Biomarker-based Bayesian randomized phase II clinical trial design to identify a sensitive patient subpopulation

Satoshi Morita[1,*], Hideharu Yamamoto[2] and Yasuo Sugitani[2]

[1] *Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan*

[2] *Clinical Research Planning Department, Chugai Pharmaceutical Co., Ltd., Tokyo, Japan*

SUMMARY

The benefits and challenges of incorporating biomarkers into the development of anti-cancer agents have been increasingly discussed. In many cases, a sensitive subpopulation of patients is determined based on pre-clinical data and/or by retrospectively analyzing clinical trial data. Prospective exploration of sensitive subpopulations of patients may enable us to efficiently develop definitively effective treatments, resulting in accelerated drug development and a reduction in development costs. We consider the development of a new molecular-targeted treatment in cancer patients. Given preliminary but promising efficacy data observed in a phase I study, it may be worth designing a phase II clinical trial that aims to identify a sensitive subpopulation. In order to achieve this goal, we propose a Bayesian randomized phase II clinical trial design incorporating a biomarker that is measured on a graded scale. We compare two Bayesian methods, one based on subgroup analysis and the other on a regression model, to analyze a time-to-event endpoint such as progression-free survival (PFS) time. The two methods basically estimate Bayesian posterior probabilities of PFS hazard ratios in biomarker subgroups. Exten-

sive simulation studies evaluate these methods' operating characteristics, including the correct identification probabilities of the desired subpopulation under a wide range of clinical scenarios. We also examine the impact of subgroup population proportions on the methods' operating characteristics. Although both methods' performance depends on the distribution of treatment effect and the population proportions across patient subgroups, the regression-based method shows more favorable operating characteristics.

Key words: Biomarker; Molecular-targeted agent; Bayesian statistics; Randomized phase II trial; Time-to-event data.

Short title: Biomarker-based Bayesian randomized phase II trial

* Address for correspondence:
Satoshi Morita, PhD
Department of Biomedical Statistics and Bioinformatics,
Kyoto University Graduate School of Medicine
Address: 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan
E-mail: smorita@kuhp.kyoto-u.ac.jp
Phone: +81-75-751-4717, Fax: +81-75-751-4767

Conflict of interest statement: There are no conflicts of interest in this study.

1. INTRODUCTION

Recently, the benefits and challenges of incorporating biomarkers into the development of anti-cancer agents have been increasingly discussed [1]. Many clinical trials are conducted to develop new molecular-targeted anticancer agents that are likely to benefit only a subset of patients. If a clinical trial is performed in a broad population of patients, which includes insensitive as well as sensitive patients, any effect of a new agent on the sensitive subset of patients may be missed. Therefore, drug development should aim to optimize the target population of patients for treatment by appropriately focusing on patients who could obtain a sufficient benefit from a molecular-targeted agent. In addition, identifying the sensitive subset of patients may be a vital process in clinical development in terms of speeding up the drug development process and reducing development costs [2, 3, 4, 5].

The following two examples of clinical development represent two different extremes in the approach to this problem. First, trastuzumab, which is a humanized monoclonal antibody with high specificity for the human epidermal growth factor receptor 2 (HER2) protein, demonstrated high anti-tumor activity in patients with HER2-overexpressing metastatic breast cancer [6, 7, 8]. Based on preclinical and clinical data that strongly supported the existence of a sensitive subpopulation of patients, the clinical development of trastuzumab prospectively focused on studying the agent in HER2-overexpressing breast cancer patients. Secondly, during the development of monoclonal antibodies targeting epidermal growth factor receptor (EGFR), such as panitumumab, and EGFR tyrosine kinase inhibitors (TKIs), such as gefitinib, patients were enrolled in clinical trials without preselection based on EGFR status or other biomarkers [6, 7]. For example, Amado *et al.* [9] retrospectively analyzed whether the effect of panitumumab on progression-free survival (PFS) in patients with metastatic colorectal cancer differed by

KRAS status and showed a significant treatment effect in the wild-type KRAS subgroup. That is, in the first case, solid prior data enabled clinical investigators to prospectively design subsequent clinical trials to develop a molecular-targeted agent in a patient subpopulation identifiable with a biomarker assay. In the other case, retrospective subgroup analysis of a phase III trial conducted in unselected patients was able to successfully identify a sensitive patient subpopulation. In many cases, however, the reality may lie in between these two cases.

If a study population of patients contains non-sensitive subpopulations, a much larger sample size would be required to establish statistically significant results in a final confirmatory phase III trial [10]. When considering the entire course of a new agent's clinical development, therefore, conducting a properly designed phase II trial may be key to raising the "success probability" of a subsequent phase III trial. In particular, pharmacogenetically developed drugs often rely on assays to measure target expression levels (e.g., HER2 or EGFR) on a graded scale; these levels are then dichotomized to define two subsets of patients with positive or negative status. We call the subset of patients with positive status the sensitive subpopulation. In this paper, we consider identifying the sensitive subpopulation using a graded-scale biomarker in a randomized phase II clinical trial to develop a new molecular-targeted agent. In order to design the phase II trial, we adopt a Bayesian approach for the decision-making flexibility it affords during the exploratory phase of clinical development. We compare two Bayesian methods, one based on subgroup analysis and the other on a regression model, in terms of their performance in identifying a sensitive subpopulation. In addition, we consider interim analyses to prematurely terminate the trial due to futility.

As reviewed by Yin [11], there is a substantial literature on study designs that are used to identify sensitive patient subpopulations, including Jiang *et al.* [10], Wang *et*

al. [12], Brannath *et al.* [13], Eickhoff *et al.* [14], and Jenkins *et al.* [15] proposed adaptive two-stage designs in which the patient subset(s) specified in the first stage is used to evaluate the treatment effect in the second stage. Their proposed study designs presume that two mutually exclusive patient subgroups are determined in advance on the basis of preclinical research or a separate exploratory study. Our focus is simply on identifying a sensitive patient subpopulation in the phase II stage, although the above study designs consider phase II/III or phase III trial settings.

This paper is organized as follows. In Section 2, we provide a motivating example. Section 3 outlines the study design of a Bayesian randomized phase II clinical trial to identify a sensitive patient subpopulation. We conduct extensive simulation studies to examine the operating characteristics of our proposed study design in Section 4. We close with a brief discussion in Section 5.

2. A MOTIVATING EXAMPLE

In this section, we present a case study based on the actual clinical development of a new molecular-targeted monoclonal antibody. Pre-clinical and clinical works suggested that anti-tumor activity of the new antibody should depend significantly on the target protein amounts. In this study the intensity of the biomarker expression is defined using a graded scale (e.g., 0, 1+, 2+, 3+), with higher values indicating higher expression. Results from a phase I dose-finding clinical trial suggested a possible association between biomarker expression and the efficacy of the antibody, that is, longer PFS time tended to be observed in patients with higher expression (e.g., 2+ and 3+). In this study, we assume monotonicity in the efficacy of the new agent with respect to the biomarker grade.

While effective first-line therapies exist for patients with advanced stages of cancer and poor prognoses, in particular hepatocellular carcinoma (HCC) and pancreatic car-

cinoma, no standard second-line treatments have yet been established. In randomized phase II clinical trials to develop second-line oncology treatments, the experimental and control arms (arms E and C) should be "best supportive case (BSC) + new agent" and "BSC + placebo", and a time-to-event outcome such as PFS time is often used as the primary endpoint [16]. In some cases, a biomarker may not only be a predictive factor for a new agent but also a prognostic factor for patients with a specific cancer type. In this study, we assume that the biomarker predicts the efficacy of the new agent, but does not predict patient prognosis. That is, we consider the situation where the efficacy in the control (placebo) arm is not modified by the biomarker. However, it is not difficult to extend our proposed study design to cases where prognosis differs between subgroups.

Under these settings, we consider designing a randomized phase II trial to assess whether the addition of a new monoclonal antibody therapy to BSC sufficiently benefits the patients in terms of prolongation of PFS time. The biomarker grade is used as a stratification factor when randomization is carried out. In order to summarize the PFS data, we basically use a hazard ratio comparing arm E to arm C, which is denoted by λ. In this study, we consider evaluating the hazard ratios in G biomarker subgroups, which are denoted by λ_g, $g = 1, \ldots, G$. Our specific goal is to find the upper subset consisting of subgroups $g \geq \kappa_0$, which meets the definition of the sensitive subpopulation, by evaluating these hazard ratios. Then, a subsequent phase III trial is to be conducted in the identified subpopulation. The value of cutoff $\kappa_0 \in \{1, \ldots, G+1\}$ is unknown and will be determined based on data observed in the trial. As one of the two extreme cases, $\kappa_0 = 1$ suggests that arm E should be beneficial for the entire population of patients, and one can make a decision to proceed to a subsequent phase III trial that enrolls the entire population of patients. On the other hand, the cutoff $\kappa_0 = G + 1$ indicates that arm E will not be beneficial for any subgroup and the "no-go" decision to a subsequent

phase III trial should be taken.

3. BIOMARKER-BASED BAYESIAN RANDOMIZED PHASE II STUDY DESIGN

We use the two Bayesian methods that are both based on a common probability model for PFS time. One method is based on a subgroup analysis (S-A method), and the other on a regression model (R-M method).

3.1 Notation, probability model for PFS time, and Bayesian posterior computation

For patient i, let x_i denote the treatment indicator, with $x_i = 1$ if patient i receives the experimental arm and $x_i = 0$ if they receive the control arm. Let T_i denote PFS time for patient i. For subgroups 1 to G defined by the biomarker grade, $z_{i,g} = 1$ if patient i is in subgroup g and 0 if not. Thus, $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,G})$ is the subgroup indicator vector for patient i. Let $\phi_1, \ldots, \phi_G$ denote the proportions of patients in subgroups $1, \ldots, G$, who would be enrolled into the phase II trial. These proportions reflect the true biomarker subgroup prevalence in the entire population of patients. Although $\Phi = (\phi_1, \ldots, \phi_G)$ is actually unknown, in the simulation study we will handle the proportions Φ as fixed values and vary the values to examine the sensitivity of simulation results to the subgroup prevalence. That is, although the proportions Φ could be handled as additional parameters to be estimated in a Bayesian study design, we will not consider them in this study.

The two Bayesian methods explained in the next subsection commonly use the following proportional hazards model. Under the proportional hazards assumption in each subgroup, the hazard at time t for patient i with x_i can be modeled as

$$h(t \mid x_i, \mathbf{z}_i) = h_0(t)\exp(\sum_{g=1}^{G} \beta_g x_i z_{i,g}), \tag{1}$$

where $h_0(t)$ denotes the baseline hazard function and β_g denotes the regression coefficient

for x_i in subgroup g. According to Sinha *et al.* [17] and Ibrahim *et al.* [18], we use the partial likelihood of the Cox proportional hazards model as the likelihood to compute the posterior distributions of the parameters in the two Bayesian methods. We used Markov chain Monte Carlo (MCMC) to compute the posteriors [19], because the joint posterior distribution of regression coefficient parameters is not readily available in closed form.

As the criteria to identify the sensitive subpopulation, we basically use the following Bayesian posterior probability given the observed data $\mathcal{D}$ from the trial,

$$p(\lambda < \eta^* \mid \mathcal{D}) > \pi^* \tag{2}$$

where η^* is the upper limit and π^* is the upper probability cutoff. These design parameters, η^* and π^*, need to be calibrated on the basis of operating characteristics of the study design, which are examined in simulation studies. More specifically, let $\mathcal{D}_g$ denote the data observed in subgroup g and $\mathcal{D}_{all}$ denote the data observed in all G subgroups.

3.2 Two Bayesian methods to analyze PFS time

The objective of the phase II trial is to prove the concept of a targeted therapy, that is, to evaluate whether higher efficacy of the new antibody is observed in patients with higher biomarker expression. Therefore we assume the monotonicity in the efficacy of the new antibody in both methods but in different ways.

The S-A method separately evaluates the hazard ratio in each subgroup using the data observed in that subgroup. Assuming the monotonic increase in $p(\lambda_g < \eta^* \mid \mathcal{D}_g)$ for $g = 1, \ldots, G$, this method sequentially assesses whether $p(\lambda_g < \eta^* \mid \mathcal{D}_g) > \pi^*$ from subgroup 1 to G. That is, if $p(\lambda_1 < \eta^* \mid \mathcal{D}_1)$ is higher than π^*, we determine $\kappa_0 = 1$. If not, we proceed to subgroup 2. If $p(\lambda_2 < \eta^* \mid \mathcal{D}_2) > \pi^*$, we determine $\kappa_0 = 2$ and decide to identify subgroups 2 to G as the sensitive subpopulations. Similar computations and decision making are then repeated up to subgroup G. If all of the posterior probabilities,

$p(\lambda_1 < \eta^* \mid \mathcal{D}_1)$, ..., $p(\lambda_G < \eta^* \mid \mathcal{D}_G)$ are lower than π^*, we determine $\kappa_0 = G + 1$. We assume a non-informative normal prior N(0,1000) for each of the regression coefficient parameters, $\beta_1, \ldots, \beta_G$, to perform these posterior computations.

The R-M method assumes a monotonic decrease in hazard ratio for the biomarker subgroups with the parameter constraint $\beta_1 > \beta_2 > \ldots > \beta_G$. In addition, this method uses the data observed in all G subgroups, $\mathcal{D}_{all}$, to evaluate the posterior distribution of λ_g for $g = 1, \ldots, G$. For computational convenience, we reparameterize $(\beta_1, \ldots, \beta_G)$ with $(\beta_1, \gamma_1, \ldots, \gamma_{G-1})$ as $\beta_1 = \beta_1$, $\beta_2 = \beta_1 - \gamma_1$, ..., $\beta_G = \beta_{G-1} - \gamma_{G-1} = \beta_1 - \gamma_1 - \gamma_2 - \ldots - \gamma_{G-1}$, where $\gamma_1 > 0, \gamma_2 > 0, \ldots, \gamma_{G-1} > 0$. Assuming a non-informative normal prior N(0,1000) for β_1 and a non-informative gamma prior Ga(0.001,0.001) with mean 1 and variance 1000 for $\gamma_1, \ldots, \gamma_{G-1}$, we compute the marginal posterior distribution of the hazard ratios. Based on the computations, we find the cutoff κ_0 to satisfy the following equation,

$$\kappa_0 = \inf_{g \in (1,\ldots,G)} \{g \mid p(\lambda_g < \eta^* \mid \mathcal{D}_{all}) > \pi^*\}. \tag{3}$$

That is, the cutoff κ_0 is specified as the minimum of the integers $g \in \{1, \ldots, G\}$ that meet $p(\lambda_g < \eta^* \mid \mathcal{D}_{all}) > \pi^*$.

Although we suppose the S-A method to have more flexibility, it may perform more poorly at identifying a sensitive subpopulation due to its subgroup-analysis approach. In contrast, although we expect the R-M method to show a higher performance owing to the parameter constraint and the use of $\mathcal{D}_{all}$, this method may be vulnerable to departures from the monotonicity assumption. We will evaluate the advantages and disadvantages of the two methods in the simulation study.

3.3 Interim study monitoring rules

It may be important to terminate a clinical trial early from ethical and practical points of view. In the randomized phase II trial, we consider early termination of the entire trial

due to futility by planning interim analyses. Although it may also be useful to consider partly terminating insensitive patient subgroups or reducing the size of those subgroups, we did not take these measures in this study. This is because it may be generally desirable to obtain sufficient data on patients in the non-selected subpopulation in order to more precisely evaluate their response to and the safety of the new treatment [20].

The number and timing of interim analyses should be determined by taking into account the practicalities of patient enrollment rates and collecting and processing of study data. In the randomized phase II trial, we consider two interim analyses with the first and second analyses occurring after 60% and 80% of patients are recruited, respectively. When using the S-A method, given the lower probability cutoff π^*_{stop}, we consider the experimental arm to have disappointingly insufficient efficacy if $p(\lambda_g < \eta^* \mid \mathcal{D}_g) < \pi^*_{stop}$ for all g. Similarly, we stop the trial early if $p(\lambda_g < \eta^* \mid \mathcal{D}_{all}) < \pi^*_{stop}$ for all g when using the R-M method. The lower cutoff π^*_{stop} needs to be calibrated on the basis of the study design operating characteristics in the same way as the upper cutoff π^*. As another interim monitoring rule, it may be useful to include early stopping for efficacy by using an efficacy stopping criterion, such as $p(\lambda_g < \eta^* \mid \mathcal{D}) > \pi^*_{stop,Eff}$. Due to the same reasons mentioned above, however, we will not apply this rule to the phase II trial.

4. EVALUATION OF OPERATING CHARACTERISTICS

4.1 Parameter calibration and simulation plan

To evaluate and compare the two Bayesian methods in the case study with four subgroups, we simulated the trial 5,000 times using extensively varying situations. We used MCMC methods to obtain samples from the posterior distributions of the parameters. In order to complete the study design, we needed to calibrate the design parameters

$(\eta^*, \pi^*, \pi^*_{stop}, N)$ on the basis of the desired type I error rate under a null hypothesis and power under an alternative hypothesis in the trial with the projected total sample size N. The detailed definitions of type I error and power are given below.

We first performed a series of simulation studies with all 12 combinations of the three fixed upper limits ($\eta^* = .70, .80, .85$), the two upper probability cutoffs ($\pi^* = .70, .80$), and the two lower probability cutoffs ($\pi^*_{stop} = .10, .20$) under $N = 500$. Although the total sample size of 500 may be too large for a phase II trial, we used N= 500 to reliably evaluate the performances of the two methods in the simulation study. The simulation results are summarized in Supplemental Tables *(see the supplementary on-line materials)*. After determining the best combination of η^*, π^*, and π^*_{stop}, we evaluated the operating characteristics using six sample size values ($N = 250, 300, 350, 400, 450, 500$) to determine the appropriate sample size for the randomized phase II trial. Furthermore, we assumed the five patterns of subpopulation proportions $\Phi = (\phi_1, \phi_2, \phi_3, \phi_4)$, as shown in Table 1, to evaluate the sensitivity of simulation results to the subgroup prevalence. We predicted that patterns 1 and 3 were more likely to be observed in the phase II trial according to the historical data.

We assumed the five clinical scenarios for the simulation study based on hazard ratios as shown in Table 1. Each scenario is characterized by the true (fixed) hazard ratios (HR_1, HR_2, HR_3, HR_4) for the four subgroups. Scenario (1) is a null case, with all hazard ratios equal to 1.0. The sensitive subpopulation, found under each scenario, is indicated in boldface. In order to define the sensitive subpopulation, we first specify the efficacy threshold so that subgroup g is contained in the sensitive subpopulation if $HR_g \leq$ the threshold. One possible way to specify the efficacy threshold may be to hold discussions with physicians regarding the published results of clinical trials, because such a specification needs to take into account the current medical environment, such

as state of the art therapy and medical costs. For example, in advanced HCC, Llovet *et al.* [21] explored the ability of several biomarkers to predict the efficacy of a new small molecule, sorafenib, using the data from the phase III sorafenib HCC assessment randomized protocol (SHARP) trial [22]. Based on this report as well as other previous data, we solicited the opinions of the two hepatologists in the study group regarding the efficacy threshold. They suggested that the efficacy threshold = 0.6 should be clinically acceptable. We will use a power value to designate the probability of correctly identifying the target subgroup(s) as the sensitive subpopulation under alternative scenarios, and a type I error to designate the probability of identifying any subgroup(s) under the null scenario.

Taking historical data on second-line therapies for HCC into account, for the simulations, we assumed that the median PFS time was 2.8 months for all four subgroups in the control arm of the trial, with 12.0 months of patient recruitment and 15.0 months of maximum follow-up (i.e., 3.0 months of minimum follow-up). In addition, we assumed that patients arrived uniformly during the recruitment period. Assuming that the patient PFS times are i.i.d. Exp(ν), exponential with parameter ν, which has pdf $f(t \mid \nu) = \nu \exp(-\nu t)$, we generated PFS times using the fixed parameter $\nu_c = 0.33$ for the control arm. For the experimental arm, we used the parameter $\nu_c HR_g$ to generate PFS times in subgroups g for $g = 1, \ldots, 4$. The SAS programs to carry out simulations using the S-A and R-M methods are provided in the Supplementary Materials (SAS for Windows release 9.3; SAS Institute Inc., Cary, NC, USA).

4.2 Simulation results

In presenting the results of the simulation studies comparing the S-A and R-M methods, we summarize the probabilities of identifying i) none of the four subgroups, ii) subgroup 4 only, iii) subgroups 3 and 4, iv) subgroups 2 to 4, and v) all four subgroups, as being in

the sensitive subpopulation; these categories are denoted by $\mathcal{P}_{none}$, $\mathcal{P}_4$, $\mathcal{P}_{3-4}$, $\mathcal{P}_{2-4}$, and $\mathcal{P}_{all}$, respectively. We chose the combination of $\eta^* = .80$, $\pi^* = .70$, and $\pi^*_{stop} = 0.2$, which were judged to provide the best operating characteristics for the two methods, based on the extensive simulations (as shown in Supplementary Tables *in the supplementary on-line materials*). Table 2 shows the simulation results with $N = 500$ under the five clinical scenarios with the five patterns of patient subpopulation proportions.

Under scenario (1) (null), the R-M method yielded extremely high probabilities of identifying none of the four groups ($\mathcal{P}_{none} = 0.98$ - 1.00), while the values of $\mathcal{P}_{none}$ with the S-A method were 0.70 to 0.80. That is, the R-M method sufficiently controlled type I error, holding it to less than 0.05 regardless of the pattern of subpopulation proportions under $N = 500$, while the S-A method did not. In addition, the R-M method resulted in early trial termination due to considerably high probabilities of identifying none of the four groups, especially at the first interim analysis. The likelihood of early termination differed significantly between the R-M and S-A methods. This may be because the R-A method analyzed the data observed in all four subgroups resulting in much sharper posterior distributions of λ_g than those obtained by the S-A method that used the data observed in each subgroup.

Under scenario (2) (linear), neither of the two methods worked sufficiently well, that is, $\mathcal{P}_{3-4}$ were at most 0.50 for both methods. In cases where an obvious sensitive subpopulation may not seem to exist, such as in this scenario that assumes that the hazard ratios change steadily over subgroups, it may be hard to definitively identify the target subpopulation using either of the methods. Under scenario (3) (step-down), although both the S-A and R-M methods performed well overall, the performance of the R-M method may depend significantly on subpopulation proportions. In pattern 4 in particular, where the number of patients enrolled in subgroup 1 (non-sensitive subpopulation)

was very slight, the R-M method was more likely to select all the subgroups resulting in poorer performance. Under scenario (4) (very high efficacy in subgroups 3 and 4), the R-M method selected subgroups 3 and 4 at sufficiently high probabilities across all patterns of subpopulation proportions and these probabilities were higher than or almost equal to those obtained by the S-A method. Under scenario (5) (very high efficacy only in subgroup 4), the two methods were almost comparable in terms of the probability of identifying subgroup 4 under pattern 1. In cases where the subpopulation proportion of subgroup 4 (sensitive subpopulation) was relatively high, such as in patterns 2 and 4, the R-M method performed much better than the S-A method, as expected. However, under patterns 3 and 5, in which the subpopulation proportion of subgroup 4 was small, the performance of the R-M method was lower than that of the S-A method.

Figure 1 indicates the type I error rates (lower circles) and power values (upper circles) provided by the R-M method for the six sample sizes (N = 250, 300, 350, 400, 450, 500) under the five patterns of subpopulation proportions. In this simulation study, we focused only on the R-M method because the S-A method could not sufficiently control the type I error rate even under $N = 500$. The R-M method held the type I error to less than 0.05 even under $N = 250$. In terms of providing 80% of the power, $N = 300$ may be sufficient for the projected total sample size of the phase II trial, considering that we actually expect the subpopulation proportions to be like pattern 1 or 3.

5. DISCUSSION

We have proposed a Bayesian approach with two alternative methods to identify a sensitive subpopulation in the setting of a randomized phase II clinical trial. Taking the simulation results into account, the R-M method may be recommended as the primary choice. The limitations of our proposed approach include: (a) the requirement of a large sample size for a phase II trial, (b) the inadequate study monitoring, (c) the monotonicity

assumption for hazard ratios of PFS for biomarker subgroups, (d) the requirement that a specific quantitative biomarker for sensitivity be established in advance, and (e) lack of experience using our proposed method in an actual clinical trial.

Considering the feasibility of patient enrollment, the projected sample size $N = 300$ may be the upper limit in a clinical trial of second-line therapies for HCC. $N = 300$ may be achievable by enrolling, for instance, 25 patients per month for one year in a multi-national trial setting. In some cases, however, it may be unrealistic to enroll such a large number of patients into a phase II trial due to the associated development costs. If we can successfully identify a sensitive subpopulation, however, the required sample size might be minimized in a subsequent phase III trial of an enriched patient population, thereby optimizing the total sample size for the entire clinical development of a new agent. In the phase II trial design, we considered early termination of the entire trial only. Because the trial is still in phase II, it may be highly recommended to monitor the safety of the new treatment. For example, a safety criterion to monitor the probability of toxicity in each subgroup, such as $p(prob(\text{Tox})_g > \eta^*_{Tox} \mid \mathcal{D}) > \pi^*_{stop,Tox}$, where η^*_{Tox} represents an acceptable toxicity level, may be useful. In addition, the efficacy and futility rules for stopping subgroups that we mentioned in Section 3.3 may help reduce the expected sample size of the phase II trial. This should be evaluated in future works. Our study design was based completely on a monotonic change in treatment efficacy for biomarker subgroups. However, such a monotonicity assumption does not necessarily work in all cases. If data observed in the phase II trial indicates a non-monotonic change, such as "V-shape", the S-A method modified to select the subgroup with the highest value of $p(\lambda_g < \eta^* \mid \mathcal{D}_g)$ may work better than the R-M method. Otherwise, we may need to develop an alternative method based on an isotonic regression model with the pool-adjacent-violator algorithm (PAVA) [23].

In this paper, we focused on identifying a sensitive subpopulation of patients in a randomized phase II trial to develop a new molecular-targeted anticancer agent. It may be useful to incorporate our proposed approach into a seamless phase II/III study design in order to maximize the probability of its successful development, an issue that will be examined in future works.

Acknowledgments

We thank Dr. Richard Simon for his helpful comments and useful suggestions. Satoshi Morita's work was supported in part by a Grant-in-Aid for Scientific Research C-24500345 from the Ministry of Health, Labour and Welfare of Japan and by a non-profit organization Epidemiological and Clinical Research Information Network. We thank the associate editor and the referees for their thoughtful and constructive comments and suggestions.

References

[1] Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res* 2004; **10**:6759-6763.

[2] Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, Ratain MJ, Le Blanc M, Stewart D, Crowley J, Groshen S, Humphrey JS, West P, Berry D. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin Cancer Res* 2010; **16**:1764-1769.

[3] McShane LM, Hunsberger S, Adjei AA. Effective incorporation of biomarkers into phase II trials. *Clin Cancer Res* 2009; **15**:1898-1905.

[4] Dancey JE, Dobbin KK, Groshen S, Jessup JM, Hruszkewycz AH, Koehler M, Parchment R, Ratain MJ, Shankar LK, Stadler WM, True LD, Gravell A, Grever MR; Biomarkers Task Force of the NCI Investigational Drug Steering Committee. Guidelines for the development and incorporation of biomarker studies in early clinical trials of novel agents. *Clin Cancer Res* 2010; **16**:1745-1755.

[5] Parmar MK, Barthel FM, Sydes M, Langley R, Kaplan R, Eisenhauer E, Brady M, James N, Bookman MA, Swart AM, Qian W, Royston P. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst* 2008; **100**:1204-1214.

[6] Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *J Biopharm Stat* 2009; **19**:530-542.

[7] Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn* 2011; **11**:171-182.

[8] Baselga J. Herceptin alone or in combination with chemotherapy in the treatment of HER2-positive metastatic breast cancer: pivotal trials. *Oncology* 2001; **61** (Suppl 2):14-21.

[9] Amado RG, Wolf M, Peeters M, Van Cutsem E, Siena S, Freeman DJ, Juan T, Sikorski R, Suggs S, Radinsky R, Patterson SD, Chang DD. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol* 2008; **26**:1626-1634.

[10] Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst* 2007; **99**:1036-1043.

[11] Yin G. *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*. Wiley: Hoboken, 2012.

[12] Wang SJ, O'Neill RT, Hung HM. Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharm Stat* 2007; **6**:227-244.

[13] Brannath W, Zuber E, Branson M, Bretz F, Gallo P, Posch M, Racine-Poon A. Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med* 2009; **28**:1445-1463.

[14] Eickhoff JC, Kim K, Beach J, Kolesar JM, Gee JR. A Bayesian adaptive design with biomarkers for targeted therapies. *Clin Trials* 2010; **7**:546-556.

[15] Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat* 2011; **10**:347-356.

[16] Korn EL, Arbuck SG, Pluda JM, Simon R, Kaplan RS, Christian MC. Clinical trial designs for cytostatic agents: are new approaches needed? *J Clin Oncol* 2001; **19**:265-272.

[17] Sinha D, Ibrahim JG, Chen MH. A Bayesian justification of Cox's partial likelihood. *Biometrika* 2003; **90**: 629-641.

[18] Ibrahim JG, Chen MH, Sinha D. Bayesian Survival Analysis. In: *Encyclopedia of Biostatistics*, John Wiley and Sons: Chichester, 2005.

[19] Gilks W, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London, 1996.

[20] US Food and Drug Administration (USFDA). *Guidance on Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products.* US FDA: Rockville, MD, 2012.

[21] Llovet JM, Pena CE, Lathia CD, Shan M, Meinhardt G, Bruix J; SHARP Investigators Study Group. Plasma biomarkers as predictors of outcome in patients with advanced hepatocellular carcinoma. *Clin Cancer Res* 2012; **18**:2290-2300.

[22] Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc JF, de Oliveira AC, Santoro A, Raoul JL, Forner A, Schwartz M, Porta C, Zeuzem S, Bolondi L, Greten TF, Galle PR, Seitz JF, Borbath I, Haussinger D, Giannaris T, Shan M, Moscovici M, Voliotis D, Bruix J; SHARP Investigators Study Group. Sorafenib in advanced hepatocellular carcinoma. *N Engl J Med* 2008; **359**:378-390.

[23] Yuan Y, Yin G. Dose-response curve estimation: a semiparametric mixture approach. *Biometrics* 2011; **67**: 1543-1554.

Table I. Patient subgroup population proportions $\Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$ and clinical scenarios characterized by the true (fixed) hazard ratios $\{HR_1, HR_2, HR_3, HR_4\}$ for subgroups 1 to 4 for the simulation study. The hazard ratio values in the sensitive subpopulation under each scenario are indicated in bold-face.

		Subgroup			
		1	2	3	4
Subpopulation proportion patterns		ϕ_1	ϕ_2	ϕ_3	ϕ_4
1	Equal	0.25	0.25	0.25	0.25
2	Higher in subgroups 1 & 4	0.35	0.15	0.15	0.35
3	Higher in subgroups 2 & 3	0.15	0.35	0.35	0.15
4	Increasing	0.05	0.15	0.30	0.50
5	Decreasing	0.50	0.30	0.15	0.05
Clinical scenarios		HR_1	HR_2	HR_3	HR_4
(1)	Null case	1.0	1.0	1.0	1.0
(2)	Linear	1.0	0.8	**0.6**	**0.4**
(3)	Step-down	1.0	**0.6**	**0.6**	**0.35**
(4)	High efficacy in subgroups 3 & 4	1.0	1.0	**0.5**	**0.3**
(5)	High efficacy only in subgroup 4	1.0	1.0	1.0	**0.5**

Table II. Probabilities of a sensitive subpopulation finding with the fixed upper limit $\eta^* = .80$ and the upper and lower probability cutoffs $\pi^* = .70$ and $\pi^*_{stop} = .20$ when the total sample size $N = 500$. The probabilities of identifying i) none of the four subgroups, ii) subgroup 4 only, iii) subgroups 3 and 4, iv) subgroups 2 to 4, v) all the four subgroups, are shown in $\mathcal{P}_{none}$, $\mathcal{P}_4$, $\mathcal{P}_{3-4}$, $\mathcal{P}_{2-4}$, $\mathcal{P}_{all}$, respectively. The probabilities of early stopping at the first and second interim analyses, which are included in $\mathcal{P}_{none}$, are also separately shown. The probability values of correct identification are indicated in bold face.

			Early stopping						
Scenario	Pattern	Method	1^{st}	2^{nd}	$\mathcal{P}_{none}$	$\mathcal{P}_4$	$\mathcal{P}_{3-4}$	$\mathcal{P}_{2-4}$	$\mathcal{P}_{all}$
(1)	1	S-A	0.04	0.05	**0.80**	0.04	0.05	0.05	0.05
		R-M	0.62	0.18	**0.99**	0.00	0.00	0.00	0.00
	2	S-A	0.04	0.04	**0.78**	0.03	0.08	0.09	0.03
		R-M	0.64	0.17	**0.99**	0.00	0.00	0.00	0.00
	3	S-A	0.04	0.04	**0.77**	0.08	0.03	0.03	0.09
		R-M	0.60	0.19	**0.99**	0.00	0.00	0.00	0.00
	4	S-A	0.03	0.03	**0.72**	0.01	0.04	0.07	0.17
		R-M	0.66	0.16	**1.00**	0.00	0.00	0.00	0.00
	5	S-A	0.03	0.03	**0.70**	0.15	0.09	0.04	0.02
		R-M	0.54	0.21	**0.98**	0.01	0.00	0.00	0.00

Table II (continued). Simulation results under scenarios (2) and (3).

			Early stopping						
Scenario	Pattern	Method	1^{st}	2^{nd}	$\mathcal{P}_{none}$	$\mathcal{P}_4$	$\mathcal{P}_{3-4}$	$\mathcal{P}_{2-4}$	$\mathcal{P}_{all}$
(2)	1	S-A	0.00	0.00	0.00	0.13	**0.54**	0.28	0.05
		R-M	0.00	0.00	0.01	0.15	**0.51**	0.23	0.10
	2	S-A	0.00	0.00	0.00	0.19	**0.48**	0.30	0.03
		R-M	0.00	0.00	0.00	0.24	**0.49**	0.23	0.04
	3	S-A	0.00	0.00	0.00	0.08	**0.55**	0.28	0.09
		R-M	0.01	0.00	0.02	0.09	**0.49**	0.22	0.18
	4	S-A	0.00	0.00	0.00	0.09	**0.50**	0.25	0.17
		R-M	0.00	0.00	0.00	0.11	**0.38**	0.16	0.34
	5	S-A	0.00	0.00	0.04	0.15	**0.49**	0.30	0.02
		R-M	0.06	0.02	0.17	0.16	**0.41**	0.23	0.03
(3)	1	S-A	0.00	0.00	0.00	0.04	0.15	**0.76**	0.05
		R-M	0.00	0.00	0.00	0.06	0.09	**0.71**	0.15
	2	S-A	0.00	0.00	0.00	0.09	0.21	**0.68**	0.03
		R-M	0.00	0.00	0.00	0.13	0.16	**0.64**	0.07
	3	S-A	0.00	0.00	0.00	0.01	0.11	**0.79**	0.09
		R-M	0.00	0.00	0.00	0.03	0.05	**0.62**	0.31
	4	S-A	0.00	0.00	0.00	0.04	0.20	**0.59**	0.17
		R-M	0.00	0.00	0.00	0.05	0.07	**0.33**	0.55
	5	S-A	0.00	0.00	0.00	0.04	0.11	**0.83**	0.02
		R-M	0.02	0.00	0.04	0.05	0.07	**0.80**	0.04

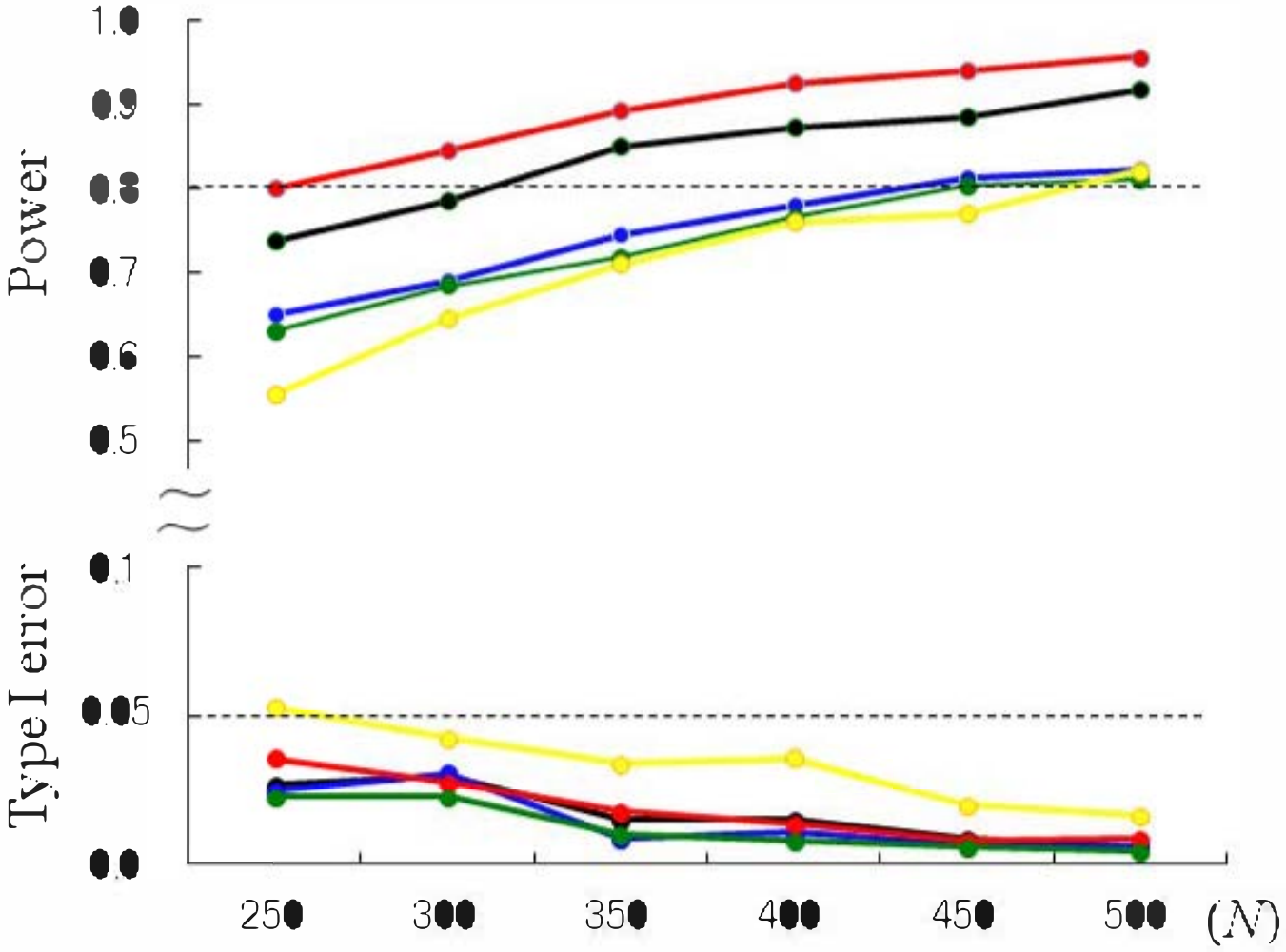
Table II (continued). Simulation results under scenarios (4) and (5).

			Early stopping						
Scenario	Pattern	Method	1^{st}	2^{nd}	$\mathcal{P}_{none}$	$\mathcal{P}_4$	$\mathcal{P}_{3-4}$	$\mathcal{P}_{2-4}$	$\mathcal{P}_{all}$
(4)	1	S-A	0.00	0.00	0.00	0.04	**0.86**	0.06	0.05
		R-M	0.00	0.00	0.00	0.04	**0.92**	0.02	0.02
	2	S-A	0.00	0.00	0.00	0.11	**0.78**	0.09	0.03
		R-M	0.00	0.00	0.00	0.13	**0.82**	0.03	0.02
	3	S-A	0.00	0.00	0.00	0.01	**0.87**	0.03	0.09
		R-M	0.00	0.00	0.00	0.01	**0.96**	0.01	0.02
	4	S-A	0.00	0.00	0.00	0.02	**0.74**	0.07	0.17
		R-M	0.00	0.00	0.00	0.03	**0.81**	0.05	0.12
	5	S-A	0.00	0.00	0.01	0.10	**0.83**	0.04	0.02
		R-M	0.04	0.01	0.07	0.09	**0.82**	0.02	0.01
(5)	1	S-A	0.00	0.00	0.04	**0.80**	0.05	0.05	0.05
		R-M	0.07	0.02	0.14	**0.82**	0.03	0.00	0.01
	2	S-A	0.00	0.00	0.01	**0.79**	0.08	0.09	0.03
		R-M	0.03	0.01	0.05	**0.87**	0.06	0.01	0.01
	3	S-A	0.00	0.00	0.10	**0.75**	0.03	0.03	0.09
		R-M	0.16	0.05	0.32	**0.66**	0.01	0.00	0.01
	4	S-A	0.00	0.00	0.00	**0.72**	0.04	0.07	0.17
		R-M	0.01	0.00	0.02	**0.94**	0.02	0.00	0.02
	5	S-A	0.01	0.01	0.29	**0.57**	0.09	0.04	0.01
		R-M	0.29	0.14	0.67	**0.31**	0.01	0.00	0.00

Figure Legends

Figure 1. Type I error rates (lower circles) and power values (upper circles) provided by R-M method for the six sample sizes (N = 250, 300, 350, 400, 450, 500) under the five patterns of subpopulation proportions; patterns 1: black, 2: blue, 3: red, 4: green, 5: yellow. In this investigation, the power is evaluated by the probability of correctly identifying subgroups 3 and 4 under scenario (4). The fixed design parameters $\eta^* = .80$, $\pi^* = .70$, and $\pi^*_{stop} = .20$ are used.

[Figure 1]


Power
1.0
0.9
0.8
0.7
0.6
0.5
Type I error
0.1
0.05
0.0
250
300
350
400
450
500
(N)