\{\theta : \hat{u}_\alpha(\theta) < R(\mathbf{X}, \theta) < \hat{u}_{1-\alpha}(\theta)\}\,. \tag{2.1}$$

The construction of (2.1) typically involves simulations to compute the quantiles as in the bootstrap method.

Since an exact method for constructing confidence regions is based on inverting a test, such a method is implicitly or explicitly linked to an ordering of the sample space of the test statistic used. The ordering defines the p-value of the test as the probability (under the null hypothesis) of more extreme values (under the ordering) of the test statistic than that observed in the sample. Under a total ordering $\leq$ of the sample space of (T, S_T), Lai and Li [32] call (t, s) a qth quantile if $P\{(T, S_T) \leq (t, s)\} = q$, which generalizes the exact method of [42] for randomly stopped sums S_T of of independent normal random variables with unknown mean μ. For the general setting where a stochastic process $\mathbf{X}_u$, in which u denotes either discrete or continuous time, is observed up to a stopping time T, Lai and Li [32] define $\mathbf{x} = \{\mathbf{x}_u : u \leq t\}$ to be a qth quantile if

$$P\{\mathbf{X} \leq \mathbf{x}\} \geq q, \quad P\{\mathbf{X} \geq \mathbf{x}\} \geq 1 - q, \tag{2.2}$$

under a total ordering $\leq$ for the sample space of $\mathbf{X} = \{\mathbf{X}_u : u \leq T\}$. For applications to confidence intervals of a real parameter θ, the choice of the total ordering should be targeted toward the objective of interval estimation. Let $\{U_r : r \leq T\}$ be real-valued statistics based on the observed process $\{\mathbf{X}_s : s \leq T\}$. For example, let U_r be an estimate of θ based on $\{\mathbf{X}_s : s \leq r\}$. A total ordering on the sample space of $\mathbf{X}$ can be defined via $\{U_r : r \leq T\}$ as follows:

$$\mathbf{X} \geq \mathbf{x} \text{ if and only if } U_{T\wedge t} \geq u_{T\wedge t}, \tag{2.3}$$

in which $\{u_r : r \leq t\}$ is defined from $\mathbf{x} = \{\mathbf{x}_r : r \leq t\}$ in the same way as $\{U_r : r \leq T\}$ is defined from $\mathbf{X}$ and which has the attractive feature that the probability mechanism generating $\mathbf{X}_t$ needs only to be specified up to the stopping time T in order to define the quantile. Section 7.4 of [6] describes how this ordering can be applied to implement resampling for secondary endpoints, and Section 7.5 describes its applications to time-sequential trials

which involve interim analyses at calendar time t_j $(1 \le j \le k)$, with $0 < t_1 < \cdots < t_k = t^*$ (the prescribed duration of the trial), and which have time to failure as the primary endpoint.

2.2. Adaptive enrichment designs in precision medicine

We begin by describing the work of Lai, Lavori and Liao [30] on enrichment designs in precision medicine. 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. Section 2 of [30] first provides an asymptotic theory for trials with fixed sample size, in which 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 a pre-defined subgroup Π_j for $j = 1, \ldots, J$, and with common known variance σ^2. Let Π_J denote the entire patient population for a traditional randomized controlled trial (RCT) comparing the two treatments, and let $\Pi_1 \subset \Pi_2 \subset \cdots \subset \Pi_J$ be the J prespecified subgroups. Since there is typically little information from previous studies about the subgroup effect size $\mu_j - \mu_{0j}$ for $j \ne J$, [30] begins with a standard RCT to compare the new treatment with the control over the entire population, but 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 $i = \hat{I}$ so that the new treatment can be claimed to be better than control for the patient subgroup $\hat{I}$ if $H_{\hat{I}}$ 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

(2.4)
$$\alpha(\boldsymbol{\theta}) = \begin{cases} P_{\boldsymbol{\theta}}(\text{reject } H_J) + P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \le 0, \text{ accept } H_J \text{ and reject } H_{\hat{I}}) & \text{if } \theta_J \le 0 \\ P_{\boldsymbol{\theta}}(\theta_{\hat{I}} \le 0, \text{ accept } H_J \text{ and Reject } H_{\hat{I}}) & \text{if } \theta_J > 0, \end{cases}$$

for $\boldsymbol{\theta} \in \Theta_0$. Subject to the constraint $\alpha(\boldsymbol{\theta}) \le \alpha$, [30, Appendix A] establishes the asymptotic efficiency of the procedure that randomly assigns n patients to the experimental treatment and the control, rejects H_J if $\mathrm{GLR}_i \ge c_\alpha$ for $i = J$, and otherwise chooses the patient subgroup $\hat{I} \ne J$ with the largest value of the generalized likelihood ratio statistic $\mathrm{GLR}_i = \{n_i n_{0i}/(n_i + n_{0i})\}(\hat{\mu}_i - \hat{\mu}_{0i})_+^2/\sigma^2$ among all subgroups $i \ne J$ and rejects $H_{\hat{I}}$ if $\mathrm{GLR}_{\hat{I}} \ge c_\alpha$, 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.

The test statistic GLR_i is the sample estimate of the Kullback-Leibler information $(np_i/4)(\mu_i - \mu_{0i})_+^2/\sigma^2$, noting that $n_i n_{0i}/(n_i + n_{0i}) \approx np_i$ as study subjects are equally likely to receive the new treatment or control. After establishing the asymptotic efficiency of the procedure in the fixed sample size case, [30] proceeds to extend it to a 3-stage sequential design by making use of the theory of Bartroff and Lai [3, 4] reviewed in the preceding paragraph. It then extends the theory from the normal setting to asymptotically normal test statistics, such as the Wilcoxon rank sum statistics.

We next consider another adaptive design for the development and testing of biomarker-guided treatment strategies, introduced by Lai, Liao and Kim [33], that attempts to select the best of k treatments for each biomarker-classified subgroup of cancer patients in phase II studies. The clinical trial has several objectives, which include (a) treating accrued patients with the best (yet unknown) available treatment, (b) developing a biomarker-guided treatment strategy for future patients, and (c) demonstrating that the strategy developed indeed has statistically significantly better treatment effect than some predetermined threshold. The group sequential design uses an outcome-adaptive randomization rule, which updates the randomization probabilities at interim analyses and use GLR statistics and modified Haybittle-Peto rules to include early elimination of inferior treatments from a biomarker class. It is shown to provide substantial improvements, besides being much easier to implement, over the Bayesian outcome-adaptive randomization design used in the BATTLE (Biomarker-integrated Approaches of Targeted Therapy for Lung Cancer Elimination) trial of personalized therapies for non-small cell lung cancer [28]. An April 2010 editorial in *Nature Reviews in Medicine* points out that 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. Besides BATTLE, another design mentioned in the editorial is that of the I-SPY2 (Investigation of Serial Studies to Predict Your Therapeutic Response with Imaging and Molecular Aanalysis) trial [7]. We use the following example to illustrate the basic idea of Lai, Liao and Kim's adaptive design of late-phase clinical trials for the development and validation of biomarker-guided personalized therapies.

As pointed out in [33, pp. 651–653, 662], targeted therapies that target the cancer cells (while leaving healthy cells unharmed) and the "right" patient population (that has the genetic or other markers for the sensitivity to

the treatment) have great promise in cancer treatments but also challenges in designing clinical trials for drug development and regulatory approval. One challenge is to identify the biomakers that are predictive of response, another is to develop a biomarker classifier that can identify patients who are sensitive to the treatments, and the third is that classical frequentist clinical trial designs and the more recent Bayesian trial designs are inadequate to address these issues and gain regulatory approval of the new treatment. This is pointed out in [33, p. 653], where it is also noted that the Bayesian "BATTLE and BATTLE-2 trials share the philosophy of the classical muti-armed bandit problem", which we have reviewed in Section 1.2. To achieve the objectives (a), (b) and (c) in the preceding paragraph, Lai et al. [33, pp. 654–655] use contextual bandit theory which we have reviewed in Section 1.2 and which we illustrate below with $J = 3$ groups of patient and $K = 3$ treatments, assuming normally distributed responses with mean μ_{jk} and known variance 1 for patients in group j receiving treatment k. Using Bartroff and Lai's adaptive design [3, 4] reviewed in Section 2.1, let n_i denote be the total sample size up to the time of the ith interim analysis, n_{ij} denote the total sample size from group j in those n_i patients, and let n_{ijk} be the total sample size from biomarker class j receiving treatment k up to the ith interim analysis. Because it is unlikely for patients to consent to being assigned to a seemingly inferior treatment, randomization in a double blind setting (in which the patient and the physician both do not know whether treatment or control is assigned) is needed for informed consent. The randomization probability $\pi_{jk}^{(i)}$, with which patient j is assigned to treatment k at the ith interim analysis, is described in Algorithm 1, in which $\hat{\mu}_{ijk}$ is the MLE of μ_{jk} and $\hat{k}_j^{(i)} = \arg\max_k \hat{\mu}_{ijk}$ is the MLE of $k_j^* = \arg\max_k \mu_{jk}$ at the ith interim analysis, and

$$
\begin{aligned}
2\ell_j^i(k, k_j^*) = {} & n_{ijk}\hat{\mu}_{ijk}^2 + n_{ij\hat{k}_j^{(i)}}\hat{\mu}_{ij\hat{k}_j^{(i)}}^2 \\
& - (n_{ijk}\hat{\mu}_{ijk} + n_{ij\hat{k}_j^{(i)}}\hat{\mu}_{ij\hat{k}_j^{(i)}})^2/(n_{ijk} + n_{ij\hat{k}_j^{(i)}}).
\end{aligned} \tag{2.5}
$$

Contextual multi-armed bandit theory suggests assigning the highest randomization probability between interim analyses i and $i+1$ to $\hat{k}_j^{(i)}$ and "nearby" treatments that are lumped into the set $\mathcal{H}_{ij}$, where $\delta_{ij} \to 0$ but $\sqrt{n_{ij}}\delta_{ij} \to \infty$, with the randomization scheme in Step 3 of Algorithm 1, in which

$$
\pi_{jk}^{(i)} = (1 - \varepsilon|\mathcal{K}_{ij} \setminus \mathcal{H}_{ij}|)/|\mathcal{H}_{ij}| \text{ for } k \in \mathcal{H}_{ij}, \tag{2.6}
$$

Algorithm 1 Randomization probabilities $\pi_{jk}^{(i)}$ and GLR tests for arm elimination at the ith interim analysis

1: Let $\mathcal{K}_{ij}$ be the set of surviving treatments in biomarker class j; "surviving" means that the treatment is not eliminated at or before interim analysis i. Compute the GLR statistic $\ell_j^i(k, k_j^*)$ by (2.5) for testing the null hypothesis $\mu_{jk} = \mu_{j,k_j^*}$ that the kth treatment has the same effect as that of the best treatment k_j^* on the jth biomarker class is $\ell_j^i(k, k_j^*)$.
2: Eliminate treatment k from $\mathcal{K}_{ij}$ at the ith interim analysis if $\ell_j^i(k, k_j^*) > 5\delta_{ij}$.
3: Let $\mathcal{H}_{ij} = \{k \in \mathcal{K}_{ij} : |\hat{\mu}_{jk}^{(i)} - \hat{\mu}_{j,\hat{k}_j^{(i)}}^{(i)}| \le \delta_{ij}\}$. Randomize treatments in $\mathcal{H}_{ij}$ with the probabilities $\pi_{jk}^{(i)}$ given by (2.6) below, and assign probability ε to each treatment in $\mathcal{K}_{ij} \setminus \mathcal{H}_{ij}$.

where $|A|$ denotes the cardinality of a finite set A. Equal randomization (with randomization probability $1/K$) for the K treatments is used up to the first interim analysis. In particular, for the simulation study with results summarized in Table 1, we choose $J = K = 3$, $\delta_{ij} = n_{ij}^{-2/5}$ (for which $\sqrt{n_{ij}}\delta_{ij} \to \infty$ as $n_{ij} \to \infty$), $n_i - n_{i-1} = 200$ and $\varepsilon = 0.1$. In context-free multi-armed bandit theory, this corresponds to the ε-greedy algorithm which has been shown by Auer et al. [2] to provide an alternative to the UCB rule for attaining the asymptotic lower bound for the regret.

We illustrate this 3-stage adaptive clinical trial design in a simulation study, the results of which are provided in Table 1 (for patients accrued to the trial) and Table 2 (for future patients); each result is based on 1000 simulations. The simulation study considers six scenarios of mean responses μ_{jk} that are listed in Table 1, which also gives in parentheses the expected number of patients in biomarker class j receiving treatment k. Note that the total number of patients in the trial is $n_3 = 600$ if there is no early stopping for futility. Scenarios S1–S3 have the biomarker class size proportional to 3:2:1 for $j = 1, 2, 3$. Scenarios S4–S6 have the same mean responses as S1–S3, but the class sizes are proportional to 1:1:1. Scenarios S1 and S4 have a best treatment that is substantially better than the others, and no negative effects for all treatments. Scenarios S2 and S5 have positive μ_{jk} (0.2, 0.3, 0.5) for $j = k$, and 0 or negative μ_{jk} for $j \ne k$. Scenarios S3 and S6 have $\mu_{jk} = 0.1$ for $j = k$, and negative $\mu_{jk} \in \{-0.05, -0.5\}$ for $j \ne k$. Table 1 shows that most of the patients in the trial are treated with the best treatment. Table 2 shows that our new test-based procedure for FWER control, described below, controls FWER in all six scenarios and also has good power for S1 and S4 (0.98, 0.95), moderate power for S2 and S5 (0.68, 0.73), and low power for S3 and S6.

Lai et al. [33] introduced a subset selection method for selecting a subset of treatments at the end of the trial to be used for future patients, with an overall probability guarantee of $1-\alpha$ to contain the best treatment for each biomarker class, and such that the expected size of the selected subset is as small as possible in some sense. Here we develop a test-based approach that is more directly related to the elimination rule and FWER control in multiple testing. For the statistical analysis after the conclusion of the trial, test for each biomarker class j the simple (multivariate) hypothesis $H_{0j} : \mu_{jk} = 0$ for $1 \le k \le K$, at significance level α. Since the probability (under H_{0j}) of early stopping for futility and accepting H_{0j} is 1, we can restrict to $i = I$ (the final analysis) for which the likelihood ratio test rejects H_{0j} if $\hat{\mu}_{Ij\hat{k}_j^{(I)}}$ exceeds some threshold c_α such that $\mathbf{P}_{H_{0j}}(\max_{1\le k\le K} \hat{\mu}_{Ijk} > c_\alpha) = \alpha$, which can be evaluated by Monte Carlo simulations (using the aforementioned "exact" resampling method for the adaptive design).

Table 1: Mean response and expected sample size (in parentheses) for scenarios S1–S6 involving $K = 3$ treatments and $J = 3$ biomarker classes

	Class	Treatment		
	j	1	2	3
S1	1	0.5 (141.33)	0.0 (29.64)	0.0 (28.93)
	2	0.0 (29.21)	0.5 (141.21)	0.0 (30.11)
	3	0.0 (28.42)	0.0 (28.72)	0.5 (142.43)
S2	1	0.2 (109.33)	−0.1 (37.97)	0.0 (52.41)
	2	−0.1 (32.78)	0.3 (126.37)	0.0 (41.63)
	3	0.0 (28.97)	−0.1 (26.18)	0.5 (144.36)
S3	1	0.1 (113.90)	−0.5 (24.88)	−0.05 (62.18)
	2	−0.5 (24.79)	0.1 (111.37)	−0.05 (63.72)
	3	−0.05 (60.72)	−0.5 (24.74)	0.1 (113.70)
S4	1	0.5 (221.77)	0.0 (39.72)	0.0 (38.54)
	2	0.0 (29.20)	0.5 (140.33)	0.0 (30.13)
	3	0.0 (19.50)	0.0 (18.59)	0.5 (62.22)
S5	1	0.2 (177.03)	−0.1 (50.18)	0.0 (72.10)
	2	−0.1 (32.64)	0.3 (126.35)	0.0 (41.43)
	3	0.0 (19.05)	−0.1 (16.95)	0.5 (64.27)
S6	1	0.1 (179.75)	−0.5 (34.52)	−0.05 (85.68)
	2	−0.5 (24.61)	0.1 (113.49)	−0.05 (61.72)
	3	−0.05 (34.04)	−0.5 (14.93)	0.1 (51.26)

2.3. Group sequential multiple testing in online experiments

For online experiments of IT companies and A/B testing, a predetermined duration is usually specified, during which we carry out T interim (including

Table 2: FWER and power of test-based subset selection method ($\alpha = 0.05$)

	S1	S2	S3	S4	S5	S6
FWER	0.03	0.01	0.01	0.02	0.01	0
Power	0.98	0.68	0.1	0.95	0.73	0.11

final) analyses. As noted in the second paragraph of Section 1.1, the sample mean difference is approximately normal with variance that can be consistently estimated. We therefore assume in the sequel that the responses to treatment (respectively, control) are i.i.d. normal. Moreover, one often finds a treatment to have markedly positive effect on a subgroup but negligible or even negative effects for other subgroups after dividing the population into disjoint subgroups $G_1, \ldots, G_J$. The A/B test also involves multiple metrics. For the metric $m (1 \le m \le M)$, let $\mu_{m,j}$ (respectively, $\mu^0_{m,j}$) be the mean of the responses to treatment (respectively, control) for subgroup G_j. To test the null hypothesis $H^{(m)}_{0j} : \mu_{m,j} = \mu^0_{m,j}$ versus the one-sided alternative $\mu_{m,j} > \mu^0_{m,j}$, compute at the tth interim analysis the test statistic

$$Z^{(m)}_{tj} = \{n_{tj} n^0_{tj}/(n_{tj} + n^0_{tj})\}^{1/2} (\hat{\mu}_{t;m,j} - \hat{\mu}^0_{t;m,j})/\hat{\sigma}_{t;m,j}, \tag{2.7}$$

which is the signed-root GLR statistic when the responses are normal and in which n_{tj} (respectively, n^0_{tj}) is the sample size of subgroup G_j up to the tth interim analysis and

$$\hat{\sigma}^2_{t;m,j} = \sum_{i=1}^{n_{tj}} (X_{i;m,j} - \hat{\mu}_{t;m,j})^2/(n_{tj} - 1) + \sum_{i=1}^{n^0_{tj}} (X^0_{i;m,j} - \hat{\mu}^0_{t;m,j})^2/(n^0_{tj} - 1)$$

is a consistent estimate of $\sigma^2_{m,j} = \mathrm{Var}(X_{i;m,j} - X^0_{i;m,j})$. Let $\mathrm{H}_m = \bigcap_{1 \le j \le J} \mathrm{H}^{(m)}_{0j}$,

$$Z^{(m)}_t = \max_{1 \le j \le J} Z^{(m)}_{tj}. \tag{2.8}$$

We use Bartroff and Lai's [5] multistage extension of Holm's step-down procedure [25], which we summarize in Algorithm 2 below with

(2.9)

$$C(\rho) = (1 - \rho)\text{th quantile of the distribution of } \max_{1 \le t \le T} Z^{(m)}_t \text{ under } \mathrm{H}_m,$$

to control the FWER in testing the multiple hypotheses $\mathrm{H}_m, 1 \le m \le M$. Note that under H_m, $\mu_{m,j} - \mu^0_{m,j} = 0$ for all j, hence the distribution of $\max_{1\le t\le T} Z_t^{(m)}$ (and therefore $C(\rho)$ also) does not depend on m and

$$\mathbf{P}_{\mathrm{H}_m}\Big\{\max_{1\le t\le T} Z_t^{(m)} \ge C(\rho)\Big\} \le \rho \text{ for } 1 \le m \le M. \tag{2.10}$$

For the values of ρ in Algorithm 2, we use Monte Carlo simulations to compute the quantile in the right-hand side of (2.9) to evaluate $C(\rho)$.

Algorithm 2 Multistage step-down procedure for testing $\mathrm{H}_m, 1 \le m \le M$, with FWER control rate α and $C(\rho)$ given by (2.9)

1: Order the test statistics as $Z_{(t,1)} \ge \ldots \ge Z_{(,M)}$ and reject $\mathrm{H}_{(1)}, \ldots, \mathrm{H}_{(m_t)}$, where

$$m_t = \max\Big\{m \ge 1 : \min_{1\le j\le m}\Big[Z_{(t,m)} - C\Big(\frac{\alpha}{M-j+1}\Big)\Big] \ge 0\Big\}.$$

2: If $t < T$ and there is no rejected hypothesis, set $t = t+1$ and GOTO Step 1. Otherwise STOP and output the set of rejected hypotheses.

From (2.10) and Theorem 2.1 of [5], the 3-stage multiple test has FWER controlled at level α. Simulation studies are conducted to illustrate the performance of the 3-stage test, which is compared with the traditional fixed sample size (FSS) step-down t-tests for multiple testing of mean differences with FWER control. The results are summarized in Tables 3 and 4. There are 5 scenarios for the alternative hypotheses in each table besides scenario S0 for the null hypothesis. In each scenario, 550 users are accrued to the online experiment, with 250 for the first interim analysis, 150 added to the second interim analysis and the remaining 150 added to the final analysis. The metrics are normally distributed with the same mean $\mu^0_{m,j} = 0$ for users assigned to the control, and different means $\mu_{m,j}$ for those assigned to the treatment. There are 50 metrics, the first six of which are the only ones affected by the treatment. Their values are listed in the vector $\boldsymbol{\mu}$ in the tables, which give the expected stopping time $E(\tau)$, the FWER, power (i.e., probability of rejecting all false null hypotheses), and the detection rate DetRate; each result is based on 1000 Monte Carlo simulations. Since an IT company would decide to launch a new feature when significant positive effects are found in some metric, we also report the probability of detecting some metric with positive treatment effect, which we call the "detection rate". The tables show that both the 3-stage test and the FSS test have FWER controlled at

the prespecified $\alpha = 0.2$ level. While Table 3 assumes the relative frequencies $r_j (j = 1, \ldots, 6)$ of the subgroups G_j to be 1/6, Table 4 assumes unequal relative frequencies $(0.2, 0.1, 0.3, 0.1, 0.1, 0.2)$ for $(r_1, r_2, r_3, r_4, r_5, r_6)$. The 3-stage test stops at the first stage with very high probability in scenarios S1-S5, although FSS has substantially higher power in S1-S4. Scenario S5 represents the situation where the treatment has negative effects on most subgroups but also has strong positive effects on some subgroups. While FSS has near-zero power and detection rate, the 3-stage test has detection rate 1 and power around 0.8.

Table 3: Comparison of 3-stage procedure in upper row with FSS test in lower row; FWER controlled at level $\alpha = 0.2$

Scenario	$E(\tau)$	FWER	Power	DetRate
S0: $\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0)$	2.82	0.00	0	0
	3	0.02	0	0
S1: $\boldsymbol{\mu} = (1, 0.8, 0.6, 0, 0, 0)$	1.06	0.01	0.48	1
	3	0.01	0.78	1
S2: $\boldsymbol{\mu} = (0.6, 0.6, 0.6, 0.6, 0.6, 0.6)$	1.26	0.00	0.35	0.98
	3	0.03	0.80	1
S3: $\boldsymbol{\mu} = (0.8, 0.6, 0.4, 0, 0, 0)$	1.40	0.01	0.31	0.97
	3	0.02	0.64	1
S4: $\boldsymbol{\mu} = (1, 1, 0.6, 0.6, 0.2, 0.2)$	1.01	0.01	0.58	1
	3	0.02	0.80	1
S5: $\boldsymbol{\mu} = (1.2, 1.2, -0.6, -0.6, -0.6, -0.6)$	1	0.00	0.79	1
	3	0.02	0.01	0.04

Table 4: Comparison in the case of unequal group frequencies

Scenario	$E(\tau)$	FWER	Power	DetRate
S0: $\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0)$	2.68	0.01	0	0
	3	0.02	0	0
S1: $\boldsymbol{\mu} = (1, 0.8, 0.6, 0, 0, 0)$	1.01	0.01	0.56	1
	3	0.02	0.8	1
S2: $\boldsymbol{\mu} = (0.6, 0.6, 0.6, 0.6, 0.6, 0.6)$	1.13	0.01	0.38	0.99
	3	0.02	0.80	1
S3: $\boldsymbol{\mu} = (0.8, 0.6, 0.4, 0, 0, 0)$	1.21	0.01	0.35	0.97
	3	0.02	0.73	1
S4: $\boldsymbol{\mu} = (1, 1, 0.6, 0.6, 0.2, 0.2)$	1.01	0.01	0.61	1
	3	0.02	0.80	1
S5: $\boldsymbol{\mu} = (1.2, 1.2, -0.6, -0.6, -0.6, -0.6)$	1	0.01	0.77	1
	3	0.02	0	0.02

3. Conclusion

This section gives some concluding remarks and ongoing work, together with further discussion of the methods and results in Section 2, which builds upon recent advances in contextual bandit theory, adaptive design of clinical trials and group sequential multiple testing to develop adaptive subset selection for biomarker-guided personalized strategies in precision medicine and A/B testing with multiple metrics and user subgroups in on-line experiments. Resampling methods play an important role to implement the likelihood ratio or generalized likelihood tests with analytically intractable distributions under the null hypothesis in the adaptive designs. It should be noted that the GLR statistics are approximate pivots for composite null hypotheses, and this explains why hybrid resampling is so effective for GLR statistics; see [6] and [15, Chapter 16].

3.1. Adaptive enrichment designs for confirmatory trials

We provide here more details and discussion of the adaptive enrichment designs reviewed in the first paragraph of Section 2.1. Although conventional randomized controlled trial (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 been recently developed by Lai, Lavori and Liao [30] in connection with the design of the DEFUSE 3 clinical trial at the Stanford Stroke Center to evaluate a new method for augmenting usual medical care with endovascular removal of the clot after a stroke, resulting in reperfusion of the area of the brain under threat, in order to salvage the damaged tissue and improve outcomes over standard medical care with intravenous tissue plasminogen activator (tPA) alone. The clinical endpoints of stroke patients are the Rankin scores, and Wilcoxon rank sum statistics are used to test for differences in Rankin scores between the new and control treatments. The DEFUSE 3 (Diffusion and Perfusion Imaging Evaluation for Understanding Stroke Evolution) trial design involves 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 Lai et al. [30, p. 195]. 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.

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. Consequently, 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. 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. The first interim analysis was scheduled after the 200 patients, and the second interim analysis after an additional 140 patients. DEFUSE 3 has a Data Coordinating Unit 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 provided 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 would also consider 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.

In June 2017 positive results of another trial DWI or CTP Assessment with Clinical Mismatch in the Triage of Wake-Up and Late Presenting Stokes 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 Albers et al. [1] that says: "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 [1], 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. 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" [39]. Early termination of DEFUSE 3 provides a concrete example of importance of 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 and related concurrent trials.

3.2. Contextual bandits and mobile health

This is an ongoing project, which is related to Section 1.2. Advances in mobile technology offer opportunities to deliver interventions that accommodate an individual's immediate needs [45]. Just-in-time adaptive interventions (JITAIs) aim to provide support for health behavior change at times when users most need the support. A key problem in designing JITAIs for mobile health is to learn decision rules from data that can map tailoring variables (e.g., user mood, time of day) to intervention options (e.g., whether a message should be sent to the user's phone right now or later). Contextual bandits provide a natural framework for sequential decision making in mobile health regarding attempts to construct decision rules with the goal of maximizing some numerical outcome (metric) following every decision point. The field of interactive machine learning encompasses applications to mobile health, and contextual bandits enable personalization such that (a) the set of possible interventions can be enlarged, thus improving their efficacy, and (b) it can take advantage of the already logged data for system optimization, without the need of a user model, to provide impactful innovations in personalized mHealth interventions [17, 36].

References

[1] Albers, G.W., Marks, M.P., Kemp, S., et al. (2018). Thrombectomy for stroke at 6 to 16 hours with selection by perfusion imaging. *N. Engl. J. Med.* 378, 708–718.

[2] Auer, P., Cesa-Bianchi, N., Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 235–256.

[3] Bartroff, J., Lai, T.L. (2008). Efficient adaptive designs with mid-course sample size adjustment in clinical trials. *Stat. Med.*, 27, 1593–1611. MR2420330

[4] Bartroff, J., Lai, T.L. (2008). Generalized likelihood ratio statistics and uncertainty adjustments in efficient adaptive design of clinical trials. *Sequent. Anal.*, 27, 254–276. MR2446902

[5] Bartroff, J., Lai, T.L. (2010). Multistage tests of multiple hypotheses. *Comm. Statist.–Theory and Methods*, 39, 1597–1607.

[6] Bartroff, J., Lai, T.L., Shih, M.C. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis*, Springer, New York. MR2987767

[7] Barker, A.D., Sigman, C.C., Kelloff, G.J., Hylton, N.M., Berry, D.A., Esserman, L. (2009). I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clin. Pharmacology. Theor.*, 86, 97–100.

[8] Bellman, R. (1971). *Introduction to the Mathematical Theory of Control Processes, Vol. II*, Academic Press, New York. MR0278767

[9] Birkett, M., Day, S. (1994). Internal pilot studies for estimating sample size. *Stat. Med.*, 13, 2455–2463.

[10] Chang, F., Lai, T.L. (1987). Optimal stopping and dynamic allocation. *Adv. Applied Prob.*, 19, 829–853. MR0914595

[11] Chen, J., Heyse, J., Lai, T.L. (2019). *Medical Product Safety Evaluation: Biological Models and Statistical Methods*, Chapman & Hall/CRC, Boca Raton, FL.

[12] Chernoff, H., Ray, S.N. (1965). A Bayes sequential sampling inspection plan. *Ann. Math. Statist.*, 1387–1407. MR0191062

[13] Chuang, C.S., Lai, T.L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika*, 85, 317–332. MR1649116

[14] Chuang, C.S., Lai, T.L. (2000). Hybrid resampling methods for confidence intervals (with discussion and rejoinder). *Statistica Sinica*, 10, 1–50. MR1742099

[15] de la Pena, V.H., Lai, T.L., Shao, Q. (2009). *Self-normalized Processes: Limit Theory and Statistical Applications*, Springer-Verlag, Berlin, Heidelberg MR2488094

[16] Denne, J.S., Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika*, 87, 125–134. MR1766833

[17] Dudík, M. (2017). 'Contextual bandit' breakthrough enables deeper personalization. https://www.microsoft.com/en-us/research/blog/contextual-bandit-breakthrough-enables-deeper-personalization/.

[18] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7, 1–26. MR0515681

[19] Efron, B. (1987). Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, 82, 171–185. MR0883345

[20] Fisher, L. (1998). Self-designing clinical trials. *Stat. Med.*, 17, 1551–1562.

[21] Gittins, J.C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B*, 41, 148–164. MR0547241

[22] Gould, A.L., Shih, W.J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Comm. Stat. Ser. A*, 21, 2833–2853.

[23] He, P., Lai, T.L., Zheng, S. (2015). Design of clinical trials with failure-time endpoints and interim analyses: An update after 15 years. *Contr. Clin. Trials*, 45, 103–112.

[24] Herson, J., Wittes, J. (1993). The use of interim analysis in sample size adjustment. *Drug Inform J.*, 27, 753–760.

[25] Holm, S. (1979). A simple Sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6, 65–70. MR0538597

[26] Jennison, C., Turnbull, B.W. (2006). Adaptive and nonadaptive group sequential tests. *Biometrika*, 93, 1–21. MR2277736

[27] Jennison, C., Turnbull, B.W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Stat. Med.*, 25, 917–932. MR2225182

[28] Kim, E.S., Herbst, R.S., Wistuba, I.I., et al. (2011). The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discovery*, 1, 44–53.

[29] Lai, T.L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.*, 15, 1091–1114. MR0902248

[30] Lai, T.L., Lavori, P.W., Liao, O.Y. (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contr. Clin. Trials*, 39, 191–200.

[31] Lai, T.L., Lavori, P.W., Tsang, K.W. (2019). Adaptive enrichment designs for confirmatory trials. *Stat. Med.*, 38, 613–624. MR3902601

[32] Lai, T.L., Li, W. (2006). Confidence intervals in group sequential trials with random group sizes and applications to survival analysis. *Biometrika*, 93(3), 641–654. MR2261448

[33] Lai, T.L., Liao, O.Y.W., Kim, D.W. (2013). Group sequential designs for developing and testing biomarker-guided personalized therapies in comparative effectiveness research. *Cont. Clinical Trials*, 36, 651–663.

[34] Lai, T.L., Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Applied Math.*, 6, 4–22. MR0776826

[35] Lai, T.L., Shih, M.C. (2004). Power, sample size and adaptation considerations in the design of group sequential clinical trials. *Biometrika*, 91(3), 507–528. MR2090619

[36] Lei, H., Tewari, A., Murphy, S.A. (2017). An actor-critic contextual bandit algorithm for personalized mobile health interventions, *arXiv:1706.09090v1 [stat.ML]*.

[37] Li, L., Chu, W., Langford, J., Schapire, R.E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Nineteenth International Conference on World Wide Web (WWW 2010)*.

[38] Lorden, G. (1983). Asymptotic efficiency of three-stage hypothesis tests. *Ann. Stat.*, 11, 129–140. MR0684871

[39] Nogueira, R.G., Jadhav, A.P., Haussen, D.C., et al. (2018). Thrombectomy 6 to 24 hours after stroke with a mismatch between deficit and infarct. *N. Engl. J. Med.*, 378, 11–21.

[40] Proschan, M., Hunsberger, S. (1995). Designed extension studies based on conditional power. *Biometrics*, 51, 1315–1324.

[41] Robbin, J.W. (1972). Topological conjugacy and structural stability for discrete dynamical systems. *Bulletin of the American Mathematical Society*, 78(6), 923–952. MR0312529

[42] Rosner, G.L., Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, 75, 723–729.

[43] Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.*, 16, 243–258. MR0013885

[44] Tang, L., Rosales, R., Singh, A., and Agarwal, D. (2013). Automatic ad

format selection via contextual bandits. In *Proc. 22nd ACM Conference on Information and Knowledge Management*, 329–338, 1587–1594.

[45] Tewari, A., and Murphy, S.A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications* (Rehg, J., Murphy, S.A., and Kumar, S., eds.), Springer, New York.

[46] Tsiatis, A.A., Mehta, C. (2003). On the efficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90, 367–378. MR1986653

[47] Wittes, J., Brittain, E. (1990). The role of internal pilots in increasing the efficiency of clinical trials. *Stat. Med.*, 9, 65–72.

[48] Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. B*, 42, 143–149. MR0583348

Tze Leung Lai
Department of Statistics
Stanford University
Stanford, CA 94305-4065
USA
E-mail address: lait@stanford.edu

Anna Choi
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen
China
E-mail address: annachoi@cuhk.edu.cn

Ka Wai Tsang
School of Science and Engineering
Chinese University of Hong Kong
Shenzhen
China
E-mail address: kwtsang@cuhk.edu.cn

Received September 16, 2019