Bayesian Adaptive Methods for Clinical Trials of Targeted Agents

Peter F. Thall

Department of Biostatistics

The University of Texas, M.D. Anderson Cancer Center

Houston, Texas, U.S.A.

To appear in **"Design and Analysis of Clinical Trials for Predictive Medicine: Applications in Cancer and Other Chronic Diseases"**

S. Matsui, M. Buyse, R. Simon (eds), Chapman Hall/CRC Press.

Abstract

This chapter presents general Bayesian concepts and some specific designs for human clinical trials of targeted agents. The designs employ decision rules that use each patient's protein or gene expression biomarkers, and possibly conventional prognostic variables, to choose an individualized treatment regime that may include one or several targeted agents. The Bayesian rules are sequentially adaptive in that they are refined repeatedly during the trial by using posteriors updated as new patient data are acquired. A design's final conclusion is not one recommended treatment regime for all patients. Rather, it is a function that maps each patient's covariates to a treatment regime targeting that patient's abnormally expressed biomarkers. Illustrations include dose-finding trials, extensions of the randomized discontinuation design, and a variety of randomized comparative group sequential trial designs.

KEYWORDS: Adaptive design; Bayesian design; Dose-finding; Individualized treatment; Phase I clinical trial; Phase I/II clinical trial; Phase II/III clinical trial; Randomized discontinuation design; Subgroup analysis; Targeted therapy

1. Introduction

This chapter provides an overview of Bayesian concepts and methods for design and conduct of clinical trials of treatment regimes including targeted agents. A targeted treatment regime may consist of one agent, a combination of agents, or a sequence of agents given over multiple stages, and it also may specify the dose, schedule, or schedule-dose combinations of each agent. The illustrations include Bayesian dose-finding based on time to toxicity, dose-finding based on both toxicity and efficacy, and randomized trials to evaluate effects of multiple targeted regimes on efficacy or an event time such as progression-free survival (PFS) or overall survival (OS) time. Patient-specific or subgroup-specific decision rules are defined in terms of a vector, $\mathbf{Z} = (Z_1, \cdots, Z_p)$, of binary or quantitative biomarkers such as gene or protein expressions, and possibly a vector, $\mathbf{X} = (X_1, \cdots, X_q)$, of conventional prognostic variables such as performance status or number of prior therapies. The decision rules are refined repeatedly during trial conduct using updated posteriors as new patient data are acquired, hence are sequentially adaptive between patients.

Denote the set of agents being evaluated by $\mathcal{T} = \{\tau_1, \cdots, \tau_J\}$. Each design's final conclusion is not a single optimal element or subset of $\mathcal{T}$ to be given to all patients. Rather, a design selects or recommends "individualized" treatment combinations that choose a subset of $\mathcal{T}$ tailored to a given patient's $(\mathbf{Z}, \mathbf{X})$. An individualized treatment regime of targeted agents is a function, ρ, from the set of all $(\mathbf{Z}, \mathbf{X})$ to the set of all 2^J subsets of $\mathcal{T}$, with $\rho(\mathbf{Z}, \mathbf{X}) = \phi$, the empty set, corresponding to "Do not treat this patient," DNT. Each set $\rho(\mathbf{Z}, \mathbf{X})$ includes τ_j's that are "targeted" at one or more gene or protein biomarkers in $\mathbf{Z}$. In this sense, the designs and rules are adaptive within patients. For example, temporarily ignoring $\mathbf{X}$, if $\mathcal{T} = \{\tau_1, \tau_2, \tau_3\}$, with τ_3 conventional therapy, and $\mathbf{Z} = (Z_1, Z_2)$ are two indicators of particular cancer cell surface markers, the optimal regime may be $\rho(0,0) = \{\tau_3\}$, $\rho(1,0) = \{\tau_1, \tau_3\}$, $\rho(0,1) = \{\tau_2, \tau_3\}$ $\rho(1,1) = \{\tau_1, \tau_2, \tau_3\}$. Alternatively, if no conventional therapy exists for the disease, then $\mathcal{T} = \{\tau_1, \tau_2\}$ and $\rho^{opt}(0,0) = \phi$, no treatment. An example is well-differentiated liposarcoma, with τ_1 targeting estrogen receptor positive disease ($Z_1 = 1$) and τ_2 targeting androgen receptor positive disease ($Z_2 = 1$). If it is certain that (1) τ_j can only benefit

patients with $Z_j = 1$, (2) there are no other agents with established efficacy, and (3) toxicity is negligible, then an optimal regime would be $\rho(1,1) = \{\tau_1, \tau_2\}$, $\rho(1,0) = \tau_1$, $\rho(0,1) = \tau_2$, $\rho(0,0) = \phi$. If any of these three assumptions are not true, then one or more clinical trials to evaluate τ_1 and τ_2 must be conducted. In early phase I or I-II evaluation, τ_j may be extended to a set $\{\tau_j(d_1), \cdots, \tau_j(d_5)\}$ of the agent given at 5 possible doses.

While a great deal of science motivates use of $\mathbf{Z}$, in clinical practice it often is important to include common, well understood prognostic covariates, $\mathbf{X}$. For example, an agent τ_1 may target an overexpressed protein represented by Z_1 with the aim to disrupt a signaling pathway leading to cancer cell growth. If τ_1 also causes immunosuppression and a patient has received $X_1 = 2$ previous immunosuppressive therapies then potential adverse effects of τ_1 in that patient must be considered along with its potential benefits. Statistical formalisms and genomic/proteomic data notwithstanding, physicians have been choosing individualized treatment regimes based on patient prognostic variables for thousands of years.

To provide a concrete frame of reference, many of the designs discussed here will refer to the problem of clinically evaluating a new molecule, M, targeting the KRAS pathway in patients with locally advanced non-small-cell lung cancer (NSCLC). The patients have approximate median DFS time 8 months with standard therapy comprised of chemotherapy with carboplatin + paclitaxel and radiation therapy (chemoradiation, C). Each component of C is given at an established dose/schedule. The two patient subgroups are KRAS+ ($Z = 1$, abnormal expression, caused by a mutated KRAS gene) and KRAS- ($Z = 0$, normal KRAS gene expression or "wild type"). Two treatments are considered, C and $C + M$.

2. Design Issues for Trials of Targeted Agents

In developing a targeted anti-cancer therapy, conventionally it first is demonstrated that a molecule designed to activate or de-activate a particular target can kill cancer cells *in vitro*, then that it can shrink tumors or extend survival in rodents that have been given the targeted cancer. Such results are not sufficient to imply that the agent will be either safe or effective in humans, or what the best dose or schedule of the agent, or possibly a combination including the agent, may be. This can be assessed only by giving the agent to humans who have the

disease and observing their outcomes in a clinical trial. While this empirical point is obvious to biostatisticians and clinical oncologists, it often is missed by laboratory-based researchers who may be overly optimistic based on pre-clinical data. Because many new targeted agents turn out to be ineffective in humans, researchers should be prepared for failure, not just success. Because many new agents are not as safe as anticipated and may have severe adverse effects, formal safety monitoring/stopping rules are essential to protect patients enrolled in clinical trials. For targeted agents showing substantive anti-disease effects in humans, this may not be due to the precise mechanism initially believed, and there may be anti-disease effects in patients not having the targeted gene or protein abnormality. Trial designs must anticipate such unexpected outcomes.

As new targeted agents flood the clinical trial arena, it is essential to utilize resources efficiently. A major feasibility issue is the time required to evaluate $\mathbf{Z}$ for each newly enrolled patient, since it is undesirable to delay therapy unduly. It also is important to harvest as much useful information per patient as possible. In addition to OS time, a patient's actual clinical outcome often is a vector of longitudinal and event time variables. In treatment of solid tumors, these often include some combination of ordinal severities of different types of toxicity (cf. Bekele and Thall, 2003), time-to-toxicity (Cheung and Chappell, 2000; Yuan and Yin, 2009), and an ordinal response, such as PD = progressive disease, SD = stable disease, PR = partial response, CR = complete response for solid tumors. These variables often are evaluated repeatedly, subject to informative discontinuation of follow up due to patient drop-out or the decision by the attending physician that toxicity or PD precludes further treatment (cf. Wang, et al., 2012). In chemotherapy of acute leukemia or lymphoma, typical outcomes include the times to infection, CR, or resistant disease (Thall, Estey and Sung, 2002) and, among patients who initially achieve a CR, subsequent DFS time (Shen and Thall, 1998). In stem cell transplant (SCT), common outcomes include the times to engraftment, disease recurrence, infection, graft-versus-host-disease, or death. SCT trial data also routinely include longitudinal counts of a variety of blood cells defined in terms of their surface biomarkers. In such settings, the common practice in trial design of characterizing

patient outcome as either one binary "response" or one right-censored event time wastes a great deal of useful information. Ignoring covariates, multiple outcomes, longitudinal data, or adaptive treatment decisions made by physicians may lead to misleading conclusions about treatment effects (cf. Hernan, Brumback, and Robins, 2000; Wahed and Thall, 2013). With complex outcomes and treatment-biomarker interactions, "treatment effect" becomes a high-dimensional object, and conventional statistical methods become inadequate. Utilizing all or most of the available information is very challenging, very time consuming, and typically leads to complex statistical models and trial designs (cf. Thall and Wathen, 2005; Thall, et al., 2007; Saville, et al., 2009; Zhao, et al. 2011).

Bayesian models and posterior decision criteria provide a practical paradigm to account for multiple sources of variability, borrow strength between related subgroups, and construct designs with multiple, sequentially adaptive decision rules (cf. Thall, Simon, and Estey, 1995). Such decision rules may (1) select an optimal treatment regime or a set of regimes; (2) terminate one or more regimes, or the entire trial, due to excessive toxicity or poor efficacy; (3) change sample size based on updated estimates of design or model parameters; or (4) change randomization probabilities adaptively within subgroups to favor empirically more successful regimes (cf. Thall and Wathen, 2005, 2007). With targeted agents, each of these decisions may be made differently for individual patients or subgroups depending on $(\mathbf{Z}, \mathbf{X})$. Optimizing each patient's regime as a function of $(\mathbf{Z}, \mathbf{X})$ is the ultimate goal of individualized, targeted treatment. Since the combination of decision rules used in a trial may be quite complex, in practice it is necessary to use computer simulation of the trial as a design tool to calibrate fixed prior or hyperprior parameters, and design parameters, to ensure that the design has good frequentist properties.

In clinical research, false negative conclusions may be far more destructive errors than false positives. Despite the deeply ingrained requirement to control Type I error in conventional clinical trials, a false positive conclusion almost certainly will be discovered if an ineffective or unsafe new treatment receives regulatory approval and subsequently is used to treat patients. How detrimental this is to patients during this second, so-called "phase IV" evaluation process

depends on what other treatments are available. Because the medical research community avidly seeks therapeutic improvements, any treatment advanced by a false positive must compete with promising new treatments, and often with previous "standard" treatments. In contrast, a putatively ineffective new agent that actually could provide substantive benefit is unlikely to be explored further or given to future patients. One prominent cause of false negatives is that, because immense resources are spent on large scale trials of very few agents, many new agents simply are not evaluated clinically. A less obvious problem is the common failure to do a good job of optimizing a new agent's dose and schedule. If an ineffective or suboptimal dose or schedule is chosen in a small early phase trial, this may cripple a new agent's ability to hit its target, and subsequent evaluation of the agent's long-term anti-disease effects may be doomed to failure (cf. Thall, 2013, Section 2).

In trials of targeted agents, these problems are more severe. Because many regime-biomarker interaction parameters must be estimated, the risks of incorrect decisions are much greater. Rather than two $C + M$ and C treatment effect parameters to be evaluated and compared, the NSCLC trial has four effects on each outcome, corresponding to $(\rho, Z) = (C + M, 1)$, $(C + M, 0)$, $(C, 1)$, $(C, 0)$, and the effect of M must be evaluated in each of the two biomarker subgroups. With J targeted agents and p binary biomarkers, there are $J \times p$ agent-biomarker interactions, and potentially 2^J regimes. Since even moderately large J and p produce intractably large numbers of targeted regimes and parameters to be evaluated, inevitably clinical trial strategies must include practical, reliable methods for dimension reduction that select agents, biomarkers, and agent-biomarker combinations to evaluate.

3. Dose Finding Trials

3.1 A Phase I Trial With KRAS+ Patients

An optimal dose of M, when combined with C, may be determined in several ways, depending on clinical outcomes, ethics, and feasibility. A 24-patient phase I trial was designed to choose an "optimal" dose, d^{opt}, from four levels, based on toxicity. Toxicity was defined to include grade 4 esophagitis, esophageal perforation, dermatitis, or nausea/vomiting; and grade 3 or 4

non-hematologic toxicities including anorexia, fatigue, infection, and pneumonitis, occurring within 70 days from the start of therapy. With $C + M$, most of the risk of toxicity may be attributed to C. Due to logistical difficulties with the long, 70-day toxicity evaluation period, the TiTE-CRM was used, with target toxicity probability 0.60. This unusually high target was chosen because there is a baseline rate of .35 with C and, as in most toxicity-based phase I trials, it was believed that a dose associated with this higher toxicity rate would also provide higher efficacy in terms of longer DFS. Accrual was restricted to KRAS+ patients, motivated by the belief that it would be unethical to include KRAS- patients since they do not have the biomarker targeted by M and thus should not be exposed to potential toxicity of M before d^{opt} has been determined. The plan was to conduct a subsequent randomized trial of C versus $C + M(d^{opt})$ including both KRAS+ and KRAS- patients, where $M(d^{opt})$ is M given at the d^{opt} determined in phase I. Excluding KRAS- patients from phase I implies that it is more desirable to deprive KRAS- patients of potential benefit of M as a trade-off for excluding the additional risk of toxicity, due to M and beyond that of C, before d^{opt} has been determined. At the same time, it was believed that, given the d^{opt} for M for which the probability of toxicity with $C + M(d^{opt})$ was closest to 0.60, it would be ethical to randomize patients between C and $C + M(d^{opt})$.

3.2 A Phase I Trial With KRAS+ and KRAS- Patients

It may be argued that, given the 8 month median DFS with C, potential improvement in DFS due to adding M is a desirable trade-off for the risk of toxicity in KRAS- patients in phase I. If both KRAS+ and KRAS- patients are included in phase I, one may consider the possibility that these two subgroups may have different toxicity rates and hence different d^{opt} values. The following phase I design accommodates this.

Define G subgroups in terms of $(\mathbf{Z}, \mathbf{X})$, indexed by $g = 0, 1, ..., G-1$ with $g = 0$ the baseline subgroup. The goal is to determine d_g^{opt} for each g. Denote the numerically standardized doses by $\mathcal{D} = \{d_1, \cdots, d_m\}$. Let T = time to toxicity, T^o = observed time to toxicity or right-censoring, and $\delta = I(T^o = T)$, so (T^o, δ) is the observed outcome. Let T^* be a fixed reference time, specified by the physician, and denote $\pi(d, g, \boldsymbol{\theta}) = \Pr(T \leq T^* \mid d, g, \theta)$, the probability

of toxicity by T^* for a patient in subgroup g given dose $d \in \mathcal{D}$, where $\boldsymbol{\theta}$ is the model parameter vector. In the NSLC trial, $m = 4$ doses, and $T^* = 70$ days. One might use $Z = \text{I(KRAS +)}$, $X_1 = \text{I(good PS)}$, and $X_2 = 0$, 1, or 2 previous treatments to determine $G = 12$ subgroups. Since it is not feasible to reliably determine $G = 2 \times 2 \times 3 = 12$ optimal doses with $N_{max} = 24$, either a larger N_{max} is needed or G must be reduced. One may obtain $G = 4$ by collapsing (X_1, X_2) into $X_{1,2} = I[X_1 = 1 \text{ and } X_2 = 0]$, an indicator of favorable prognosis. In practice, a design's reliability should be investigated for several (G, N_{max}) pairs by computer simulation during the design process. The trade-off between practical limitations on N_{max} and the desire to investigate larger G or $\mathbf{Z}$ is central to trial design for individualized therapy and targeted agents.

The pdf and survivor function of $[T \mid d, g, \theta]$ are $f(t \mid d, g, \theta)$ and $\mathcal{F}(t \mid d, g, \boldsymbol{\theta}) = \Pr(T \geq t \mid d, g, \boldsymbol{\theta})$ for $t > 0$, so $\pi(d, g, \boldsymbol{\theta}) = 1$ - $\mathcal{F}(T^* \mid d, g, \boldsymbol{\theta})$. The distribution of T may be chosen based on prior knowledge about the form of the toxicity hazard function over $[0, T^*]$. Some practical choices are a Weibull, which has a monotone increasing, decreasing, or constant (exponential) hazard, a gamma, or a lognormal, which may have a non-monotone hazard. Model choice depends on flexibility, tractability and robustness, and should be studied by computer simulation. A linear term characterizing how T varies with (d, g) is

$$\eta(d, g, \theta) = \mu + \alpha d + \sum_{g=1}^{G-1} (\beta_g + d\gamma_g),$$

with $\boldsymbol{\theta} = (\mu, \alpha, \beta_1, \cdots, \beta_{G-1}\gamma_1, \cdots, \gamma_{G-1})$, so $\dim(\boldsymbol{\theta}) = 2G$. For the lognormal, $\eta(d, g, \theta)$ is the mean of $log(T)$, for the exponential or Weibull $\eta(d, g, \theta)$ acts on the log hazard domain, etc. The baseline subgroup $(g = 0)$ dose effect is α and, in each subgroup $g \geq 1$, $\mu + \beta_g$ is the main effect and $\alpha + \gamma_g$ is the dose effect, so γ_g is the dose-subgroup interaction. It is essential to include the γ_g 's since, if in fact $\gamma_g^{true} \neq 0$, then assuming a model without the γ_g 's can lead to erroneous conclusions and a design with poor performance. Each patient's data are (T^o, δ, d, g), and the likelihood is the usual form for right-censored event time data,

$$\mathcal{L}(T^o, \delta | d, g, \theta) \;=\; \{f(T^o \mid d, g, \theta)\}^{\delta} \; \{\mathcal{F}(T^o \mid d, g, \theta)\}^{1-\delta}.$$

For n patients, $data_n = \{(T_i^o, \delta_i, d_{[i]}, g_i), i = 1, ...n\}$, the likelihood is $\mathcal{L}_n = \prod_i \mathcal{L}(T_i^o, \delta_i | d_{[i]}, g_i, \theta)$,

and given prior $p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, the posterior is $p(\boldsymbol{\theta} \mid data_n) \propto \mathcal{L}_n \times p(\boldsymbol{\theta} \mid \tilde{\boldsymbol{\theta}})$, computed by Monte Carlo Markov chain methods (Robert and Cassella, 1999).

For priors, one may assume $\mu \sim N(\tilde{\mu}, \tilde{\sigma}_\mu^2)$, $\alpha \sim N_0(\tilde{\alpha}, \tilde{\sigma}_\alpha^2)$, a normal truncated below at 0 to ensure that $\alpha > 0$, $\beta_1, \cdots, \beta_{G-1} \sim iid\ N(\tilde{\beta}, \tilde{\sigma}_\beta^2)$ and $\gamma_1, \cdots, \gamma_{G-1} \sim iid\ N(\tilde{\gamma}, \tilde{\sigma}_\gamma^2)$, with the additional constraints that all $\alpha + \gamma_g > 0$. In the lognormal case where $log(T) \sim N(\eta(d, g, \theta), \sigma_T^2)$ the variance σ_T^2 may have an inverse gamma or uniform prior. There are several strategies for establishing $\tilde{\boldsymbol{\theta}}$ (cf. Thall and Cook, 2004; Thall et al. 2011). One approach is to elicit prior means of $\pi(d_r, g, \boldsymbol{\theta})$ for all mG pairs of (d_r, g), use nonlinear least squares or pseudo-sampling (Thall and Nguyen, 2012) to solve for the hyperparameter means, and calibrate the hyperparameter variances during the computer simulation to obtain $\tilde{\boldsymbol{\theta}}$ that gives sensible priors for the $\pi(d, g, \boldsymbol{\theta})$'s, and a design with good operating characteristics. Prior effective sample size (Morita, Thall and Mueller, 2008, 2010) may be used as a tool in this process.

Generalizing the TiTE-CRM, given a fixed target π^*, each $d^{opt}(g) = d^{opt}(g, data)$ may be defined as the dose minimizing $|\pi^* - \pi(d_r, g, data)|$, where $\pi(d_r, g, data) = E\{\pi(d_r, g, \theta) \mid data\}$, $r = 1, ..., m$. If desired, different subgroup-specific fixed targets $\pi^*(g), g = 0, 1, ..., G-1$ may be specified. E.g., in the NSCLC trial with $G = 4$ based on $Z = \{+1, -1\}$ and $X_{1,2}$, one may use $\pi_g^* = .35$ in the two subgroups with $X_{1,2} = 1$ and $\pi_g^* = .50$ in the two subgroups with $X_{1,2} = 0$. The simplest model ignores $\mathbf{X}$ and has $G = 2$, so $\boldsymbol{\theta} = (\mu, \alpha, \beta_1, \gamma_1)$.

The trial may be conducted as follows. In subgroup g, treat the first 3 patients enrolled at that subgroup's specified starting dose, and for each patient thereafter give the dose $d^{opt}(g, data_n)$ based on the current posterior using $data_n$ from all subgroups. One also may impose the constraint that, within each subgroup, an untried dose level may not be skipped when escalating. In subgroup g, if the lowest dose is unacceptably toxic, formally $\Pr\{\pi(d, g, \boldsymbol{\theta}) > \pi_g^* \mid data_n\} > p_{g,U}$, then accrual to that subgroup is terminated with no dose selected; otherwise, at the end of the trial, $d^{opt}(data_{g,N_{max}})$ is chosen.

3.3 A Phase I-II Trial With Both KRAS+ and KRAS- Patients

A Bayesian phase I-II method that bases dose-finding on $\mathbf{Y} = (Y_E, Y_T)$, where Y_T indicates

toxicity and Y_E indicates efficacy, accounting for prognostic covariates $\mathbf{X}$, was proposed by Thall, Nguyen, and Estey (2008). This may be extended to include a binary biomarker, Z, as follows. For a patient with covariates $(Z, \mathbf{X})$ treated with dose d, let $\pi_k(d, Z, \mathbf{X}, \boldsymbol{\theta})$ $= \Pr(Y_k = 1 \mid d, Z, \mathbf{X}, \boldsymbol{\theta})$, $k = E, T$, with $\boldsymbol{\pi}(d, Z, \mathbf{X}, \boldsymbol{\theta}) = (\pi_E(d, Z, \mathbf{X}, \boldsymbol{\theta}), \pi_T(d, Z, \mathbf{X}, \boldsymbol{\theta}))$. Denote these for brevity by π_E, π_T and $\boldsymbol{\pi}$ when no meaning is lost. The method requires an informative prior on $\mathbf{X}$ effect parameters, obtained from historical data. In contrast, non-informative priors on any effects associated with either Z or d should be used. Rather than choosing one best dose, the trial data are used to select optimal $(Z, \mathbf{X})$-specific doses.

The data from the trial's first n patients are $\mathcal{D}_n = \{(\mathbf{Y}_i, Z_i, \mathbf{X}_i, d_{[i]}), i = 1, \cdots, n\}$. Denote the historical data by $\mathcal{H} = \{(\mathbf{Y}_i, \mathbf{X}_i, \tau_{[i]}), i = 1, \cdots, n_H\}$, where $\{\tau_1, \cdots, \tau_m\}$ are historical treatments and $\tau_{[i]}$ is the i^{th} patient's treatment. Unsubscripted τ denotes either a dose or historical treatment. Denote $\mathbf{X}^+ = (Z, \mathbf{X})$. The following Bayesian model provides a basis for using $\mathcal{H}$ to learn about covariate effects and, during the trial, account for joint effects of $(d, \mathbf{X}^+)$ on $\boldsymbol{\pi}$ based on $\mathcal{D}_n^{\mathcal{H}} = \mathcal{D}_n \cup \mathcal{H}$. For a patient with covariates $\mathbf{X}^+ = (Z, \mathbf{X})$ treated with τ, let $\pi_{a,b}(\tau, \mathbf{X}^+, \boldsymbol{\theta}) = \Pr(Y_E = a,\ Y_T = b \mid \tau, \mathbf{X}^+, \boldsymbol{\theta})$, for $a, b \in \{0, 1\}$, with $\pi_k(\tau, \mathbf{X}^+, \boldsymbol{\theta}) = \Pr(Y_k = 1 \mid \tau, \mathbf{X}^+, \boldsymbol{\theta})$ for $k = E, T$. For link function ϕ, denote the linear terms $\eta_k = \phi(\pi_k)$. A model is determined by the marginals $\pi_E = \phi^{-1}(\eta_E)$ and $\pi_T = \phi^{-1}(\eta_T)$ and one association parameter, ψ. For a bivariate model, one may use Gumbel-Morgenstern copula to obtain

$$\pi_{a,b} \ = \ \pi_E^a \, (1 - \pi_E)^{1-a} \, \pi_T^b \, (1 - \pi_T)^{1-b} + (-1)^{a+b} \, \psi \, \pi_E \, (1 - \pi_E) \, \pi_T \, (1 - \pi_T), \tag{1}$$

with $-1 \leq \psi \leq 1$. For fitting $\mathcal{H}$, the linear terms are

$$\eta_k(\tau_r, \mathbf{X}, \boldsymbol{\theta}) = \mu_{k,r} + \boldsymbol{\beta}_k \mathbf{X} + \boldsymbol{\xi}_{k,r} \mathbf{X}, \qquad \text{for } r = 1, \cdots, m_H \ \text{ and } k = E, T. \tag{2}$$

The covariate main effects are $\boldsymbol{\beta}_k = (\beta_{k,1}, \cdots, \beta_{k,q})$, interactions between $\mathbf{X}$ and historical treatment τ_r are $\boldsymbol{\xi}_{k,r} = (\xi_{k,r,1}, \cdots, \xi_{k,r,q})$, and the m_H historical treatment main effects are $\boldsymbol{\mu}_k$ $= (\mu_{k,1}, \cdots, \mu_{k,m_H})$. For the trial data, the linear terms are

$$\eta_k(d, \mathbf{X}^+, \boldsymbol{\theta}) \ = \ \phi_k(d, \boldsymbol{\alpha}_k) + \boldsymbol{\beta}_k^+ \mathbf{X}^+ + d \, \boldsymbol{\gamma}_k^+ \mathbf{X}^+, \qquad \text{for } k = E, T. \tag{3}$$

For each k, covariate main effects are $\boldsymbol{\beta}_k^+ = (\beta_{k,Z}, \boldsymbol{\beta}_k)$, and dose-covariate interactions are $\boldsymbol{\gamma}_k^+ = (\gamma_{k,Z}, \boldsymbol{\gamma}_k)$. Main dose effects on π_E and π_T are characterized by $\phi_E(d, \boldsymbol{\alpha}_E)$ and $\phi_T(d, \boldsymbol{\alpha}_T)$, which should be formulated to reflect the application. For cytotoxic agents, $\phi_T(x, \boldsymbol{\alpha}_T) = \alpha_{T,0} + \alpha_{T,1}x$ with $\Pr(\alpha_{T,1} > 0) = 1$ ensures $\pi_T(d, \mathbf{Z}, \boldsymbol{\theta})$ increases in d, while π_k non-monotone in d may be appropriate for biologic agents.

The likelihood for the current trial data is

$$\mathcal{L}(\mathcal{D}_n \mid \boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} \prod_{a=0}^{1} \prod_{b=0}^{1} \{\pi_{a,b}(d_{[i]}, \mathbf{Z}_i, \boldsymbol{\theta})\}^{\mathbf{1}\{\mathbf{Y}_i=(a,b)\}}.$$

and the posterior based on $\mathcal{D}_n^{\mathcal{H}}$ is

$$p(\boldsymbol{\alpha}, \boldsymbol{\gamma}^+, \boldsymbol{\beta}^+, \psi \mid \mathcal{D}_n^{\mathcal{H}}) \;\propto\; \mathcal{L}(\mathcal{D}_n \mid \boldsymbol{\theta})\, p(\boldsymbol{\alpha}, \boldsymbol{\gamma}^+)\, p(\boldsymbol{\beta}, \psi \mid \mathcal{H}). \tag{4}$$

To determine a prior for the model used in trial conduct, one starts by fitting $\mathcal{H}$ to obtain an informative posterior $p(\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\xi}, \psi \mid \mathcal{H})$. The marginal posterior $p(\boldsymbol{\beta}, \psi \mid \mathcal{H})$ for the prognostic covariates is used as an informative prior on $(\boldsymbol{\beta}, \psi)$ at the start of the trial. Since nothing is known about effects $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ of the experimental agent, and $\beta_{k,Z}$, $\gamma_{k,Z}$ are biomarker effects, their priors all should be non-informative. For prior means, set $\mathrm{E}(\boldsymbol{\gamma}^+) = \mathbf{0}$, and obtain prior $\mathrm{E}(\boldsymbol{\alpha})$ by eliciting means of $\pi_E(d_j, \mathbf{X}^+, \boldsymbol{\theta})$ and $\pi_T(d_j, \mathbf{X}^+, \boldsymbol{\theta})$ at several values of d_j and solving for $\mathrm{E}(\boldsymbol{\alpha})$, using one of the least squares or pseudo sampling methods noted earlier. As before, prior variances may be calibrated to control the ESS of the priors on $p\{\pi_k(d_j, \mathbf{X}^+, \boldsymbol{\theta})\}$ for all $k = E, T$ and d_j, and obtain a design with good OCs.

During the trial, $(Z, \boldsymbol{X})$-specific doses are chosen adaptively using quantities computed from the posterior, $p(\boldsymbol{\alpha}, \boldsymbol{\gamma}^+, \boldsymbol{\beta}^+, \psi \mid \mathcal{D}_n^{\mathcal{H}})$. To account for both d and $\mathbf{X}^+$, two decision criteria are used. The first determines whether d is *acceptable* for given $(Z, \mathbf{X})$. The second is the *desirability* of each d given $(Z, \mathbf{X})$, using a function quantifying trade-off between π_E and π_T. Constructing covariate-specific acceptability bounds from elicited values, while straightforward, is somewhat involved. Several possible geometric methods may be used to define the desirability $\zeta_n(d, Z, \mathbf{X})$ of d for a patient with biomarker Z and prognostic covariates $\mathbf{X}$, based on efficacy-toxicity trade-offs. Detailed explanations are given in Thall, Nguyen and Estey (2008). Given $\mathcal{D}_n$, the set $\mathcal{A}_n(\mathbf{X}^+)$ of acceptable doses for a patient with covariates

$\mathbf{X}^+ = (Z, \mathbf{X})$ consists of all d satisfying

$$\Pr[\pi_E(d, \mathbf{X}^+, \boldsymbol{\theta}) < \underline{\pi}_E(\mathbf{X}^+) \mid \mathcal{D}_n^{\mathcal{H}}] < p_E \text{ and } \Pr[\pi_T(d, \mathbf{X}^+, \boldsymbol{\theta}) > \overline{\pi}_T(\mathbf{X})^+ \mid \mathcal{D}_n^{\mathcal{H}}] < p_T. \quad (5)$$

The cut-offs p_T and p_E should be calibrated to obtain good OCs. During the trial, $\mathcal{A}_n(\mathbf{X}^+)$ changes adaptively for each $\mathbf{X}^+ = (Z, X_{1,2})$, and if $\mathcal{A}_n(\mathbf{X}^+)$ is empty no dose is acceptable for that patient. To conduct the trial, $\mathcal{A}_n(\mathbf{X}^+)$ is computed for each new patient, and the following decision rules are applied. If $\mathcal{A}_n(\mathbf{X}^+)$ is empty, the patient is not treated. If $\mathcal{A}_n(\mathbf{X}^+)$ consists of a single dose then that dose is used by default. If two or more dose are acceptable, the the dose maximizing $\zeta_n(d, Z, X_{1,2})$ is given.

As an illustration, using $(Z, X_{1,2})$ from the NSCLC trial, suppose the doses of M are $\{100, 200, 300, 400, 500\}$ mg/m^2. A possible set of optimal doses if the agent is active in both biomarker subgroups, the KRAS+ patients need less of the agent to obtain the same anti-disease effect, and Good prognosis patients can tolerate a higher dose, is as follows:

Z	$X_{1,2}$	$d^{opt}(Z, X_{1,2})$
KRAS+	Good	400
KRAS+	Poor	200
KRAS-	Good	500
KRAS-	Poor	300

Using $d^{opt}(Z, X_{1,2})$ is an individualized version of targeted agent M when administered in combination with C.

4. A General Structure for Learning, Refining, and Confirming

It is useful to think about the process of developing and clinically evaluating new treatment regimes as having "learning" and "confirmation" stages (cf. Sheiner, 1997). In the conventional paradigm, one may be consider phases I and II learning and phase III confirmation. A sequence of trials of targeted agents is more complex, with more stages, that may overlap. Learning what belongs in $\mathbf{Z}$ via genomics/proteomics, "discovery," is not the same thing as using $\mathbf{Z}$ to learn about effects of ρ on clinical outcomes, although they certainly are related. A possible multi-stage strategy for the process of clinical evaluation is as follows. In practice,

each ρ must include a small number of elements of $\mathcal{T}$ in order to be therapeutically feasible. E.g., a five-agent combination may be suggested by $\mathbf{Z}$, but new five-agent dose combinations are very difficult to ethically and feasibly evaluate for safety in humans.

Delivery Optimization. Based on early safety and possibly efficacy, determine or optimize dose or dose-schedule, for each $\tau \in \mathcal{T}$, possibly certain 2-agent or 3-agent combinations, possibly combined with standard therapy, such as $C + M$ in the NSCLC example.

Randomized Comparative Evaluation. Following delivery optimization of all $\rho(\mathbf{Z})$ to be studied, use group sequential (GS) decision-making to comparatively evaluate the regimes using *weeding, selection*, and *confirmation*. *Weeding* is a process of dropping agents, subgroups, or agent-subgroup combinations due to either low efficacy or excessive toxicity. Adaptive "futility" rules are used to do this. *Selection* is a process of choosing a feasibly small set of ρ that are promising within specific $\mathbf{Z}$ subgroups. *Confirmation* aims to obtain conclusions regarding effects of single or multi-agent regimes within specific $\mathbf{Z}$ subgroups that will motivate final decisions or actions, such as promulgating conclusions about what $\mathbf{Z}$ and $\rho(\mathbf{Z})$ are a reliable basis for clinical practice. One may think of weeding as dropping the lower end and selection as moving forward the upper end of an ordered set of of treatment regimes, in subgroup $\mathbf{Z}$, based on a treatment effect parameter $\theta(\rho, \mathbf{Z})$, where the ordering may vary substantially with $\mathbf{Z}$. A Bayesian subgroup-specific futility rule stops assignment of ρ in subgroup Z if $\theta(\rho, \mathbf{Z})$ is likely to be substantively smaller than $\theta(\rho', \mathbf{Z})$ for at least one $\rho' \neq \rho$, according to some posterior criterion. Similarly, ρ may be selected in subgroup Z if, *a posteriori*, $\theta(\rho, \mathbf{Z})$ is likely to be larger than most $\theta(\rho', \mathbf{Z})$ for $\rho' \neq \rho$.

5. Two Arm Trials with Biomarker Subgroups

5.1 A Two-Arm NSCLC Trial

Comparing $C + M$ to C while accounting for one binary Z illustrates a simple 2×2 case. Given an individualized dose function for M, such as $d^{opt}(Z, X_{1,2})$ in the previous illustration, consider the question of whether $C + M$ improves $T =$ PFS time compared to C. Denote the

indicator of $C + M$ by ρ. To simplify the discussion, we suppress the fact that ρ and M are functions of $(Z, \mathbf{X})$, and only consider effects of Z. Suppose the distribution of $[T \mid \rho, Z, \theta]$ has been determined by goodness-of-fit analyses of historical data, and assume that the model's linear term takes the form

$$\eta(\rho, Z, \boldsymbol{\theta}) \;=\; \theta_0 + \theta_1\rho + \theta_2 Z + \theta_{12}\rho Z. \tag{6}$$

For example, denoting $\mu_{\tau,Z}$= $E(T \mid \tau, Z, \theta)$, under an exponential distribution $\eta(\tau, Z, \boldsymbol{\theta}) = \log(\mu_{\tau,Z})$, under a lognormal $\eta(\tau, Z, \boldsymbol{\theta}) = \mathrm{E}\{\log(T) \mid \tau, Z, \theta\}$, and so on. For $Z = 0$, the effect of adding the optimized targeted agent M to C is θ_1, while for $Z = 1$ this effect is $\theta_1 + \theta_{12}$, and θ_{12} is the *RAS-M* interaction. Assume that the parameters are defined so that larger θ_1 or θ_{12} correspond to superiority (longer mean PFS time) with $C + M$ compared to M. The assumption that M will have no effect in KRAS- patients says $\theta_1 = 0$. Under this assumption, one would conduct a randomized trial comparing $C+M$ to C in KRAS+ patients only. However, if in fact M is effective in KRAS- patients ($\theta_1 > 0$), excluding KRAS- patients would guarantee a false negative in this subgroup. The further assumption that $C + M$ will provide a substantive improvement over C in KRAS+ patients ($\theta_{12} > \delta$ for large $\delta > 0$) implies that only $C + M$ should be given to KRAS+ patients, and a randomized clinical trial should not be conducted since giving C alone would be unethical. Such assumptions replace clinical evaluation of a targeted agent in humans by subjective inferences based on pre-clinical data in rodents or cell lines. Many laboratory-based scientists have precisely this sort of belief about a targeted agent that they have developed in laboratory experiments.

5.2 Designs that Deal with Biomarker-Subgroup Interactions

Maitournam and Simon (2005) compared a conventional randomized trial design, an untargeted design, to a targeted design restricted to patients who are biomarker positive (Z =1), and showed that the relative power of the two approaches depends on the biomarker prevalence, Pr(Z =1), the magnitudes of the treatment effects in the two subgroups, and reliability of evaluation of Z, i.e. assay sensitivity and specificity. For multiple biomarkers, $\mathbf{Z}$, that all are putatively associated with sensitivity to a targeted agent, Freidlin and Simon

(2005) proposed a two-stage design where a "biomarker positive" classifier is developed in stage 1 and two tests for the effect of M are done in stage 2, one overall and the other in the biomarker sensitive subgroup. In the survival time setting with one biomarker, Karuri and Simon (2012) proposed a logically similar two-stage Bayesian design, with point mass distributions on comparative treatment effects. Prior to stage 2 the design drops subgroups, either both or only biomarker negative patients, if the stage 1 data show it is likely that there is no treatment effect in the subgroup.

Much of the literature on frequentist designs is devoted to the technical problem of determining design parameters given pre-specified GS test size and power. Song and Chi (2007) applied closed testing to obtain a two-stage procedure wherein a test of overall effect is carried out and, if the global null is rejected, a test is then carried out in a subgroup of interest, allowing treatment-subgroup interaction. For binary outcomes, Tang and Zhao (2013) randomized patients between two treatment arms in two stages using unbalanced randomization with probabilities chosen to minimize the expected overall number of failures, given specified size and power, also accounting for classification error in Z.

5.3 Two-Arm Bayesian Designs with Biomarker-Subgroup Interactions

A Bayesian randomized trial to compare $C + M$ to C in two subgroups defined by Z may be conducted is as follows. All of the following rules may be applied group-sequentially. The monitoring schedule and sample size are very important since they play central roles in determining the design's operating characteristics, along with the decision rules. Futility rules are applied throughout the trial, but superiority rules are applied only in the latter portion of the trial. Given a minimum desired improvement δ_1 in mean DFS from adding M to C, a futility rule stops accrual to subgroup Z if

$$\Pr\{\mu_{C+M,Z} > \mu_{C,Z} + \delta_1 \mid data_n\} < p_{L,Z}, \tag{7}$$

for small lower decision cut-off $p_{L,Z}$. Since patient safety is never a secondary concern in an ethical clinical trial, similar stopping rules for adverse events may be constructed, and should be applied throughout the trial (cf. Thall, Simon, and Estey, 1995). E.g., if $\pi_{\rho,Z}$ denotes the

probability of toxicity with $\rho = C + M$ or C in subgroup Z, and $\bar{\pi}$ is a fixed upper limit based on clinical experience or historical data, then accrual should be stopped in subgroup Z if

$$\Pr\{\pi_{\rho,Z} > \bar{\pi} \mid data_n\} > p_{U,Z,tox}. \tag{8}$$

One may declare $C + M$ promising compared to C in subgroup Z if

$$\Pr(\mu_{C+M,Z} > \mu_{C,Z} + \delta_2 \mid data_n) > p_{U,Z}, \tag{9}$$

for slightly larger $\delta_2 > \delta_1$, using upper decision cut-off $p_{U,Z}$. The same sort of criterion may be used confirm that $C + M$ is superior to C in subgroup Z for substantially larger $\delta_3 > \delta_2$. Given a monitoring schedule, the cut-offs of these one-sided decision rules and sample size should be calibrated via computer simulation to obtain desired overall type I error and power, and possibly also each within-subgroup false negative rate. If desired, a symmetric two-sided version of this procedure could be defined by including similar rules with the roles of C and $C + M$ reversed. Rules of this sort may be replaced by analogous Bayesian rules based on predictive probabilities (cf. Anderson, 1999) or Bayes factors (cf. Spiegelhalter, et al., 2004).

If no standard therapy exists and one wishes to evaluate two targeted agents, $\mathcal{T} = \{\tau_1, \tau_2\}$, with corresponding biomarker indicators $\mathbf{Z} = = (Z_1, Z_2)$, then there are 4 biomarker subgroups, $\mathbf{Z}$ = (1,1), (1,0), (0,1), and (0,0). A modified version of the above design with symmetric rules randomizes patients between τ_1 and τ_2, and uses futility rules to stop accrual to τ_j in subgroup $\mathbf{Z}$ if

$$\Pr\{\mu_{\tau_j,\mathbf{Z}} > \mu_0 + \delta_1 \mid data_n\} < p_{L,\mathbf{Z}}, \tag{10}$$

where μ_0 is the historical mean DFS. There are 8 such futility rules, one for each combination of agent τ_j and biomarker signature $\mathbf{Z}$. Weeding out unpromising τ_j-$\mathbf{Z}$ combinations is important so that the remaining combinations may be enriched. If neither τ_1 nor τ_2 is stopped due to futility in subgroup $\mathbf{Z}$, then τ_1 may be declared superior to τ_2 in this subgroup if

$$\Pr\{\mu_{\tau_1,\mathbf{Z}} > \mu_{\tau_2,\mathbf{Z}} + \delta_3 \mid data_n\} > p_{U,\mathbf{Z}}, \tag{11}$$

with the symmetric subgroup-specific rule used to declare τ_2 superior to τ_1. An elaboration of this design might also include the combination $\tau_1 + \tau_2$ for a three-arm trial, and thus require

a more complex model and three pairwise comparative rules of the form (11), or possibly posterior probabilities of the form $\Pr\{\mu_{\tau_1+\tau_2,Z} > \max\{\mu_{\tau_1,Z}, \mu_{\tau_2,Z}\} + \delta_S \mid data_n\}$. A very different design is motivated by the assumption that τ_j can benefit only patients with $Z_j = 1$ for each $j = 1, 2$. This design would not randomize, but rather would use $\tau_1 + \tau_2$ to treat all patients with (Z_1, Z_2) = (1,1), τ_1 to treat all patients with (Z_1, Z_2) = (1,0), and τ_2 to treat all patients with (Z_1, Z_2) = (0,1).

The decision cut-offs may be elaborated as parametric functions that vary with sample size, to facilitate optimization with regard to expected sample size for given overall Type I and Type II error rates. For a Bayesian two-arm trial to compare survival or PFS time with adaptive model selection, Wathen and Thall (2008) use the boundary functions $p_U(data_n)$ $= a_U - b_U(N^+(data_n)/N)^{c_U}$ and $p_L(data_n) = a_L + b_L(N^+(data_n)/N)^{c_L}$, where $N^+(data_n)$ is the number of events observed through n patients, and $p_L(data_n) \leq p_U(data_n)$. To adapt their decision rules to accommodate biomarker subgroups, denote $p_{\tau_1>\tau_2,Z,\delta,n} = \Pr(\mu_{\tau_1,Z} > \mu_{\tau_2,Z} + \delta \mid data_n)$ and $p_{\tau_2>\tau_1,Z,\delta,n}$ similarly. Decision rules may be defined as follows, where δ_1 is a minimal $|\mu_{\tau_1,Z} > \mu_{\tau_2,Z}|$ effect and δ_2 is a larger, clinically meaningful effect.

1: *Futility.* If $\max\{p_{\tau_1>\tau_2,Z,\delta_1,n}, p_{\tau_2>\tau_1,Z,\delta_1,n}\} < p_L(data_n)$ then stop accrual in subgroup Z and conclude there is no meaningful $\tau_1 - \tau_2$ effect in this subgroup.

2: *Superiority.* If $p_{\tau_1>\tau_2,Z,\delta_2,n} > p_U(data_n) > p_{\tau_2>\tau_1,Z,\delta_2,n}$ then stop accrual in subgroup Z and conclude $\tau_1 > \tau_2$ in this subgroup.

Otherwise, continue accrual in subgroup Z. If accrual is stopped in one or more subgroups, the overall sample size should not be reduced, so that the remaining subgroups are enriched. In practice, the rules are applied group sequentially at successive points where $N^+(data_n)$ equals pre-specified values. As suggested earlier, with 4 or more subgroups, it may be useful to only apply the futility rules initially, and apply superiority rules for larger n.

5.4 Potential Consequences of Ignoring Subgroups

Most conventional clinical trial designs implicitly assume homogeneity by ignoring subgroups. Statistical models and methods that ignore subgroups produce decisions based on treatment

effect estimates that actually are averages across subgroups. A simple example of the consequences of ignoring patient heterogeneity in the single-arm, phase II setting of an experimental treatment E versus historical control C was given by Wathen et al. (2008). Denote the probability of response with treatment $\tau = E$ or C in subgroup $Z = 0$ or 1 by $\pi_{\tau,Z}$. Under a Bayesian model with $\text{logit}\{\pi_{\tau,Z}\} = \eta_{\tau,Z}$ of the form (6), accrual is stopped in subgroup Z if $\Pr(\pi_{E,Z} > \pi_{C,Z} + \delta_Z \mid data) < p_Z$. A simulation study was conducted of a 100 patient trial with $\Pr(Z = 0) = \Pr(Z = 1) = 0.50$, and prior means 0.25 for both $\pi_{E,0}$ and $\pi_{C,0}$, and 0.45 for both $\pi_{E,1}$ and $\pi_{C,1}$. These correspond to historical response rates of 25% and 45% in the two subgroups. The targeted improvements were $\delta_0 = \delta_1 = 0.15$, and the decision cut-offs p_0, p_1 were calibrated to ensure within-subgroup incorrect stopping probabilities 0.10 for $Z = 1$ if $\pi_{E,1}^{true} = 0.45 + 0.15 = 0.60$, and also 0.10 for $Z = 0$ if $\pi_{E,0}^{true} = 0.25 + 0.15 = 0.40$. Comparison of this design to the analogous design that ignores Z and uses null mean $\pi_E = (0.25 + 0.45)/2 = 0.35$ showed that, in the treatment-subgroup interaction case where $\pi_{E,1}^{true} = 0.60$ (E gives improvement 0.15 over C if $Z = 1$) and $\pi_{E,1}^{true} = 0.25$ (E gives no improvement over C if $Z = 0$) the design ignoring subgroups stopped the trial and rejected E with probability 0.42. This implies a false negative probability of 0.42 if $Z = 1$ and a false positive probability of 1 - 0.42 = 0.58 if $Z = 0$ in this case. In practical terms, with this treatment-subgroup interaction, one could do about as well as a design that ignores subgroups by not bothering to conduct a clinical trial and simply flipping a coin. Similar results hold for randomized trials, and also were found by Thall, Nguyen and Estey (2008) in the phase I-II dose-finding setting for dose-covariate interactions. The general point is extremely important. If there is in fact a treatment-subgroup interaction, then ignoring subgroups can produce extremely unreliable conclusions. This is particularly problematic for trials of multiple targeted agents since a vector of J binary biomarkers implies up to 2^J subgroups, although they are far from being disjoint.

6. Randomized Discontinuation Designs

The randomized discontinuation design (RDD, Kopec et al., 1993; Rosner, et al., 2002) for targeted agents that aim to achieve stable disease (SD) or better categorizes patient outcome

Y_s at each of $s = 1$ or 2 stages of therapy as R (response), SD, or PD (progressive disease). All patients are given the targeted agent, τ, in stage 1. If $Y_1 = R$ then τ is given in stage 2; if $Y_1 = PD$ then the patient is taken off study; if $Y_1 = SD$ then the patient is randomized between τ and placebo (discontinuation). In practice, PD also includes toxicity that precludes further treatment with τ. This is an example of an "enrichment" design in the sense that patients more likely to benefit from τ are more likely to be kept on the agent for stage 2. Rosner, et al. (2002) presented the RDD in the context of cytostatic agents, where stable disease or better, $SD^+ = (SD$ or $R)$, is considered success. Freidlin and Simon (2005) found that, compared to a conventional randomized design, the RDD design has substantial power loss for comparing τ to placebo in terms of $\Pr(Y_2 = PD)$. The RDD is an elaboration of a simple single-arm phase IIA activity trial for τ based on stage 1 of therapy alone (cf. Gehan, 1961; Thall and Sung, 1998) that includes an additional second stage of therapy where treatment is chosen adaptively using Y_1. In this sense, the RDD is a randomized comparison of the two-stage dynamic treatment regimes $\rho_1 = (\tau^{(1)}, \tau^{(2)})$ and $\rho_2 = (\tau^{(1)}, DNT^{(2)})$, where $\tau^{(1)}$ means "Give τ in stage 1," $\tau^{(2)}$ means "Give τ in stage 2 if Y_1 was SD^+" and $DNT^{(2)}$ means "Do not treat or give placebo in stage 2 if Y_1 was SD^+." The RDD randomizes patients between ρ_1 and ρ_2. An elaboration might also specify salvage treatments for PD, and treatments for toxicity.

Some Bayesian extensions of the RDD are as follows. If the clinical payoff for comparing the two regimes is Y_2 then, denoting $\pi_{2,\tau} = \Pr(Y_2 = SD^+ \mid Y_1 = SD,\ \tau$ in stage 2) and $\pi_{2,DNT} = \Pr(Y_2 = SD^+ \mid Y_1 = SD,\ DNT$ in stage 2), a large value of $\Pr(\rho_1 > \rho_2 \mid data)$ $= \Pr(\pi_{2,\tau} > \pi_{2,DNT} \mid data\}$, say above .95 or .99, would lead to the conclusion that giving τ is better than not treating the patient in stage 2 if SD is seen with τ in stage 1. Values of $\Pr(\rho_1 > \rho_2 \mid data)$ near 0.50 correspond to no difference, and values near 0 to ρ_2 being superior. It may be useful to add a Bayesian futility rule that stops the trial early if

$$\Pr\{\pi_{1,\tau} > \pi_1^* \mid data_n\} < p_{1,L} \tag{12}$$

where $\pi_{1,\tau} = \Pr(Y_1 = SD^+ \mid \tau$ in stage1) and π_1^* is a fixed minimum stage 1 activity level in terms of SD^+, say 0.20.

To accommodate competing targeted agents, say τ_1 and τ_2, a generalization of the RDD

might randomize patients between τ_1 and τ_2 for stage 1. If τ_1 is given in stage 1, then the stage 2 adaptive rule might be to give τ_2 if $Y_1 = PD$; randomize between τ_1 and τ_2 if $Y_1 = SD$; and repeat τ_1 in stage 2 if $Y_1 = R$. The two regimes being compared are $\rho_1 = (\tau_1, \tau_2)$ and $\rho_2 = (\tau_2, \tau_1)$, where ρ_1 says to start with τ_1 in stage 1, repeat τ_1 in stage 2 if $Y_1 = SD^+$, and switch to τ_2 in stage 2 if $Y_1 = PD$. The regime ρ_2 is obtained by switching the roles of τ_1 and τ_2. Schematically, ρ_1 may be expressed as $(\tau_1,\ Y_1 = PD \rightarrow \tau_2,\ Y_1 = SD^+ \rightarrow \tau_1)$. Bayesian comparison of ρ_1 and ρ_2 may be done as for the 2 regimes in the RDD, above. For $J > 2$ agents, however, there would be $J(J-1)/2$ such two-stage regimes, so even for $J = 3$ there are 6 regimes. For $J \geq 3$, stage 1 futility rules of the form (12) become very important. For example, dropping τ_1 due to stage 1 futility would eliminate both (τ_1, τ_2) and (τ_1, τ_3), and thus allow more patients to be randomized to the remaining 4 regimes. This may be thought as "between patient enrichment" of multi-stage targeted regimes.

7. Multiple Agents and Multiple Targets

None of the above extensions of the RDD account for $\mathbf{Z}$, and elaborations that do so unavoidably are much more complicated. For example, subgroup-specific stage 1 futility rules might be used, based on $\pi_{1,\tau_j(Z)} = \Pr(Y_1 = SD^+ \mid \tau_j,\ \mathbf{Z})$ for each τ_j and biomarker subgroup $\mathbf{Z}$. More generally, when either $\mathbf{Z} = (Z_1, \cdots, Z_p)$ has $p > 1$ entries or $\mathcal{T} = \{\tau_1, \cdots, \tau_J\}$ has $J > 1$ targeted agents, practical issues of discovery, delivery optimization, and obtaining reliable comparative evaluations are much more difficult. Ignore known prognostic covariates $\mathbf{X}$ for simplicity. Even with $p = 2$ targets and $J = 2$ targeted agents, where putatively τ_j targets Z_j for each $j = 1, 2$ the NSCLC trial has four biomarker-defined subgroups $\{(0,0), (1,0), (0,1), (1,1)\}$ for (Z_1, Z_2), and four possible treatment combinations, $\{C, C+\tau_1, C+\tau_2, C+\tau_1+\tau_2\}$. It is tempting to simply randomize patients with $(Z_1, Z_2) =$ (1,1) between C and $C+\tau_1+\tau_2$, patients with $(Z_1, Z_2) =$ (1,0) between C and $C+\tau_1$, patients with $(Z_1, Z_2) =$ (0,1) between C and $C+\tau_2$, and use C to treat all patients with $(Z_1, Z_2) =$ (0,0), controlling the sample sizes in the four treatment combination in some fashion. This strategy is motivated by the assumption that each τ_j has potential anti-disease activity only in patients with $Z_j = 1$, which often is incorrect. A simpler strategy is to only include two

arms, C and $C+\tau_1+\tau_2$. While this may seem very appealing, it cannot discover whether, for example, an observed improvement of $C+\tau_1+\tau_2$ over C in mean PFS could be achieved with $C+\tau_1$, that is, τ_2 provides no additional clinical benefit. Moreover, the issues of toxicity and determining acceptable doses for combinations must be addressed. Even for $p=2$, optimizing dose pairs is well known to be an extremely complex and difficult problem in single-arm phase I trials, and very little work has been done for $p \geq 3$. Recall the example in the 2×2 case of huge false positive and false negative error rates if homogeneity of treatment effect across Z is assumed but in fact there are substantial τ-Z interactions.

Inevitably, some strategy for dimension reduction must be devised. Michiels et al. (2011) propose permutation tests for a confirmatory two-arm trial based on survival time under a Weibull distribution with multiple biomarkers, where treatment-biomarker interactions are of interest, controlling the overall type I error for multiple tests. Tests are obtained by and computing an overall biomarker score $\mathbf{wZ} = w_1 Z_1 + \cdots + w_J Z_J$ for each patient and permuting $\mathbf{Z}$ among the patients within each treatment group. This sort of approach works if $Z_1, \cdots, Z_J$ all go in the same direction, i.e. larger Z_j corresponds to the hypothesis that τ_j has greater anti-disease effect. With K targeted agents, $\tau_1, \cdots, \tau_K$, after applying weeding rules to drop unpromising τ_j-$\mathbf{Z}$ combinations, one may focus attention on the most promising combinations in terms of the posteriors of $\mu_{\tau_j,\mathbf{Z}}$ and select a small subset for further evaluation. For example, one may rank order these based on $E(\mu_{\tau_j,\mathbf{Z}} \mid data_n)$ or $\Pr(\mu_{\tau_j,\mathbf{Z}} > \mu^* \mid data_n)$ for fixed μ^* and select the largest m, where m is a small, feasible number to evaluate. A comparative rule might select τ_j for further evaluation in subgroup $\mathbf{Z}$ if

$$\Pr(\mu_{\tau_j,\mathbf{Z}} > \min_{r,\mathbf{Z}}\{\mu_{\tau_r,\mathbf{Z}}\} \mid data_n) > p_{U,Z}. \tag{13}$$

If ranking is the objective, then an advantage of the Bayesian approach is that the posteriors of the ranks themselves may be computed (cf. Laird and Louis, 1989). In terms of the means, for $j = 1, \cdots, K$, the rank of τ_j in subgroup $\mathbf{Z}$ is

$$R(\tau_j, \mathbf{Z}) \;=\; \sum_{l=1}^{K} I(\mu_{\tau_j,\mathbf{Z}} \geq \mu_{\tau_l,\mathbf{Z}})$$

One may base decisions, similar to those given above in terms of parameters, on the joint posterior of $R(\tau_1, \mathbf{Z}), \cdots, R(\tau_K, \mathbf{Z})$.

For targeted regimes $\tau_1, \cdots, \tau_K$ with J biomarker subgroups, assume a model with linear terms $\boldsymbol{\eta} = \{\eta_{\tau_r,j},\ r = 1, 2, j = 1, ..., J\}$, where each real-valued $\eta_{\tau_r,j} = \text{link}(\mu_{\tau_r,j})$ for mean outcome $\mu_{\tau_r,j}$ of treatment τ_r in subgroup j. If it is realistic to assume that these effects are exchangeable across subgroups within each treatment, one may assume the Level 1 priors $\eta_{\tau_r,1}, \cdots, \eta_{\tau_r,J} \sim$ iid $N(\tilde{\mu}_{\tau_r}, \tilde{\sigma}^2_{\tau_r})$ for each $r = 1, \cdots, K$. For treatment τ_r, the deviation of treatment effect from the overall mean due to subgroup j is $\Delta_{j(r)} = \mu_{\tau_r,j} - \tilde{\mu}_{\tau_r}$, so $\Delta_{1(r)}, \cdots, \Delta_{K(r)} \sim$ iid $N(0, \tilde{\sigma}^2_{\tau_r})$ for each r. This model is saturated, with KJ parameters $\boldsymbol{\eta}$ and $2K$ fixed hyperparameters, $\tilde{\boldsymbol{\theta}} = (\tilde{\mu}_{\tau_1}, \cdots, \tilde{\mu}_{\tau_K}, \tilde{\sigma}^2_{\tau_1}, \cdots, \tilde{\sigma}^2_{\tau_K})$. If one further assumes a hierarchical model with Level 2 priors (hyperpriors) $\tilde{\mu}_{\tau_1}, \cdots, \tilde{\mu}_{\tau_K}, \sim$ iid $N(a, b)$ and $\tilde{\sigma}^2_{\tau_1}, \cdots, \tilde{\sigma}^2_{\tau_K}, \sim$ uniform $[0, U_{\sigma^2}]$, then there are three fixed level 2 hyperparameters, $\boldsymbol{\phi} = (a, b, U_{\sigma^2})$, regardless of K and J. This model shrinks the estimated posterior mean treatment effects toward each other, and shrinks the subgroup effects toward each other within treatments.

A futility rule to stop accrual in subgroup j may take the form

$$\Pr\{\max_{r \neq r'} |\eta_{\tau_r,j} - \eta_{\tau_{r'},j}| < \delta_1 \mid data_n\} < p_L. \tag{14}$$

Identifying a substantive treatment-subgroup effect might be done relative to a historical value η^H based on $\Pr(\eta_{\tau_r,j} > \eta^H + \delta_2 \mid data) > p_U$. A similar rule using only the trial data would be $\Pr(\eta_{\tau_r,j} > \max\{\eta_{\tau_m,l} : (m,l) \neq (r,j)\} + \delta_2 \mid data)$. The overall effect of τ_r is $\bar{\eta_r} = \sum_j w_j \eta_{r,j}$ where w_j is the probability of subgroup j. A comparison of overall treatment effects between τ_r and $\tau_{r'}$ could be based on $\Pr(|\bar{\eta_r} - \bar{\eta_{r'}}| > \delta_2 \mid data) > p_U$. The fact that there are $K(K-1)/2$ such pairwise comparisons would create the usual multiplicity issues. With all of these rules, however, shrinkage of posteriors among biomarker subgroups or treatment arms may help to control the overall false positive rates.

A final point pertains to uncertainty about $\mathbf{Z}$, which can take at least two forms. The first pertains to whether a particular Z_j should have been included in a given gene or protein signature $\mathbf{Z}$, or was included erroneously. It is very undesirable to treat a patient with an agent targeting an element of $\mathbf{Z}$ that was included erroneously or, alternatively, to fail to use an agent targeting a protein that should have been included but was either not discovered or excluded erroneously. All of the methods discussed here could be elaborated by including

a vector $\mathbf{p}_Z$ where each entry $p(Z_j)$ is the probability that Z_j is correct, e.g. using a beta prior if Z_j is binary. Such indexes of uncertainty might be obtained from previous genomic discovery studies. The second source of uncertainty assumes that $\mathbf{Z}$ is qualitatively correct, but pertains to whether each entry of a particular patient's $\mathbf{Z}_i$ was measured with error, specifically whether each binary $Z_{i,j}$ was incorrectly scored as a false positive or false negative, or continuous $Z_{i,j}$ is actually $Z_{i,j}^{true} + \epsilon_{i,j}$ where $\epsilon_{i,j}$ is, say, Gaussian measurement error. Given that some $\mathbf{Z}$ is assumed to be qualitatively correct, each patient's $\mathbf{Z}_i$ could have an associated probability distribution $q(\mathbf{Z}_i)$to account for possible misclassification or measurement error, and here a Bayesian hierarchical model assuming that patients are exchangeable would be appropriate.

ACKNOWLEDGMENT

This research was partially supported by NIH grant RO1 CA 83932.

Bibliography

1. J.D. Andersen. Use of predictive probabilities in phase II and phase III clinical trials. *J Biopharmaceutical Statistics*. 9(1) 67-79, 1999.

2. J.S. Babb and A. Rogatko. Patient specific dosing in a phase I cancer trial. *Statistics in Medicine* 20: 2079-2090, 2001.

3. B.N. Bekele and P.F. Thall. Dose-finding based on multiple toxicities in a soft tissue sarcoma trial. *J American Statistical Assoc*, 99:26-35, 2004.

4. B. Freidlin and R. Simon. Evaluation of randomized discontinuation design. *J Clin Oncol*, 23:50948, 2005

5. B. Freidlin and R. Simon. Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Res* 11(21): 7872-7878, 2005.

6. N. Friedman, L. MIchal, I. Naxchman and D. Pe'er. Using Bayesian networks to analyze expression data. *J Computational Biology* 7:601-620, 2000.

7. S.W. Karuri and R. Simon. A two-stage Bayesian design for co-development of new drugs and companion diagnostics. *Statistics in Medicine.* 31:901-914, 2012.

8. J.A. Kopec, M. Abrahamowicz, and J.M. Esdaile. Randomized discontinuation trials: utility and efficiency. *J Clin Epidemiol* 46:95971, 1993.

9. N. Laird and T.A. Louis. Empirical Bayes ranking methods. *Journal of Educational Statistics* 14:29–46, 1989.

10. A. Maitournam and R. Simon. On the efficiency of targeted clinical trials, *Statistics in Medicine*, 24:329-339, 2005.

11. S. Michiels, R.E. Potthoff and S.L. George. Multiple testing of treatment-effect-modifying biomarkers in a randomized clinical trial with a survival endpoint. *Statistics in Medicine*, 30:1502-1518, 2011.

12. R. Mick and M.J. Ratain. Model-guided determination of maximum tolerated dose in phase I clinical trials: Evidence of increased precision. *Journal of the National Cancer Institute* 85: 217-223, 1993.

13. S. Morita, P.F. Thall, and P. Mueller. Determining the effective sample size of a parametric prior. *Biometrics* 64: 595-602, 2008.

14. S. Morita, P.F. Thall and P. Mueller. Evaluating the impact of prior assumptions in Bayesian biostatistics. *Statistics in Biosciences.* 2:1-17, 2010.

15. J. O'Quigley, M. Pepe, and L. Fisher. Continual reassessment method: A practical design for Phase I clinical trials in cancer. *Biometrics* 46: 33-48, 1990.

16. C.P. Robert and G. Cassella. *Monte Carlo Statistical Methods.* New York: Springer, 1999.

17. G.L. Rosner, W. Stadler, and M.J. Ratain. Randomized discontinuation design: application to cytostatic antineoplastic agents. *J Clin Oncol* 20:447884, 2002.

18. B. Saville, H. Ah, and G. Koch. A robust method for comparing two treatments in a confirmatory clinical trial via multivariate time-to-event methods that jointly incorporate information from longitudinal and time-to-event data. *Statistics in Medicine* 29:75-85, 2009.

19. M.R. Sharma, W.M. Stadler, M.J. Ratain. Randomized phase II trials: A Long-term investment with promising returns. *J National Cancer Institute* 103:10931100, 2011.

20. L.B. Sheiner. Learning versus confirming in clinical drug development. *Clin Pharmacol Ther*, 61:27591, 1997.

21. Y. Song and Y.H. Choi. A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine*, 26:3535-3549, 2007.

22. Y. Shen and P.F. Thall. Parametric likelihoods for multiple non-fatal competing risks and death. *Stat in Medicine*, 17:999-1016, 1998.

23. R. Simon and A. Maitournam. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res*, 10:675963, 2004

24. D.J. Spiegelalter, K.R. Abrams, and J.P. Myles. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* John Wiley and Sons, Chichester, 2004.

25. L. Tang and X-H. Zhao. A general framework for marker design with optimal allocation to assess clinical utility. *Statistics in Medicine* 32:620-630, 2013.

26. P.F. Thall, H.Q. Nguyen, and E.H. Estey. Patient-specific dose-finding based on bivariate outcomes and covariates. *Biometrics* 64: 1126-1136, 2008.

27. P.F. Thall and H-G. Sung. Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*, 17:1563-1580, 1998.

28. P.F. Thall, H.-G. Sung and E.H. Estey. Selecting therapeutic strategies based on efficacy and death in multi-course clinical trials. *J American Statistical Assoc*, 97:29-39, 2002.

29. P.F. Thall and J.K. Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, 24:1947-1964, 2005.

30. P.F. Thall and J.K. Wathen. Practical Bayesian adaptive randomization in clinical trials. *European J Cancer*. 43:860-867, 2007.

31. P.F. Thall, L.H. Wooten, C.J. Logothetis, R. Millikan and N.M. Tannir. Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statistics in Medicine*. 26:4687-4702, 2007.

32. A. Wahed and P.F. Thall. Evaluating joint effects of induction-salvage treatment regimes on overall survival in acute leukemia. *Journal of Royal Statistical Society, Series C*. In press, 62, 2013.

33. L. Wang, A. Rotnitzky, X. Lin, R. Millikan, and P.F. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *J American Statistical Assoc.* 107:493-508, (with discussion, pages 509-517; rejoinder, pages 518-520), 2012.

34. J.K. Wathen and P.F. Thall. Bayesian adaptive model selection for optimizing group sequential clinical trials. *Statistics in Medicine* 27:5586-5604, 2008.

35. Y. Yuan and G. Yin. Bayesian dose-finding by jointly modeling toxicity and efficacy as time-to-event outcomes. *Journal of the Royal Statistical Society: Series C* 58:719-736, 2009.

36. Y. Zhang. A novel Bayesian graphical model for genome-wide multi-SNP association mapping. Genet Epi, 36:36-37, 2011.

37. Y. Zhao, D. Zeng, M.A. Socinski, and M.R. Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics* 67: 14221433, 2011.