ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Adaptive Enrichment Designs in Clinical Trials

Peter F. Thall

Department of Biostatistics, M.D. Anderson Cancer Center, University of Texas, Houston, Texas 77030, USA; email: rex@mdanderson.org

Annu. Rev. Stat. Appl. 2021. 8:393–411

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-040720-032818

Keywords

adaptive signature design, Bayesian design, biomarker, clinical trial, group sequential design, precision medicine, subset selection, targeted therapy, variable selection

Abstract

Adaptive enrichment designs for clinical trials may include rules that use interim data to identify treatment-sensitive patient subgroups, select or compare treatments, or change entry criteria. A common setting is a trial to compare a new biologically targeted agent to standard therapy. An enrichment design's structure depends on its goals, how it accounts for patient heterogeneity and treatment effects, and practical constraints. This article first covers basic concepts, including treatment-biomarker interaction, precision medicine, selection bias, and sequentially adaptive decision making, and briefly describes some different types of enrichment. Numerical illustrations are provided for qualitatively different cases involving treatment-biomarker interactions. Reviews are given of adaptive signature designs; a Bayesian design that uses a random partition to identify treatment-sensitive biomarker subgroups and assign treatments; and designs that enrich superior treatment sample sizes overall or within subgroups, make subgroup-specific decisions, or include outcome-adaptive randomization.

1. INTRODUCTION

1.1. Basic Concepts and Examples

It is widely recognized in oncology that the molecular biology of virtually all tumors is heterogeneous. For clinical evaluation of targeted agents, this has motivated the use of genomic, proteomic, or other biological variables obtained from new technologies such as microarrays, cytometry by time of flight, or DNA sequencing. If an experimental agent, E, is designed to hit a biological target thought to be associated with the disease being treated, whether a target is present in a patient may be represented by a function of a vector $\boldsymbol{Z}$ of biomarkers. The idea is that by hitting a biological target or targets, E may disrupt a functional pathway of cancer cells and thus increase the chance of a favorable clinical outcome. Major goals are determining a function of $\boldsymbol{Z}$ that characterizes a subgroup of patients likely to respond to a new targeted agent and comparing E to a standard control treatment, C, in the identified subgroup. To address these issues, the US Food and Drug Administration (FDA) has provided guidelines for evaluating new cancer vaccines (FDA 2011).

There is a large, growing literature on tumor heterogeneity, which may be between different patients who have the same disease, between different tumors in one patient, within a given tumor over time, or within a tumor and determined by tissue samples taken from different locations. Numerous clustering methods have been proposed, with a review given by Weber & Robinson (2016). An important modern approach to characterizing tumor heterogeneity is application of Bayesian feature allocation models (FAMs), first formulated by Griffiths & Ghahramani (2006) and reviewed by Griffiths & Ghahramani (2011). As an illustrative example, consider a setting where each cancer cell in one or more samples may or may not have any of a set of biomarkers, such as surface markers identified by cytometry. A FAM accounts for the fact that a tumor cell population in a sample may include different cell subpopulations, each characterized by a set of biomarkers. The FAM identifies cell subpopulations by using a random binary matrix, with rows corresponding to biomarkers and columns corresponding to cell subpopulations. The FAM uses latent (unobserved) feature variables to represent the unknown subpopulations, and it includes a feature allocation prior on the probabilities of whether or not each biomarker is expressed in each subpopulation. A FAM is more general than a clustering algorithm, since a feature is a set of the biomarkers, and a given biomarker may belong to more than one cluster. For example, maximum a posteriori estimation with FAM priors was proposed by Xu et al. (2015) for identifying haplotypes and subclones. Lee et al. (2016) developed a FAM framework for identifying subclonal copy number and within-patient single nucleotide mutations over time. Reviews of the biological bases for tumor heterogeneity and possible strategies for targeted therapies were written by Fisher et al. (2013) and Dagogo-Jack & Shaw (2018), among many others.

Consider a new agent, E, that has been designed to hit a particular biological target, a binary variable defined with $Z = 1$ if a patient has a biomarker for the target and $Z = 0$ if not, and the primary clinical outcome of interest, denoted by Y. Patients with $Z = 1$ are said to have biomarker positive disease, or to be E-sensitive. Usually, Y is an early treatment response indicator or survival time. If the new molecule behaves as it was designed, then on average, patients who receive E will live longer, and this effect will be larger if $Z = 1$ than if $Z = 0$; that is, E-sensitive patients can be expected to benefit more from E. More generally, Y may denote a vector including two or more coprimary endpoints, such as an indicator Y_{RES} of early antidisease effect (response), an indicator Y_{TOX} of a severe adverse event (toxicity), and Y_{S} = survival time, so in this more general case, $Y = (Y_{\mathrm{RES}}, Y_{\mathrm{TOX}}, Y_{\mathrm{S}})$. This sort of outcome is used, for example, by Chapple & Thall (2019) to construct a hybrid phase I-II-III design. While such designs using multidimensional Y can be very useful, this article focuses on enrichment trials, so only designs with one primary outcome are discussed. If the effects of E on Y in the subgroups with $Z = 1$ and $Z = 0$ differ, then Z is predictive. Given

an active control treatment, C, denote the treatment indicator $\tau_E = 1$ if a patient is treated with $E + C$ and $\tau_E = 0$ if treated with C. In some settings, E is not combined with C, but this distinction is ignored here for simplicity.

As a prototype example, suppose that the disease is de novo acute myelogenous leukemia (AML), C is the intravenous chemotherapy agent cytosine arabinoside (ara-C), and E is a designed molecule that may aim to either inhibit cancer cell proliferation by targeting a mutation in the FLT3 tyrosine kinase enzyme or enhance ara-C-induced cell death. The trial would randomize patients between ara-C alone (C) and the targeted agent + ara-C ($E + C$). In this setting, Y may be the indicator of complete remission (CR) within 42 days, or survival time. The aim is that E-sensitive patients treated with $E + C$ should have larger $p = \text{Prob(CR)}$ if Y is CR, or longer mean survival if Y is survival time. In the extreme case, if $Z = 0$, then $E + C$ provides no benefit at all over C. In a regression model for the distribution of Y as a function of (τ_E, Z), a linear term accounting for the possible effects of both E and Z may take the form

$$\eta = \mu + \gamma\, \tau_E + \beta\, Z + \xi\, \tau_E\, Z.$$

In η, the effect of E is $\gamma + \xi$ if $Z = 1$ and γ if $Z = 0$. This implies that the treatment-biomarker interaction quantifying the additional effect of E in biomarker positive patients is $(\gamma + \xi) - \gamma = \xi$. If $\xi = \gamma = 0$, then E has no antidisease effect at all, regardless of biomarker status. In the AML example, if Y is a binary indicator of CR, then one might assume the logistic model $\eta = \log\{p/(1 - p)\}$. If Y is survival time, then η would appear in the model for the logarithm of the hazard of death. For example, a Weibull distribution may be assumed since it has the flexible hazard function $h(t) = \lambda\, \alpha\, t^{\alpha - 1}$, which is the death rate at time $t > 0$, where λ is a rate parameter, α is a shape parameter, and one may assume $\lambda = \exp(\eta)$ to model the effects of E and biomarker status Z on survival time.

1.2. Enrichment Based on Many Biomarkers

In practice, a vector $\mathbf{Z}$ that includes many candidate biomarkers often is available, rather than only one binary Z, and an important statistical problem is to identify a discrimination function, $f(\mathbf{Z})$, such that $f(\mathbf{Z}) > c$ for a fixed cutoff c identifies a patient as being E-sensitive. Statistical methods for identifying such a function are discussed below in Sections 4 and 6. If an E-sensitive patient subset in which $E + C$ has a substantively larger antidisease effect than C can be determined, then $\mathbf{Z}$ may be used by practicing physicians to guide precision medicine, wherein the physician uses each patient's biomarker vector, $\mathbf{Z}$, to guide treatment decisions. The availability of modern biomarkers notwithstanding, physicians have been using patient covariates to guide their therapeutic decision making for thousands of years.

To account for these possibilities prospectively, an adaptive enrichment design, probability model, and parameter estimation method must address two closely related statistical problems. These are (*a*) identification of a subpopulation of E-sensitive patients, and (*b*) estimation (evaluation) or testing (validation) of the effects of E on Y in both biomarker positive and biomarker negative patients. Doing both identification and evaluation reliably in the same clinical trial, or in a series of trials, is a challenging problem. Preclinical in vitro data on the molecular biology of the disease and in vivo data on the effects of E in rodents xenografted with the disease may suggest which elements of $\mathbf{Z}$ are more likely to identify E-sensitive patients. However, data on humans treated with $E + C$ and C from a properly designed randomized clinical trial are needed. Strategies and designs for clinical identification, evaluation, and validation are reviewed below in Sections 3 through 6.

While it might appear counterintuitive to evaluate possible effects of E in biomarker negative patients, in terms of clinical outcomes, a new targeted agent often does not affect a particular human disease as expected based on preclinical data. Moreover, any statistical rule for dichotomizing patients into E-sensitive and non-E-sensitive subgroups based on an available biomarker vector $\boldsymbol{Z}$ is not perfect because it is based on data, and data are subject to random variation. Consequently, E may turn out to have a substantive antidisease effect in biomarker negative patients. Numerical examples of this are given below.

In most enrichment designs, decisions and actions include identifying a patient subgroup considered to be sensitive to E, restricting enrollment to E-sensitive patients, and testing whether $E + C$ provides a substantive benefit over C either within the E-sensitive subgroup or overall. More generally, the two main goals of clinical trials, including adaptive enrichment trials, are to benefit the patients in the trial and to provide high-quality data for making statistical inferences to benefit future patients. The central statistical problem in adaptive enrichment designs is that the statistical decisions include some combination of selection of a subvector $\boldsymbol{Z}^*$ of elements of $\boldsymbol{Z}$; construction of a function $f(\boldsymbol{Z}^*)$ that is used to define an E-sensitive subgroup, or possibly more than two subgroups; and one or more comparative tests to assess the effect of E, conducted in sequence and possibly within different subgroups. The conventional type I error and power of one such test viewed in isolation are incorrect and misleading since all of the other statistical decisions have been ignored. Consider the probability of the following two actions:

- Step 1: Determine $\boldsymbol{Z}^*$, $f(\boldsymbol{Z}^*)$, and a cutoff c such that $[f(\boldsymbol{Z}^*) > c] = [E\text{-sensitive}]$.
- Step 2: Test whether E is superior to C in the E-sensitive patient subset determined in step 1.

Denote the response probabilities, or mean survival times, for the two treatments by θ_{E+C} and θ_C. The probability of both step 1 and step 2 for a given value θ_{E+C} larger than θ_C may be called the generalized power (GP) of the procedure. The GP is smaller than the probability of the test (step 2) considered alone, as if the subset of E-sensitive patients were known and not determined from data. A conventional power figure is misleading if the test has been preceded by decisions such as variable selection, subset selection, or treatment selection. The relevant quantity is the GP of the entire decision process. Moreover, to compute a GP, one must assume a true subset of E-sensitive patients as well as a parameter θ_{E+C} in that subset.

1.3. Treatment Selection and Estimation Bias

While the use of biomarkers to guide application of targeted agents is the most common idea of an enrichment design, such designs may include a variety of different types of sequentially adaptive decisions. They are adaptive in that treatment, outcome, and covariates $(\tau, Y, \boldsymbol{Z})$ from previous patients, and each newly enrolled patient's $\boldsymbol{Z}$, may be used to make interim decisions. Many adaptive enrichment designs are based on (τ, Y) only and do not involve patient covariates. An adaptive futility rule stops accrual to a treatment found to be ineffective, and an adaptive safety rule stops accrual to a treatment found to be unsafe. An example is a randomized phase II-III select-and-test trial of three experimental agents, E_1, E_2, and E_3, and a control, C, that includes adaptive futility and safety rules. In the AML setting, if one wishes to evaluate three different targeted agents in the same trial, then C again denotes the ara-C arm, and each E_j denotes a targeted agent given in combination with ara-C. If interim data show that, for example, the E_1-versus-C efficacy effect is negligible, or that the E_1-versus-C toxicity rate is unacceptably high, then one of the rules may drop E_1 and increase (enrich) the sample sizes of E_2 and E_3, followed by confirmatory comparison of one or both of $\{E_2, E_3\}$ to C. This approach was taken in the STAMPEDE (Systemic

Therapy in Advancing or Metastatic Prostate Cancer: Evaluation of Drug Efficacy) trial (James et al. 2009) in men with advanced prostate cancer, a very large randomized trial including many experimental treatments. After an initial stage for safety testing, interim treatment comparisons are based on time to failure, defined as disease progression or death, allowing termination of arms showing poor performance compared with the control arm, with final comparisons to the control based on survival time.

Estimation bias is a major issue after doing adaptive futility or safety monitoring, treatment selection, or subgroup selection, and it is an inherent problem in enrichment trials. Temporarily ignoring subgroups, consider a conventional group sequential design for a randomized two-arm trial of E versus C (see, e.g., Jennison & Turnbull 2007). If a futility rule is included in the decision scheme, the fact that the difference $\hat{\Delta} = \hat{\theta}_E - \hat{\theta}_C$ between the estimated response probabilities or expected survival times must be sufficiently large at each interim test to continue the trial causes the final $\hat{\Delta}$ statistic to overestimate the true Δ. To see this, suppose first that the trial was conducted without any adaptive interim rules to stop the trial early. Then the fact that patients were randomized between E and C would ensure, by a standard statistical argument, that the distribution of the estimator $\hat{\Delta}$ based on the final data will have a distribution with mean (expected value) equal to the true Δ, written as $\mathrm{E}(\hat{\Delta}) = \Delta$. This is the definition of an unbiased estimator. Now suppose that, instead, the trial has two stages with equal sample sizes, and a futility rule is applied after stage 1. If the futility rule does not stop the trial early, then only interim stage 1 data that give values of $\hat{\Delta}_1$ large enough to continue the trial are possible. So $\mathrm{E}(\hat{\Delta}_1)$ must be larger than the true parameter Δ. Denote the estimator based on the stage 2 data by $\hat{\Delta}_2$. So, if the trial is not stopped early by the futility rule, then since the final estimator is $\hat{\Delta} = 0.50\hat{\Delta}_1 + 0.50\hat{\Delta}_2$, the upward bias in $\hat{\Delta}_1$ causes the final $\hat{\Delta}$ to overestimate Δ.

A similar problem arises in a two-arm trial comparing E to C where patient subgroups $\{S_1, \ldots, S_m\}$ have been identified before the start of the trial, possibly based on a biomarker vector $\boldsymbol{Z}$. If one uses the final data to select the best subset, S_{j^*}, defined as having the largest estimated E-versus-C effect, $\hat{\Delta}_{j^*}$, then the estimate of Δ in that subset will be upwardly biased. If, in fact, the subsets have no effect whatsoever, then the estimators $\hat{\Delta}_1, \ldots, \hat{\Delta}_m$ will have identical distributions, all with mean Δ. But the fact that the maximum $\hat{\Delta}_{j^*}$ must be larger than all of the other $\hat{\Delta}_j$s implies that $\mathrm{E}(\hat{\Delta}_{j^*}) > \Delta$. For example, if 10 independent random variables are uniformly distributed between 0 and 1, so that each has mean 0.50, then the maximum of the 10 has expected value 0.91, rather than 0.50. If there are no actual treatment-subset interactions, in this case a nominally best subset will be identified purely due to the play of chance. Ignoring this basic fact is likely to lead to serious errors when making inferences about a selected best E_{j^*}, including estimation bias, miscalculation of a test's power, and the incorrect conclusion that E_{j^*} provides a treatment advance over C when in fact it does not. The common practice of selecting a best subset based on post hoc data analyses, sometimes called cherry picking or data dredging, is one of the major reasons why such results often cannot be replicated in later studies. In the context of adaptive enrichment trials, methods to correct for bias due to using the same data set for developing a classifier $C(\boldsymbol{Z}, \tau_E)$ and doing parameter estimation are given, for example, by Bai et al. (2017) and Zhang et al. (2017).

The idea of GP also arises in settings where a comparative test is preceded by treatment selection, even if one assumes that patients are homogeneous. Consider a multi-arm randomized select-and-test trial of experimental treatments E_1, E_2, and E_3 and an active control C, which first selects the treatment E_{j^*} having the largest estimated E_j-versus-C effect for later comparison to C. For such two-stage phase II-III designs, as given by Thall et al. (1988), if $\hat{\theta}_{E_{j^*}} - \hat{\theta}_C$ is sufficiently large in stage 1, then additional data are obtained in a second stage by randomizing patients between the selected E_{j^*} and C, thus enriching the E_{j^*} arm, and a final test is done, based on all of the data, to decide whether E_{j^*} is superior to C. In the global null case where all four treatments

have identical means, $\theta_{E_1} = \theta_{E_2} = \theta_{E_3} = \theta_C$, the statistical estimate $\hat{\theta}_{j^*}$ will have an expected value larger than the true value. Again, this is because a selected maximum produces upward estimation bias. If a two-arm test comparing E_{j^*} to C is constructed while ignoring the preliminary selection, it will not have the nominal size and power figures, since it ignores the fact that j^* is a statistic that depends on data from all four treatment arms. The relevant quantity is not conventional power but the GP, which in this setting is the probability of (*a*) correctly identifying an E_{j^*} that truly provides an improvement over C in stage 1 and (*b*) concluding that $\theta_{E_j^*} > \theta_C$ in stage 2. Since the GP is the probability of a smaller event than just the event (stage 2) assuming that j^* is known, it is harder to achieve a numerical GP value close to conventional power figures. A similar two-stage phase II-III design based on $Y =$ survival time that allows more than one E_j to be selected for stage 2 is given by Schaid et al. (1990). This design controls the pairwise type I error rate and power when testing $\theta_{E_j^*} = \theta_C$ for each selected E_j^*. A multistage version is given by Stallard & Todd (2003).

1.4. Other Forms of Enrichment

An extreme form of enrichment is outcome-adaptive randomization (OAR), which repeatedly unbalances randomization probabilities for the treatment arms by using the interim data to favor the arm, or arms, seen to have more favorable outcomes. This continuously enriches the arms that have superior performance based on the interim data. The main motivation of OAR is to enroll a greater proportion of patients to the treatment arms that, during the trial, show higher response rates. Some unexpected properties of OAR designs are discussed in Section 7.

Another type of enrichment is done in sequentially adaptive early-phase trials that choose an optimal regime, which may be a dose, dose pair, schedule, or dose-schedule combination, for successive patient cohorts. This repeatedly enriches the regimes seen interimly to have superior outcomes in terms of the optimization criterion that is used. Acceptable regimes are further enriched by adaptive rules that drop unsafe or ineffective regimes. If the outcomes include some combination of two or more efficacy and toxicity variables, the class of designs is called phase I-II, since they hybridize conventional phase I dose escalation trials based on toxicity and phase II trials based on response. This topic is reviewed briefly by Yan et al. (2018) and Gauthier et al. (2019) and is covered extensively in the book by Yuan et al. (2016). Refinements of such designs may account prospectively for patient subgroups, allowing for regime-subgroup interactions and possibly assigning different optimal regimes to different subgroups. Lee et al. (2019) provide a utility-based phase I-II design that uses restricted randomization to adaptively optimize the dose of natural killer cells to treat hematologic malignancies. Dose optimization is done within each of six subgroups defined by disease type and severity, with subgroup-specific rules that stop accrual to unsafe doses. Lin et al. (2020) propose a phase I-II design that optimizes (dose, schedule) regimes within ordered disease subgroups based on an efficacy–toxicity tradeoff. A Bayesian phase I trial design based on time to toxicity that adaptively chooses optimal subgroup-specific doses while using latent subgroup membership variables to combine similar subgroups is given by Chapple & Thall (2018).

In the setting of multistage therapies, also known as dynamic treatment regimes (DTRs), a within-patient adaptive treatment decision may be to choose a patient's dose or treatment in the second or later stages of therapy based on the patient's previous treatments and outcomes. These within-patient sequential treatment decisions may be informed by updated covariate values or recent outcomes used as tailoring variables, as well as data from other patients. This may be regarded as within-patient sequential enrichment. The idea is to give each patient the best sequence of treatments by using the accumulating data both within and between patients. For example, Lee

Table 1 Mean survival times for each of the four possible combinations of treatment arm and biomarker status, and *E*-versus-*C* effects in the biomarker subgroups

Treatment	Biomarker status	
	Positive	Negative
E	$\theta_{E,\,pos}$	$\theta_{E,\,neg}$
C	$\theta_{C,\,pos}$	$\theta_{C,\,neg}$
E-versus-C effect[a]	Δ_{pos}	Δ_{neg}

[a]Calculated as $\Delta_{pos} = \theta_{E,\,pos} - \theta_{C,\,pos}$ and $\Delta_{neg} = \theta_{E,\,neg} - \theta_{C,\,neg}$.

et al. (2015) give a Bayesian phase I-II design that jointly optimizes the doses (d_1, d_2) of an agent given in two stages of therapy, where each d_s may be a dose or the action to not treat at any dose if, for example, the patient has experienced unacceptable toxicity at the lowest dose being considered. DTRs have been applied to optimize sequences of adaptive interventions in mobile health devices to treat behavioral disorders, drug and alcohol dependence, and chronic diseases, with methods described by Nahum-Shani et al. (2017). A randomized oncology trial designed to evaluate multistage chemotherapy regimes for advanced prostate cancer was reported by Thall et al. (2007). A sequential multiple assignment randomized trial was designed to evaluate precision medicine in which burn victims were repeatedly and adaptively rerandomized to different plastic surgery methods at multiple scheduled intervention points (Hibbard et al. 2018). There is an extensive literature on methods for optimizing DTRs (see, for example, Murphy 2005, Kosorok & Moodie 2016, and Tsiatis et al. 2019).

2. NUMERICAL EXAMPLES OF TREATMENT-BIOMARKER INTERACTIONS

There are many possible cases when using biomarkers for enrichment during a clinical trial and doing precision medicine based on its results. The following toy numerical examples are constructed to illustrate some qualitatively different cases. Consider a simple setting where $E + C$ is compared with C in terms of $Y =$ survival time and a binary biomarker, Z, is available. Each patient is either biomarker positive, $pos = (Z = 1)$, if the biomarker is detected as being present, and otherwise is biomarker negative, $neg = (Z = 0)$. **Table 1** gives the mean survival time $\theta_{\tau,\,Z}$ for each $(\tau, Z) =$ (treatment, biomarker) combination, and the E-versus-C effect in each biomarker subgroup. The $\theta_{\tau,\,Z}$s could instead represent response probabilities if $Y =$ response rather than survival were the primary outcome.

Denote the proportion of patients who test positive by $p_{pos} = \Pr(Z = 1)$. If there is no testing error, then p_{pos} is the prevalence of E-sensitive patients in the population of patients with the disease. For simplicity, assume that the test for Z is perfectly accurate, so $Z = 1$ implies that the patient must be *pos*. The expected effects of E are $\Delta_{pos} = \theta_{E,\,pos} - \theta_{C,\,pos}$ in biomarker positive patients and $\Delta_{neg} = \theta_{E,\,neg} - \theta_{C,\,neg}$ in biomarker negative patients. The overall effect of E in the patient population is the subgroup prevalence weighted average $\Delta = \Delta_{pos}\, p_{pos} + \Delta_{neg}\,(1 - p_{pos})$, and the vector of all parameters is $\boldsymbol{\theta} = (\theta_{E,pos}, \theta_{E,neg}, \theta_{C,pos}, \theta_{C,neg}, p_{pos})$.

Different numerical configurations of a statistical estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ may motivate different precision treatment decisions by a physician, depending on each patient's biomarker status. This underscores the importance of conducting a randomized trial of $E + C$ versus C that includes both *pos* and *neg* patients, although as data are accumulated during an enrichment trial, the sample sizes of the four subgroups may be changed by the adaptive decision rules. **Table 2** illustrates eight cases, each defined by fixed numerical values of $\boldsymbol{\theta}$. In case 1, E has no effect, either overall or

Table 2 Numerical illustration of possible cases involving a binary biomarker that is either positive (*pos*) or negative (*neg*) for each patient; an experimental treatment, *E*; and an active control treatment, *C*

Case	Mean survival times (months)							
	$\theta_{E,\,pos}$	$\theta_{E,\,neg}$	$\theta_{C,\,pos}$	$\theta_{C,\,neg}$	p_{pos}	Δ_{pos}	Δ_{neg}	Δ
1	24	24	24	24	NA	0	0	0
2	36	36	24	24	NA	12	12	12
3	36	24	24	24	0.10	12	0	1.2
4	84	24	24	24	0.01	60	0	0.6
5	36	24	24	24	0.50	12	0	6
6	36	30	24	24	0.50	12	6	9
7	25	24	24	24	0.50	1	0	0.5
8	84	24	24	24	0.40	60	0	24

Each case is characterized by the biomarker subgroup-specific mean survival times for *E* and *C*, and prevalence p_{pos} of *E*-positive patients. Within-subgroup *E*-versus-*C* effects are $\Delta_{pos} = \theta_{E,\,pos} - \theta_{C,\,pos}$ and $\Delta_{neg} = \theta_{E,\,neg} - \theta_{C,\,neg}$, and the overall effect is the weighted average is $\Delta = p_{pos}\Delta_{pos} + (1 - p_{pos})\Delta_{neg}$. Abbreviation: NA, not applicable.

within subgroups, and *Z* is irrelevant. In case 2, $E + C$ provides a 12-month (50%) mean survival improvement over *C* in both subgroups, but again, *Z* is irrelevant.

Case 3 illustrates a biomarker-specific effect, where $E + C$ provides a 12 month improvement over *C* if $Z = 1$, but *E* has no effect if $Z = 0$. While *Z* is very important in case 3, only the 10% of patients who are biomarker positive benefit from *E*, so if this were known, then it would not make sense to give *E* to biomarker negative patients. For example, the BRAF mutation, associated with colorectal cancer (CRC), encodes a serine/threonine protein kinase that is a downstream effector of activated KRAS. Thus, BRAF often is targeted by therapies for CRC, but the BRAF mutation has low prevalence. Hsieh et al. (2012) cited BRAF rates of 6% to 13% in Spanish CRC populations and 7% in Chinese and Greek CRC populations. An important point illustrated by case 3 in **Table 1** is that the overall average effect $\Delta = 1.2$, considered alone, is extremely misleading, since what matters for therapeutic decision making is that $\Delta_{pos} = 12$ and $\Delta_{neg} = 0$. Thus, adaptively identifying and enriching a biomarker positive subgroup by increasing its sample size during a trial and decreasing the number of *neg* patients who receive *E* are highly desirable in settings like case 3. This would be triggered by interim estimates showing that $\hat{\Delta}_{pos}$ is much larger than $\hat{\Delta}_{neg}$ and that $\hat{\pi}_{pos}$ is small. Case 4 is a more extreme version of case 3, with $\Delta_{pos} = 60$ months and $p_{pos} = 0.01$, so *E* is a home run that extends expected survival from to one to five years, but only in a tiny subpopulation of 1% of patients. For example, in case 4, a sample of 1,000 patients randomized fairly between $E + C$ and *C* would yield a subsample with an expected size of only 10 biomarker positive patients, so the reliability of an estimate $\hat{\Delta}_{pos}$ or statistical test of $\Delta_{pos} = 0$ would be very low.

Case 4 shows that a very small value of the prevalence p_{pos} may make it very difficult to make a reliable inference about Δ_{pos}, even with a large overall sample. Thus, adaptive enrichment by increasing the biomarker positive subgroup sample size is desirable to improve the reliability of inferences about the benefit Δ_{pos}. Case 5 also is like case 3, but $p_{pos} = 0.50$ rather than 0.10. The potential overall benefit of using $E + C$ for treating biomarker positive patients is much larger in case 5 because, on average, there are a lot more of them. Here, subsample size enrichment is less useful. In case 6, $E + C$ provides a substantial advantage over *C* in both subgroups, but the expected survival benefit is twice as large in biomarker positive patients, 12 versus 6 months. This shows why a randomized trial should include both biomarker positive and negative patients since, despite

what may be believed based on preclinical studies and the way that E was designed, Δ_{neg} may be meaningfully large. Case 7 is like case 5 in that $E + C$ provides a benefit over C only in biomarker positive patients, but the expected improvement is very small, only $\Delta_{pos} = 1$ month. In case 7, while a nominally significant p-value for a subgroup-specific test of $\Delta_{pos} = 0$ may be obtained with a large sample, the clinical benefit of E in terms of Δ_{pos} is a 1-month improvement in expected survival time, which renders addition of E to C nearly useless and is a waste of resources if E is expensive. In case 8, a subgroup of positive superresponders have a 60-month expected survival benefit with $E + C$, they comprise 40% of all patients, and the E effect is $\Delta_{neg} = 0$ in the 60% of patients who are biomarker negative. In this case, ideally, interim adaptive enrichment during a trial would be to conclude that $\Delta_{neg} = 0$, conclude that Δ_{pos} is large, stop the trial, and treat all future patients having $Z = 1$ with $E + C$ and all future patients having $Z = 0$ with C. Case 8 is easy to deal with statistically, but one still may go astray if a conventional design is used. If Z is ignored in case 8, as in a conventional trial, then a large estimate of the overall mean improvement $\hat{\Delta} = 24$ months, which doubles the overall mean survival time with C from 24 to 48 months, is very misleading, as in case 3. This is because adding E to C only benefits biomarker positive patients. In case 8, once reliable estimates of $\boldsymbol{\theta}$ are obtained, it would be a waste of resources to continue to treat biomarker negative patients with $E + C$ rather than C. Case 8 illustrates the importance of reliably identifying an E-sensitive subgroup if it exists, estimating its prevalence, and not mistakenly concluding that $E + C$ is superior to C in all patients when an overall beneficial effect is due entirely to a biomarker positive subgroup. In all cases where it can be inferred interimly that Δ_{pos} is substantively larger than Δ_{neg}, enrichment of the $E + C$ subgroup is beneficial both during and after the trial. Recall that all of these cases are based on the simplifying assumption that Z is a perfect classifier for E-sensitive and non-E-sensitive patients. As noted earlier, in practice, the statistical problem of determining a reliable classifier is far from trivial and is a central issue in adaptive enrichment trials. More generally, before conducting a clinical trial, one simply does not know which case is true.

Given the fact that many targeted agents are very expensive to produce, the question of whether a pharmaceutical company may consider it feasible or desirable to pursue development of a given targeted E is quite important. In many settings, the costs of preclinical experimentation to develop E are quite substantial. In terms of clinical evaluation, if the prevalence p_{pos} of E-sensitive patients turns out to be very small, then even if E turns out to be highly effective in E-sensitive patients, it may not appear to be economically worthwhile to produce the agent unless patients or insurance companies pay a very high price for treatment with E. A broader, more useful perspective is obtained by also considering the prevalence of the disease and the number of people affected. For example, if the disease has annual prevalence 0.001 in a population of 400 million people and $p_{pos} = 0.10$, then one may expect 400,000 people to have the disease each year and 40,000 of these to be sensitive to E. So the market for E is substantial despite the apparently low disease prevalence and small value of p_{pos}. Since lower values of either the disease prevalence or p_{pos} will reduce the number of people who have the disease and may benefit from E, from either a scientific or an economic viewpoint, this underscores the importance of obtaining reliable estimates of p_{pos} and Δ_{pos}. To place this in context, since the annual prevalence of influenza is about 0.05 to 0.20, although a given case of the flu may be due to numerous different strains of this class of viral diseases, the potential benefits of a highly effective vaccine or treatment targeting one or more specific strains are immense.

In some settings, optimism motivated by preclinical data about the effects of E at the molecular or cellular level, or promising results using the agent in E to treat xenografted rodents, may lead to the use of a dysfunctional clinical trial design where a randomized trial of $E + C$ versus C is conducted in biomarker positive patients only. Worse, a single-arm trial of $E + C$

alone may be conducted in what are presumed to be *E*-sensitive patients. Such presumptive enrichment is very bad scientific practice, as seen above in case 1, and may lead to serious errors. Observations at the molecular or cellular level or in experiments with rodents provide a basis for designing a clinical trial, but they cannot guarantee that a desired effect of *E* will be seen in humans. Presumptively excluding nonsensitive patients uses preclinical data as a basis for assuming that the effect Δ_{neg} in humans is negligible or 0. For a trial that does presumptive enrichment by ignoring the subgroups (*E*, *neg*) and (*C*, *neg*), inferences cannot be made about either Δ_{pos} or Δ_{neg}. Such a trial does not provide data if *Z* has little or no relationship to clinical outcome because $\Delta_{pos} - \Delta_{neg}$ is 0 or small (cases 1, 2, or 7), or if $\Delta_{pos} = \Delta_{neg} = \Delta > 0$ is clinically meaningful (case 2), or if $\Delta_{pos} > \Delta_{neg} > 0$ and Δ_{neg} is clinically meaningful (case 6). In these cases, biomarker negative patients would be deprived of the benefit of *E*. This raises the question of whether conducting such a presumptive enrichment trial based on preclinical data alone is ethically reasonable.

These numerical examples show the importance of obtaining reliable, unbiased statistical estimates of the relevant parameters in $\boldsymbol{\theta}$. This requires randomizing patients, determining an *E*-sensitive subgroup reliably if it exists, and reliably estimating comparative subgroup-specific treatment effects. In cases 3, 5, or 7 in **Table 1**, given reliable interim data, an adaptive rule to treat a larger proportion of *pos* patients, or to restrict enrollment to *pos* patients, would provide more reliable estimates of Δ_{pos}. The possibility of cases 1, 3, 4, 5, or 7 suggests the desirability of using adaptive subgroup-specific rules to terminate enrollment of *neg* patients, or all patients, due to futility. However, the risk from using subgroup-specific futility rules is that, in cases 2 or 6, incorrectly terminating accrual of *neg* patients and concluding that *E* is not superior to *C* in those patients would deprive future *Z*-negative patients of the benefit of *E*.

A major reason for conducting a clinical trial is to find out which case actually is true. **Table 1** shows that, given a reliable estimate of $\boldsymbol{\theta}$ from a well-designed trial, in some cases, physicians would have an informed way to make precision treatment decisions using each patient's *Z*. The therapeutic value of *Z* would be high in cases 3, 4, 5, and 8, where the best decisions would be to treat biomarker positive patients with $E + C$ and negative patients with *C* and to reduce treatment costs of adding *E* unnecessarily when it is not beneficial. In case 6, $E + C$ would be the best choice for all patients, but a better outcome could be expected in biomarker positive patients, and here *Z* also would be useful for predicting patient survival times.

Enrichment need not rely on frequentist tests and *p*-values to make inferences. Under a Bayesian model, adaptive enrichment rules may increase the sample size of *pos* patients if a sufficiently large value is seen for $\Pr(\Delta_{pos} > \Delta_{neg} + \delta | data)$ based on interim data, where $\delta > 0$ is a clinically meaningful improvement. This is the posterior probability, given the observed data, that the improvement due to giving *E* to *E*-sensitive patients is at least δ larger than the improvement in non-*E*-sensitive patients. A Bayesian design may also adaptively reduce the sample size of *E* or terminate enrollment entirely of *neg* patients if $\Pr(\Delta_{neg} < \epsilon | data)$ is large for a given small $\epsilon > 0$ (see, for example, Trippa et al. 2012 or Simon & Simon 2018).

Randomization is essential to obtain unbiased treatment comparisons, both overall and within biomarker-defined *pos* and *neg* subgroups. Failing to randomize between $E + C$ and *C* ignores the bias that is inherent in the comparison of a single-arm trial of $E + C$ to historical data on *C*. If a single-arm trial of $E + C$ is conducted, it provides data to estimate $\theta_{E,\,pos}$ and $\theta_{E,\,neg}$, but unbiased *E*-versus-*C* comparisons cannot be made. If historical data on *C* are available, then some sort of bias correction method may be used, such as pair matching (see Rosenbaum & Rubin 1985) or inverse probability of treatment weighting (see Robins et al. 2000). However, this is much less desirable than conducting a randomized trial in the first place.

Because adaptive enrichment designs prospectively account for possible treatment-subgroup interactions, they avoid post hoc data analyses to identify subsets where E provides substantive benefit. When subgroup analyses are not planned ahead of time, discovery of a large treatment effect in some subgroup is very difficult to defend. As explained above in the discussion of selection bias, with many subgroups, the maximum estimated treatment-subgroup effect will be large even if, in fact, there is no effect at all. Unfortunately, the practice of selecting a maximum without understanding its potentially misleading consequences is very common in medical research. There are many methods to correct for selection bias in post hoc searches for subsets with large treatment effects or treatment-covariate interactions. Bayesian hierarchical model–based approaches have been proposed by Dixon & Simon (1991, 1992).

Causal analysis is based on potential and counterfactual outcomes, which may be imagined but may not actually be observed. Causal methods have been proposed by Foster et al. (2011) by fitting random forests with cross-validation, and by Zhang et al. (2013) under a generalized linear mixed regression model for the joint distribution of each patient's two potential outcomes. Other methods for dealing with adaptive subset selection were discussed by Lipkovich et al. (2017), Ondra et al. (2016), and Lai et al. (2019). Approaches to enrichment in particular clinical trial settings were given by Rosenblum & Hanley (2017) for stroke trials, by Rosenblum et al. (2016) to combine group sequential decision making with reallocation, by Steingrimsson et al. (2019) for three-arm trials, and by Flehinger et al. (1972) for reducing the number of inferior treatments.

3. ADAPTIVE SIGNATURE DESIGNS

Freidlin & Simon (2005) proposed the adaptive signature design (ASD) for randomized trials of targeted agents, which has the goals to (*a*) determine a subset of E-sensitive patients and (*b*) compare $E + C$ to C overall and within the determined E-sensitive subset. For $\theta_i = \Pr(response)$ and candidate biomarkers $\boldsymbol{Z}_i = (Z_{i,1}, \ldots, Z_{i,K})$ in the ith patient, $i = 1, \ldots, N$, the logistic regression model $\log\{\theta_i/(1-\theta_i)\} = \mu + \tau_E(\lambda + \sum_{k=1}^{K} \gamma_k Z_{i,k})$ is assumed, so λ is the main E effect and γ_k is the interaction between E and $Z_{i,k}$. The design accrues N_s patients in stage $s = 1, 2$, uses the stage 1 data to develop a classifier, and applies it at the end of stage 2 to identify a subset of E-sensitive patients. The final analysis has two tests, (*a*) overall comparison of $E + C$ to C with a test having type I error probability α_1 and (*b*) comparison of $E + C$ to C in the subset of E-sensitive patients accrued in stage 2 by a test having type I error probability α_2. Since signature development and testing in the sensitive subset are done using disjoint subpopulations, the overall type I error is $\alpha_1 + \alpha_2$. For example, one may use $\alpha_1 = 0.03$ and $\alpha_2 = 0.02$ to have an overall type I error rate of 0.05. If either test is significant, the trial is considered positive. Freidlin & Simon (2005) suggested the following two-step algorithm for defining E-sensitive patients: Step 1 is to use the fitted logistic regression model to declare biomarker Z_j significant if $\hat{\gamma}_k > \eta_1$, where η_1 is a fixed cutoff. Step 2 is to classify a patient as E-sensitive or not for the stage 2 test, using the biomarkers selected in step 1. For example, one may declare the ith patient sensitive if $\hat{\gamma}_k Z_{i,k} > \eta_2$ for at least G of the significant biomarkers. This design has parameters $(N_1, N_2, \alpha_1, \alpha_2, \eta_1, \eta_2, G)$, which may be determined using various criteria, including achieving given power figures for the two tests or deciding how large N_1 should be for given overall $N = N_1 + N_2$. Jiang et al. (2007) extended the ASD to settings with prespecified continuous biomarkers and combined the test for an overall treatment effect with the establishment, validation, and estimation of a cutpoint for the biomarker.

Freidlin et al. (2010) proposed a cross-validated ASD (CV-ASD) to improve reliability. This was motivated by the problems that reliably determining a signature for E-sensitive patients with large-dimensional $\boldsymbol{Z}$ requires a large sample, and that if the proportion of E-sensitive patients is

low, then a large sample also is required for the test in the E-sensitive subset to have reasonably high power. The CV-ASD uses K-fold cross-validation by first randomly partitioning the sample into K validation cohorts, $V_1, \ldots, V_K$, each consisting of $M = N/K$ patients. For each k, a signature based on $\boldsymbol{Z}$ is determined using the data in the complement of V_k, the development cohort $D_k = \cup_{r=1}^{K} V_r - V_k$. The signature obtained from D_k is applied to identify the subset, S_k, of E-sensitive patients in V_k. Since the sample consists of $\cup_{k=1}^{K} V_k$, each patient is classified as E-sensitive or not, and $S = \cup_{k=1}^{K} S_k$ is the set of all sensitive patients. A permutation test comparing $E + C$ to C then is carried out in S. Freidlin et al. (2010) show that the CV-ASD has much larger overall power than the ASD to detect very large treatment effects in cases where 10% of patients are sensitive, and the response probability is $\theta_{E,\,neg} = 0.25$ with E in the nonsensitive patients and $\theta_C = 0.25$. In this case, the CV-ASD using $K = 10$ validation cohorts has overall power (*a*) 0.71 compared with 0.35 with the ASD if the response probability $\theta_{E,\,pos} = 0.80$ in E-sensitive patients and (*b*) overall power 0.91 compared with 0.60 with the ASD if $\theta_{E,\,pos} = 0.90$ in E-sensitive patients. Thus, the CV-ASD is useful when the prevalence of sensitive patients is small, but the E effect is much larger in sensitive patients compared with nonsensitive patients.

4. ENRICHMENT DESIGNS USING ONE BIOMARKER

Simon & Simon (2013) propose a general adaptive enrichment framework for developing a classifier and using it to restrict enrollment when doing treatment comparisons, with the main focus on preserving the overall type I error rate. Denote a single biomarker by Z_i, treatment indicator $\tau_i = 1$ for $E + C$ and $\tau_i = 0$ for C, and response indicator Y_i for each patient $i = 1, \ldots, N$. Let $\theta_\tau(Z_i) = \Pr(Y_i = 1 \mid \tau, Z_i)$ denote the probability of response for a patient with biomarker Z_i treated with τ. A discrimination function is defined as $f(Z_i) = I[\theta_C(Z_i) < \theta_{E+C}(Z_i)]$, so $f(Z_i) = 1$ if the response probability for a patient with biomarker Z_i is larger with $E + C$ than with C, and $f(Z_i) = 0$ otherwise. Denoting an estimator based on the interim data from m patients who have been treated and evaluated by $\hat{f}_m(\boldsymbol{Z})$, the proposed method is to (*a*) randomize m_0 patients fairly between $E + C$ and C and compute $\hat{f}_{m_0}$ from their data; (*b*) for $m > m_0$, compute the updated estimate $\hat{f}_m$ based on all accumulated data $\{(\boldsymbol{Z}_i, \tau_i, Y_i), i = 1, \ldots, m\}$; and (*c*) restrict trial entry to patients for whom $\hat{f}_m(\boldsymbol{Z}_i) = 1$, continuing this until a prespecified number N patients have been enrolled, and perform a final test. Defining the global null hypothesis H_0: $\theta_{E+C}(Z) = \theta_C(Z)$ for all Z, Simon & Simon (2013) propose the test statistic $T = \sum_{i=1}^{N} [\tau_i Y_i + (1 - \tau_i)(1 - Y_i)] =$ (number of successes with $E + C$)+ (number of failures with C). Since T follows a binomial distribution with parameters $(N, 0.5)$ under H_0, by using appropriate cutoffs from this null binomial distribution to perform a test, the type I error probability is controlled, regardless of how the discrimination function for adaptively changing the enrollment based on Z is defined.

To do adaptive threshold enrichment with a single biomarker, Z, Simon & Simon (2013) suggest a practical approach for modeling f, since there are infinitely many possibilities for a true f. They start by assuming that the E effect $\Delta(Z) = \theta_{E+C}(Z) - \theta_C(Z)$ equals either 0 or δ, is monotone nondecreasing in Z, and jumps from 0 to δ at one of a set of candidate cutpoints $c_1 < c_2 < \cdots < c_K$. At each interim decision, the cutpoint c_{j*} maximizing the likelihood of the current observed data is used to estimate the true cutpoint, and the rules $f(Z) = 1$ if $Z \geq c_{j*}$ and $f(Z) = 0$ if $Z < c_{j*}$ are used to restrict enrollment. As a practical matter, if the optimal cutpoint is not selected, but a nearby cutpoint is selected that still provides good discrimination between patients for whom the true $\Delta(Z)$ is large and those for whom it is small, then the methodology has succeeded.

Simon & Simon (2013) provide simulations of this enrichment design for clinical trials with 200 patients; the biomarker uniformly distributed between 0 and 1; candidate cutpoints $1/(K+1)$,

$2/(K+1), \ldots, K/(K+1)$; clinical outcome as a binary response; and the adaptive enrichment rule applied once interimly at 100 patients. Accrual for the last 100 patients is restricted to E-sensitive patients having biomarkers $Z_i > c_{j^*}$. Their simulations show that, compared with a design that does not restrict enrollment using the biomarker, if 75% of patients are more likely to benefit from E, then the adaptive enrichment design has much larger power to detect differences $\Delta = \theta_{E+C} - \theta_C = 0.25$ or 0.30 in most cases considered. The price for restricting enrollment to E-sensitive patients defined in this way is that, for the sample size of 200, the trial duration will be increased, and this increase in duration will be larger if the proportion p_{pos} of E-sensitive patients is smaller. Simon & Simon (2013) also discuss extensions to a group sequential design with more than one interim decision and settings with a continuous outcome rather than a binary response indicator.

5. AN ENRICHMENT DESIGN BASED ON PREDEFINED SUBGROUPS

Magnusson & Turnbull (2013) propose a phase II-III group sequential enrichment design for subgroups (GSDS). The design requires that, at the start, a partition be provided that classifies patients into disjoint subgroups $\{S_1, \ldots, S_K\}$, with the effect of E potentially differing between the subgroups. Since the K subgroups are a partition, by definition they are disjoint, and every patient belongs to exactly one subgroup. The partition may be predetermined from preclinical data, previous clinical trial results, or possibly by using a biomarker vector $\boldsymbol{Z}$. For convenience, let $\mathcal{I} = \{1, \ldots, K\}$ denote the subgroup indices, and denote the prevalence of subgroup k by p_k, so $p_1 + \cdots + p_K = 1$. A set of subgroups, $S \subset \mathcal{I}$, called a subpopulation, may be represented by the subgroup indices, and the average improvement in response rate, or mean survival time, due to E in S is $\Delta_S = \sum_{j \in S} p_j \Delta_j$. For example, if $\mathcal{I} = \{1, 2, 3, 4, 5\}$ and $S = \{1, 4, 5\}$, then $\Delta_S = p_1\Delta_1 + p_4\Delta_4 + p_5\Delta_5$.

The strategy underlying GSDS is that, after an initial stage of the trial with no restriction of accrual, a subpopulation, S, of E-sensitive patients is determined adaptively by combining subsets; thereafter, enrollment is restricted to S, and the hypotheses $H_{0,S} : \Delta_S = 0$ versus $H_{a,S} : \Delta_S > 0$ are tested using efficient score statistics. In stage 1, the GSDS begins using equal randomization (ER) and uses the efficient score statistics of the K subgroups to adaptively identify an optimal E-sensitive subpopulation, S^*. The GSDS allows H_{0,S^*} to be rejected early, after only one stage, with the conclusion that $\Delta_{S^*} > 0$; that is, E provides an improvement in the identified subpopulation S^*. If not, then stage 2 proceeds with enrollment restricted to S^*, excluding future patients who do not appear likely to benefit from E. The subsequent comparative tests of the group sequential design use all available data. Overall family-wise error rate (FWER) is defined as the maximum probability of incorrectly rejecting at least one $H_{0,S}$, where $\Delta_j = 0$ for all $j \in S$. The GSDS controls the FWER by using a bootstrap algorithm to obtain point and interval estimates of treatment effect parameters, which then are adjusted for selection bias.

6. SUBA: A BAYESIAN RANDOM PARTITION DESIGN

A subgroup-based Bayesian adaptive enrichment design (SUBA) was proposed by Xu et al. (2016) to choose each patient's treatment from a set $\tau \in \{1, \ldots, T\}$ of T possible candidates based on the patient's vector $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$ of real-valued biomarkers. SUBA identifies and successively refines a random partition, Π, of the possible values of $\boldsymbol{Z}$. The partition is used to repeatedly identify prognostic subgroups and assign each new patient to the subgroup-specific treatment that is best based on the trial's most recent data. SUBA was motivated, in part, by the fact that the targeted agent trial BATTLE (biomarker-integrated approaches of targeted therapy for lung cancer

elimination), described by Kim et al. (2011), which used five predefined biomarker subgroups and started with ER followed by OAR, resulted in a lower observed response rate with OAR than that obtained during the initial ER period. This result was the opposite of what may be expected based on the putative ethical attraction of OAR that, on average, OAR should result in better outcomes for the patients enrolled in the trial compared with ER. Possible reasons for undesirable, counterintuitive behavior of OAR are identified by the simulation studies reported by Thall & Wathen (2007), Thall et al. (2015), and Wathen & Thall (2017), which are discussed in Section 7.

SUBA is based on a binary response variable Y, the treatment set, and $\boldsymbol{Z}$. The underlying motivation is that there may exist subgroups of patients who respond differentially to each of the T treatments. The SUBA design has the goals of optimizing treatment selection for patients enrolled in the trial and optimizing the final rules used to select treatments for future patients. Rather than using predefined subgroups, SUBA derives and repeatedly refines patient subgroups adaptively during the trial as new data are obtained. A random partition $\Pi = \{\mathcal{Z}_1, \ldots, \mathcal{Z}_M\}$ of the K-dimensional set of possible $\boldsymbol{Z}$ is defined, and this determines M patient subgroups. The partition Π is obtained by constructing a tree from recursive binary splits of the elements of $\boldsymbol{Z}$, and a prior on Π is constructed using probabilities that define the algorithm in the tree's splitting rules. In the tree algorithm, a subset is not split at all with probability ν_0, and it is split into two new subsets using biomarker Z_k with probability ν_k. Each node of the tree is a subset of the K-dimensional set R^K of possible $\boldsymbol{Z}$ vectors. Given Π, the ith patient is placed in mth subgroup if their covariate vector $\boldsymbol{Z}_i$ is in $\mathcal{Z}_m$. As the data in the trial accumulate, at each step, the posterior of Π is updated based on the current data $\mathcal{D}_n = \{(\tau_i, Y_i, \boldsymbol{Z}_i) : i = 1, \ldots, n\}$ from all previous patients. The posterior predictive probability of response under treatment τ for a future $n + 1$st patient with biomarker profile $\boldsymbol{Z}_{n+1}$ is

$$q(\tau, \boldsymbol{Z}_{n+1}) = \Pr(Y_{n+1} = 1 \mid \boldsymbol{Z}_{n+1}, \tau_{n+1} = \tau, \mathcal{D}_n),$$

computed by averaging over the posterior of Π. An optimal treatment for the $n + 1$st patient is chosen by maximizing $q(\tau, \boldsymbol{Z}_{n+1})$. The probabilities $q(\tau, \boldsymbol{Z})$ also are used to construct $\boldsymbol{Z}$-specific rules for dropping inferior treatments from the trial. Thus, SUBA does sequentially adaptive treatment discovery, futility monitoring, and biomarker-specific treatment assignment.

The number of sets in the partition must be reasonably small to avoid subsets with small numbers of patients. For example, Xu et al. (2016) construct a tree for a breast cancer trial with $M = 8$ subsets. To obtain data for reliable adaptive decision making at the start, SUBA begins with a burn-in with patients randomized fairly among the T treatments, followed by continuous adaptive decision making in which (*a*) treatments having uniformly inferior $\pi_\tau(\boldsymbol{Z})$ for $\boldsymbol{Z}$ in all subgroups are discarded, and the remaining treatments are enriched; (*b*) patients are assigned to treatments adaptively as described above; and (*c*) at the end of the trial, the final partition for optimal treatment allocation is reported. A simulation study given by Xu et al. (2016) shows that SUBA compares very favorably to designs using ER or OAR and a design based on probit regression of Y on $\boldsymbol{Z}$ and τ. An important property of SUBA is that, in contrast with GSDS, rather than beginning with a partition of patient subsets that are provided at the start, SUBA derives and repeatedly refines the subsets adaptively using $\boldsymbol{Z}$.

7. OUTCOME-ADAPTIVE RANDOMIZATION

7.1. General Definitions

An early form of enrichment, OAR, was proposed by Thompson (1933), who suggested that, for success probabilities θ_A and θ_B of treatments A and B, under a Bayesian model, the next patient enrolled in a clinical trial should receive A with probability $r(A, n) = \Pr(\theta_B < \theta_A \mid data)$ and B with

probability $1 - r(A, n)$. Numerous OAR methods have been proposed, and a number of clinical trials have been conducted using various OAR methods (see, e.g., Maki et al. 2007 or Kim et al. 2011). One may define the modified randomization probability

$$r(A, n, 0.5) = \frac{r(A, n)^{0.5}}{r(A, n)^{0.5} + \{1 - r(A, n)\}^{0.5}},$$

which shrinks the probabilities toward 0.50. For example, $r(A, n) = 0.80$ gives $r(A, n, 0.5) = 0.67$, so the modified OAR proportions are 2:1 rather than 4:1. OAR may be regarded as a compromise between ER and the extreme form of enrichment in which the next patient enrolled in a trial is given the treatment having the larger estimated response rate, known as play the winner (PTW). To illustrate a fundamental flaw with PTW, suppose that initially, two patients are treated with each of A and B, and thereafter, PTW is used. If the true response probabilities are $\theta_A = 0.50$ and $\theta_B = 0.25$, and the initial data are 0/2 responses with A and 1/2 responses with B, then PTW will assign all future patients to B, the inferior treatment. PTW is an example of a greedy algorithm, and the problem illustrated above is known as stickiness. Reviews of a wide variety OAR methods are given by Rosenberger et al. (2012) and Sverdlov (2015). OAR remains quite controversial for both ethical and methodological reasons (see, for example, Korn & Freidlin 2011, Yuan & Yin 2011, or Hey & Kimmellman 2015, among many others).

7.2. Simulation Studies of Outcome-Adaptive Randomization

Thall & Wathen (2007) reported a simulation study of two-arm trials with binary response outcomes that compared several Bayesian OAR methods to a group sequential design using ER. Their simulations showed that, compared with ER, OAR methods often have a much lower probability of selecting a truly superior treatment arm (PSEL), produce much larger estimation bias, and give sample size distributions having much greater variability and skewness. For example, in a 200-patient trial, if the true response probabilities are $\theta_A = 0.25$ and $\theta_B = 0.45$, the OAR methods based on $r(A, n)$ or $r(A, n, 0.5)$ that control the type I error probability at 0.05 have PSEL figures 0.35 and 0.40, compared with PSEL = 0.86 with the group sequential design using ER. The OAR trials also have a nontrivial probability of unbalancing the achieved sample sizes, N_A and N_B, in favor of the inferior treatment, which is the opposite of the intended effect of OAR. While OAR produces a larger expected sample size for the superior treatment, the drawbacks noted above are a consequence of the greater variability and skewness of the sample size distributions produced by OAR. These effects are a consequence of the fact that OAR probabilities, such as $r(A, n)$ and $r(A, n, 0.5)$, are statistics computed from the trial data and hence are highly variable, whereas with ER, the randomization probabilities are fixed constants. In practice, often only the mean sample sizes from simulations of OAR are reported. This may be very misleading. A general caveat with OAR in designs accounting for patient heterogeneity, say by defining E-sensitive and non-E-sensitive subgroups, is that if subgroup-specific OAR probabilities are defined, then the smaller subgroup-specific sample sizes create even greater variability in the OAR probabilities. Because numerous OAR methods have been defined, the particular OAR method and specifics of a trial design can greatly affect a design's performance. Thus, simply referring to OAR as if it were one method, without providing specific details, is very bad scientific practice.

Motivated by the controversy about the merits and flaws of OAR and a suggestion that the greatest benefit of OAR may be seen in multi-arm trials, Wathen & Thall (2017) conducted a simulation study evaluating four Bayesian OAR methods and ER in five-arm trials with maximum sample size 250, either including a control arm, C, as a comparator or not. They studied designs that included (*a*) futility rules that terminate accrual to experimental arms, E_js, seen to have

response rates inferior to C, (*b*) enrichment of the remaining arms if one or more E_js are terminated early, and (*c*) selection of the best E_j^* at the end of the trial. The simulations showed that, in the case where three E_js are equivalent to C, with $\theta_1 = \theta_2 = \theta_3 = \theta_C = 0.20$, but E_4 is superior, $\theta_4 = 0.40$, the four OAR methods have probabilities 0.44, 0.46, 0.48, and 0.67 of correctly selecting E_4 compared with 0.66 with ER. In the more realistic and more difficult case where $\theta_C = 0.20, \theta_1 = 0.25, \theta_2 = 0.30, \theta_3 = 0.35$, and $\theta_4 = 0.40$, ER has the largest probability, 0.66, of correctly selecting E_4, compared with 0.40, 0.45, 0.45, and 0.61 for the four OAR methods. The best-performing OAR method restricts randomization probabilities to the interval 0.10–0.90.

8. PRACTICAL CONSIDERATIONS AND CAVEATS

Adaptive enrichment designs have the potential to greatly benefit both the patients enrolled in the trial and future patients in settings where a targeted therapy is evaluated. Ideally, a properly designed and conducted enrichment trial can lead to precision medicine in the clinic by enabling a physician to use each patient's biomarker vector $\boldsymbol{Z}$ to assist them in making the best treatment decision for that patient. To facilitate the application of a particular design, a user-friendly computer program should be made available to clinicians that allows a patient's $\boldsymbol{Z}$ vector to be input easily and outputs the optimal treatment, possibly accompanied by predicted survival distributions with each treatment that was evaluated in the trial.

In practice, clinical trials with adaptive designs do not always play out as planned, and there always is a risk of human error. For example, Thall & Wathen (2005) constructed a Bayesian design for a multicenter trial to compare two chemotherapies for metastatic soft tissue sarcomas. The design included covariate-adjusted OAR and, thus, covariate-dependent treatment assignment rules. A website with a graphical user interface (GUI) was constructed, and personnel at each participating site were trained to use the GUI prior to trial initiation. The trial results were reported by Maki et al. (2007). Despite this careful preparation, analyses of the final data showed that baseline covariates used by the OAR were input incorrectly for substantial numbers of patients at two participating sites. This corrupted the adaptive trial design by giving it incorrect data needed for the OAR-based treatment assignment and decision making. Through pure luck, no patients were harmed, essentially because there were no treatment-covariate interactions. This example illustrates what can go wrong in even the most carefully designed trials that use $\boldsymbol{Z}$ to assign treatments adaptively.

Another important issue is that poor choice of a primary endpoint may lead to flawed inferences. Thall (2020, chapter 7.1) gave an example where using a response indicator as the primary outcome may be very misleading. In the example, a new treatment, E, has response probability 0.40, doubling the probability of 0.20 with standard therapy, S, for a disease where response increases the 12-month survival probability from 0.40 to 0.60. While E may seem very promising, an elementary probability computation shows that the expected 12-month survival probabilities are 0.48 with E and 0.44 with S, so E actually gives the trivial improvement $0.48 - 0.44 = 0.04$. This illustrates that early treatment response may be a poor surrogate for survival time.

Several additional practical issues must be addressed when planning an adaptive enrichment trial. The time required to evaluate $\boldsymbol{Z}$ is critical. If it takes two weeks to evaluate $\boldsymbol{Z}$ and the accrual rate is five patients per month, then it is not feasible to delay each patient's treatment in order to apply a $\boldsymbol{Z}$-adaptive rule. In such settings, a conventional design without adaptive enrichment may be more appropriate. A major issue is the financial cost of evaluating $\boldsymbol{Z}$ since, if this is prohibitively expensive, then a trial that requires $\boldsymbol{Z}$ to make adaptive decisions is not feasible. The success of a trial using each new patient's $\boldsymbol{Z}$ and previous patients' data to make adaptive decisions also depends on timely and accurate data entry during the trial. This requires specialized computer software,

including a database and a program to apply the design's decision rules; a GUI to connect personnel in the clinic to the computer programs; and pretrial training of those involved in trial conduct. As seen in the trial reported by Maki et al. (2007), if incorrect patient data are entered into the database, then the adaptive design's rules will not function properly. Still, the data input process is very similar to what has been done for decades in conventional group sequential trials, so adaptive enrichment trials are not more difficult to conduct. The greater difficulty resides in constructing a design, which requires accounting for much greater complexity. Given the large potential benefit provided by adaptive enrichment designs, this should be well worth the effort.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The author is grateful to the reviewer for providing many helpful and constructive comments on an earlier draft.

LITERATURE CITED

Bai X, Tsiatis AA, Lu W, Song R. 2017. Optimal treatment regimes for survival endpoints using a locally-efficient doubly-robust estimator from a classification perspective. *Lifetime Data Anal.* 23:585–604

Chapple A, Thall P. 2018. Subgroup-specific dose finding in phase I clinical trials based on time to toxicity allowing adaptive subgroup combination. *Pharm. Stat.* 17:734–49

Chapple A, Thall P. 2019. A hybrid phase I-II/III clinical trial design allowing dose re-optimization in phase III (with discussion). *Biometrics* 75:371–81

Dagogo-Jack I, Shaw A. 2018. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15:81–94

Dixon DO, Simon R. 1991. Bayesian subset analysis. *Biometrics* 47:871–81

Dixon DO, Simon R. 1992. Bayesian subset analysis in a colorectal cancer clinical trial. *Stat. Med.* 11:13–22

FDA (US Food Drug Admin.). 2011. *Clinical considerations for therapeutic cancer vaccines*. Guid. Ind. FDA-2009-D-0427, US Food Drug Admin., Silver Spring, MD

Fisher R, Pusztai L, Swanton C. 2013. Cancer heterogeneity: implications for targeted therapeutics. *Br. J. Cancer* 108:479–85

Flehinger B, Louis T, Robbins H, Singer B. 1972. Reducing the number of inferior treatments in clinical trials. *PNAS* 69:2993–94

Foster J, Taylor J, Ruberg S. 2011. Subgroup identification from randomized clinical trial data. *Stat. Med.* 30:2867–80

Freidlin B, Jiang W, Simon R. 2010. The cross-validated adaptive signature design. *Clin. Cancer Res.* 16(2):691–98

Freidlin B, Simon R. 2005. Evaluation of randomized discontinuation design. *J. Clin. Oncol.* 23:5094–98

Gauthier J, Thall P, Yuan Y. 2019. Bayesian phase 1/2 trial designs and cellular immunotherapies: a practical primer. *Cell Gene Ther. Insights* 5:1483–94

Griffiths T, Ghahramani Z. 2006. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, ed. Y Weiss, B Schölkopf, JC Platt, pp. 475–82. Cambridge, MA: MIT Press

Griffiths T, Ghahramani Z. 2011. The Indian buffet process: an introduction and review. *J. Mach. Learn. Res.* 12:1185–224

Hey SP, Kimmellman J. 2015. Are outcome-adaptive allocation trials ethical? *Clin. Trials* 12:102–6

Hibbard J, Friedstat J, Thomas S, Edkins R, Hultman C, Kosorok M. 2018. Liberti: a smart study in plastic surgery. *Clin. Trials* 15:286–93

Hsieh LL, Erl TK, Chen CC, Hsieh JS, Chang JG, Liu TC. 2012. Characteristics and prevalence of KRAS, BRAF, and PIK3CA mutations in colorectal cancer by high-resolution melting analysis in Taiwanese population. *Clin. Chim. Acta* 413:1605–11

James N, Sydes M, Clarke N, Mason M, Dearnaley D, et al. 2009. Systemic therapy for advancing or metastatic prostate cancer (STAMPEDE): a multi-arm, multistage randomized controlled trial. *BJU Int.* 103:464–69

Jennison C, Turnbull B. 2007. Adaptive seamless designs: selection and prospective testing of hypotheses. *J. Biopharm. Stat.* 17:1135–61

Jiang W, Freidlin B, Simon R. 2007. Biomarker-adaptive threshold design: A procedure for evaluating treatment with possible biomarker-defined subset effect. *J. Natl. Cancer Inst.* 99:1036–43

Kim E, Herbst R, Wistuba I, Lee JJ, Blumenschein GR Jr., et al. 2011. The BATTLE trial: personalizing therapy for lung cancer. *Cancer Discov.* 1:44–53

Korn EL, Freidlin B. 2011. Outcome-adaptive randomization: Is it useful? *J. Clin. Oncol.* 29:771–76

Kosorok M, Moodie EEM, eds. 2016. *Adaptive Treatment Strategies in Practice: Planning Trials and Analyzing Data for Personalized Medicine*. Philadelphia: SIAM

Lai TL, Lavori P, Tsang K. 2019. Adaptive enrichment designs for confirmatory trials. *Stat. Med.* 38:613–24

Lee J, Mueller P, Sengupta S, Gulukota K, Ji Y. 2016. Bayesian inference for intratumor heterogeneity in mutations and copy number variation. *J. R. Stat. Soc. B* 65:547–63

Lee J, Thall PF, Ji Y, Müller P. 2015. Bayesian dose-finding in two treatment cycles based on the joint utility of efficacy and toxicity. *J. Am. Stat. Assoc.* 110:711–22

Lee J, Thall PF, Rezvani K. 2019. Optimizing natural killer cell doses for heterogeneous cancer patients based on multiple event times. *J. R. Stat. Soc. B* 68:461–74

Lin R, Thall P, Yuan Y. 2020. An adaptive trial design to optimize dose–schedule regimes with delayed outcomes. *Biometrics* 76:304–15

Lipkovich I, Dmitrienko A, D'Agostino R. 2017. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* 36:136–96

Magnusson B, Turnbull B. 2013. Group sequential enrichment design incorporating subgroup selection. *Stat. Med.* 32:2695–714

Maki R, Wathen J, Hensley M, Patel S, Priebat D, et al. 2007. An adaptively randomized phase III study of gemcitabine and docetaxel versus gemcitabine alone in patients with metastatic soft tissue sarcomas. *J. Clin. Oncol.* 25:2755–63

Murphy S. 2005. An experimental design for the development of adaptive treatment strategies. *Stat. Med.* 24:1455–81

Nahum-Shani I, Smith S, Spring B, Collins L, Witkiewitz K, et al. 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Ann. Behav. Med.* 52:446–62

Ondra T, Dmitrienko A, Friede T, Graf A, Miller F, et al. 2016. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J. Biopharm. Stat.* 26:99–119

Robins JM, Hernán MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11:550–60

Rosenbaum P, Rubin D. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* 39:33–38

Rosenberger W, Sverdlov O, Hu F. 2012. Adaptive randomization for clinical trials. *J. Biopharm. Stat.* 22:719–36

Rosenblum M, Hanley D. 2017. Topical review: adaptive enrichment designs for stroke clinical trials. *Stat. Med.* 33:2021–25

Rosenblum M, Qian T, Du Y, Qiu H, Fisher A. 2016. Multiple testing procedures for adaptive enrichment designs: combining group sequential and reallocation approaches. *Biostatistics* 17:650–62

Schaid DJ, Wieand S, Therneau TM. 1990. Optimal two stage screening designs for survival comparisons. *Biometrika* 77:507–13

Simon N, Simon R. 2013. Adaptive enrichment designs for clinical trials. *Biostatistics* 14:613–25
Simon N, Simon R. 2018. Using Bayesian modeling in frequentist adaptive enrichment designs. *Biostatistics* 19:27–41
Stallard N, Todd S. 2003. Sequential designs for phase III clinical trials incorporating treatment selection. *Stat. Med.* 22:689–703
Steingrimsson J, Betz J, Qian T, Rosenblum M. 2019. Optimized adaptive enrichment designs for three-arm trials: learning which subpopulations benefit from different treatments. *Biostatistics*. **https://doi.org/10.1093/biostatistics/kxz030**
Sverdlov O. 2015. *Modern Adaptive Randomized Clinical Trials: Statistical and Practical Aspects*. Boca Raton, FL: CRC, Taylor and Francis
Thall P. 2020. *Statistical Remedies for Medical Researchers*. New York: Springer
Thall P, Fox P, Wathen J. 2015. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann. Oncol.* 26:1621–28
Thall P, Logothetis C, Pagliaro L, Wen S, Brown M, et al. 2007. Adaptive therapy for androgen independent prostate cancer: a randomized selection trial including four regimens. *J. Natl. Cancer Inst.* 99:1613–22
Thall PF, Simon R, Ellenberg SS. 1988. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 75(2): 303–10
Thall PF, Wathen JK. 2005. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Stat. Med.* 24:1947–64
Thall PF, Wathen JK. 2007. Practical Bayesian adaptive randomisation in clinical trials. *Eur. J. Cancer* 43:859–66
Thompson W. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. *Biometrika* 25:285–94
Trippa L, Rosner G, Mueller P. 2012. Bayesian enrichment strategies for randomized discontinuation trials. *Biometrics* 68:203–11
Tsiatis A, Davidian M, Holloway S, Laber E. 2019. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. Boca Raton, FL: CRC
Wathen JK, Thall PF. 2017. A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clin. Trials* 14:432–40
Weber LM, Robinson MD. 2016. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A* 89:1084–96
Xu Y, Mueller P, Yuan Y, Gulukota K, Ji Y. 2015. Mad Bayes for tumor heterogeneity—feature allocation with exponential family sampling. *J. Am. Stat. Assoc.* 110:503–14
Xu Y, Trippa L, Mueller P, Ji Y. 2016. Subgroup-based adaptive SUBA designs for multi-arm biomarker trials. *Stat. Biosci.* 8:159–80
Yan F, Thall PF, Lu K, Gilbert M, Yuan Y. 2018. Phase I-II clinical trial design: a state-of-the-art paradigm for dose finding with novel agents. *Ann. Oncol.* 29:694–99
Yuan Y, Nguyen H, Thall P. 2016. *Bayesian Designs for Phase I–II Clinical Trials*. Boca Raton, FL: CRC
Yuan Y, Yin G. 2011. On the usefulness of outcome adaptive randomization. *J. Clin. Oncol.* 29:771–76
Zhang Z, Li M, Soon G, Greene T, Shen C. 2017. Subgroup selection in adaptive signature designs of confirmatory clinical trials. *J. R. Stat. Soc. C* 66:345–61
Zhang Z, Wang C, Nie L, Soon G. 2013. Assessing the heterogeneity of treatment effects via potential outcomes of individual patients. *J. R. Stat. Soc. C* 62:687–704

Annual Review of Statistics and Its Application

Volume 8, 2021

Contents

Modeling Player and Team Performance in Basketball
Zachary Terner and Alexander Franks 1

Graduate Education in Statistics and Data Science: The Why, When, Where, Who, and What
Marc Aerts, Geert Molenberghs, and Olivier Thas 25

Statistical Evaluation of Medical Tests
Vanda Inácio, María Xosé Rodríguez-Álvarez, and Pilar Gayoso-Diz 41

Simulation and Analysis Methods for Stochastic Compartmental Epidemic Models
Tapiwa Ganyani, Christel Faes, and Niel Hens 69

Missing Data Assumptions
Roderick J. Little 89

Consequences of Asking Sensitive Questions in Surveys
Ting Yan 109

Synthetic Data
Trivellore E. Raghunathan 129

Algorithmic Fairness: Choices, Assumptions, and Definitions
Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum 141

Online Learning Algorithms
Nicolò Cesa-Bianchi and Francesco Orabona 165

Space-Time Covariance Structures and Models
Wanfang Chen, Marc G. Genton, and Ying Sun 191

Extreme Value Analysis for Financial Risk Management
Natalia Nolde and Chen Zhou 217

Sparse Structures for Multivariate Extremes
Sebastian Engelke and Jevgenijs Ivanovs 241

Compositional Data Analysis
Michael Greenacre 271

Distance-Based Statistical Inference
Marianthi Markatou, Dimitrios Karlis, and Yuxin Ding 301

A Review of Empirical Likelihood
Nicole A. Lazar 329

Tensors in Statistics
Xuan Bi, Xiwei Tang, Yubai Yuan, Yanqing Zhang, and Annie Qu 345

Flexible Models for Complex Data with Applications
Christophe Ley, Slađana Babić, and Domien Craens 369

Adaptive Enrichment Designs in Clinical Trials
Peter F. Thall 393

Quantile Regression for Survival Data
Limin Peng 413

Statistical Applications in Educational Measurement
Hua-Hua Chang, Chun Wang, and Susu Zhang 439

Statistical Connectomics
Jaewon Chung, Eric Bridgeford, Jesús Arroyo, Benjamin D. Pedigo, Ali Saad-Eldin, Vivek Gopalakrishnan, Liang Xiang, Carey E. Priebe, and Joshua T. Vogelstein 463

Twenty-First-Century Statistical and Computational Challenges in Astrophysics
Eric D. Feigelson, Rafael S. de Souza, Emille E.O. Ishida, and Gutti Jogesh Babu 493

Errata

An online log of corrections to *Annual Review of Statistics and Its Application* articles may be found at http://www.annualreviews.org/errata/statistics