Therapeutic Innovation & Regulatory Science
http://dij.sagepub.com/

Incorporating Historical Data in Bayesian Phase I Trial Design: The Caucasian-to-Asian Toxicity Tolerability Problem

Kentaro Takeda and Satoshi Morita

Therapeutic Innovation & Regulatory Science published online 18 August 2014
DOI: 10.1177/2168479014546333

The online version of this article can be found at:
http://dij.sagepub.com/content/early/2014/08/18/2168479014546333

Published by:
SAGE
http://www.sagepublications.com

On behalf of:

Drug Information Association

Original Article

www.diahome.org

Therapeutic Innovation
& Regulatory Science
1-7

DOI: 10.1177/2168479014546333
tirs.sagepub.com

Incorporating Historical Data in Bayesian Phase I Trial Design: The Caucasian-to-Asian Toxicity Tolerability Problem

Kentaro Takeda, MS[1,2], and Satoshi Morita, PhD[1,3]

Abstract

Following phase I dose-finding oncology trials completed in Western countries, Asian investigators often conduct further phase I trials to determine the maximum tolerated dose for Asian patients. This may be due to concerns about possible differences in treatment tolerability between Caucasian and Asian patient groups. Our proposed approach aims to appropriately borrow strength from a previous Caucasian trial to improve the maximum tolerated dose determination in an Asian population of patients. We design an Asian phase I trial using the Bayesian continual reassessment method. First we analyze toxicity data from a Caucasian trial to derive the prior distributions for a subsequent Asian trial. Then, we calibrate the informativeness of the prior distributions according to prior effective sample size defined by Morita et al. Extensive simulation studies demonstrate favourable operating characteristics of the proposed method, compared with two methods based on power and noninformative priors, respectively.

Keywords

dose finding, phase I study design, historical data, prior effective sample size, continual reassessment method

Introduction

In this paper, we propose an approach to incorporating historical data to establish prior distributions for a dose-finding clinical trial to develop an anticancer agent. Following phase I dose-finding trials completed in Western countries, Asian investigators often conduct further phase I trials to determine the maximum tolerated dose (MTD) for Asian patients. This may be due to concerns about possible differences in treatment tolerability between Caucasian and Asian patient groups. In several cases, different treatment MTDs were estimated for Asians and Caucasians.[1,2] For example, phase I studies of capecitabine (Xeloda) monotherapy undertaken in Caucasians identified 1657 mg/m^2 as the MTD.[3,4] After these studies were completed, a phase I trial in Japanese patients determined a higher dose level, 2510 mg/m^2, as the MTD for Japanese patients.[5] Taking into account the recent trend of the globalization of new drug development, it may be worth considering the relevant use of historical data from a previous trial to design and conduct a subsequent trial in a new region. It should, however, be noted that an overly use of prior information may rather degrade the study design of a subsequent trial.

Our proposed approach aims to appropriately borrow strength from a previous Caucasian trial to improve the MTD determination in an Asian population of patients. We design an Asian phase I dose-finding trial using the Bayesian continual reassessment method,[6,7] even if other Bayesian designs can be used. The continual reassessment method is a model-based method that enables us to utilize all available prior information

Supplementary material for this article is available on the journal's website at http://tirs.sagepub.com/supplemental.

[1] Department of Biostatistics and Epidemiology, Graduate School of Medicine, Yokohama City University, Yokohama, Japan
[2] Biostatistics Group, Data Science, Global Development, Astellas Pharma Inc, Tokyo, Japan
[3] Department of Biomedical Statistics and Bioinformatics, Graduate School of Medicine, Kyoto University, Kyoto, Japan

Submitted 05-Apr-2014; accepted 15-Jul-2014

Corresponding Author:
Kentaro Takeda, Biostatistics Group, Data Science, Global Development, Astellas Pharma Inc, 2-5-1, Nihonbashi-Honcho, Chuo-ku, Tokyo 103-8411, Japan.
Email: kentaro.takeda@astellas.com

through prior distributions of the model parameters. First, we analyze toxicity data from a Caucasian trial to derive the priors for a subsequent Asian trial. We suppose that the Caucasian trial is conducted using a traditional "3 + 3" cohort design[8] and that the same dose levels are tested commonly in Caucasian and Asian trials. Second, we calibrate the informativeness of the priors according to a prior effective sample size (ESS).[9,10] Morita et al[10] wrote that the prior ESS provides a useful tool for understanding the impact of prior assumptions in Bayesian study design and data analysis. We call these priors based on the prior ESS "ESS priors." Finally, we conduct the Asian phase I trial using the continual reassessment method with the ESS priors.

In our study, we compare our proposed method with two methods based on power and noninformative priors, respectively, in terms of their performance in estimating MTD in a subsequent Asian dose-finding study. The power prior was proposed by Ibrahim and Chen[11] to allow investigators to incorporate and downweight historical data. The power prior raises the likelihood of historical data to a power parameter, $a_0 \in [0, 1]$, that controls how much strength to borrow from the historical data: "no borrowing ($a_0 = 0$)" to "full borrowing ($a_0 = 1$)."

This paper is organized as follows. In the next section, we outline the Bayesian study designs of an Asian phase I trial incorporating historical data from a previously conducted Caucasian phase I trial. We conduct extensive simulation studies to examine the operating characteristics of our proposed method in the subsequent section. We close with a brief discussion.

Probability Model and Study Designs

We compare the methods embedded with the 3 types of priors: the ESS, power, and noninformative priors. Note that the difference among the 3 methods is only in establishing the priors that are to be assumed in the Asian trial.

Preliminary Notation and Probability Model for Toxicity

Let $\mathcal{D}_C$ and $\mathcal{D}_A$ denote data from the Caucasian and Asian trials, respectively. That is, $\mathcal{D}_C$ and $\mathcal{D}_A$ correspond to the historical data and the current study data, respectively. Suppose that both Caucasian and Asian phase I trials are conducted to investigate a single anticancer agent with the same dose levels. Each patient receives one of J doses denoted by $d_1, \ldots, d_J$, with standardized doses $x_j = d_j/s.d.(d_1, \cdots, d_J)$, where, *s.d.* abbreviates standard deviation. As described in the introduction, we suppose that, for simplicity, the same dose levels are tested commonly in Caucasian and Asian trials. However, it is not difficult to extend our proposed method to cases where different dose levels are examined between two populations of patients.

We use a two-parameter logistic model to derive the priors based on the previous Caucasian data, as well as to design and conduct a subsequent Asian phase I trial. The outcome variable is the indicator $Y_1 = 1$ if a patient i suffers toxicity, 0 if not. Denoting the probability of toxicity under dose x_i by $\pi(x_i, \alpha, \beta)$, we assume the following two-parameter logistic model,

$$\pi(x_i, \alpha, \beta) = \Pr(Y_i = 1 | x_i, \alpha, \beta) = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)} \quad (1)$$

with the intercept and slope parameters α and β. We assume a normal prior for α as

$$\alpha \sim N(\mu_\alpha, \gamma_\alpha) \quad (2)$$

To constrain β to be positive, we assume a gamma prior for β as

$$\beta \sim Ga(g_1(\mu_\beta, \gamma_\beta), g_2(\mu_\beta, \gamma_\beta)), \quad (3)$$

where μ_β and γ_β are the prior mean and variance of β, respectively, and $g_1(s, t) = s^2/t$ and $g_2(s, t) = s/t$. We assume that α and β are a priori independent. We use Markov chain Monte Carlo to compute the posteriors,[12] because the joint posterior distribution of regression coefficient parameters is not readily available in closed form.

Establishing ESS Prior

By analyzing the historical data $\mathcal{D}_C$ using the two-parameter logistic model (1), we compute the posterior means and variances of α and β that are denoted by $(\tilde{\mu}_{\alpha,C}, \tilde{\mu}_{\beta,C})$ and $(\tilde{\gamma}_{\alpha,C}, \tilde{\gamma}_{\beta,C})$, respectively. For the priors of the model parameters in the Asian phase I trial, we propose to assume

$$\begin{aligned} \alpha &\sim N(\tilde{\mu}_{\alpha,C}, w \cdot \tilde{\gamma}_{\alpha,C}), \\ \beta &\sim Ga(g_1(\tilde{\mu}_{\beta,C}, w \cdot \tilde{\gamma}_{\beta,C}), g_2(\tilde{\mu}_{\beta,C}, w \cdot \tilde{\gamma}_{\beta,C})), \end{aligned} \quad (4)$$

where w is a constant for the prior calibration. Then we calibrate the prior distributions by tuning w so that the priors have a prior ESS, m.[9,10] That is, we use the prior ESS as a guide to calibrate the priors. A prior ESS quantifies the prior information in terms of an equivalent number of hypothetical patients. As described in the next section, in the simulation study we will vary the values of m (e.g, $m = 1, 2, \ldots, 10$) to examine the impact of the prior informativeness on the operating characteristics of the study design. The algorithm to derive the prior distributions is summarized as follows:

Step 1: Retrospectively analyze $\mathcal{D}_C$ to estimate $\tilde{\mu}_{\alpha,C}$, $\tilde{\mu}_{\beta,C}$, $\tilde{\gamma}_{\alpha,C}$, and $\tilde{\gamma}_{\beta,C}$ using the model ,
Step 2: Calibrate the informativeness of the priors (4) by tuning w according to the prior ESS.

In step 1, we use the priors (2) and (3) to stabilize the retrospective analysis of $\mathcal{D}_C$. We obtain the two estimates of the probabilities of toxicity at two doses, the second lowest (d_2)

Table 1. True dose-toxicity relationships (true toxicity probability at 6 doses) under 2 scenarios in Caucasian patients and 6 scenarios in Asian patients.

	Dose Level					
Scenario	d_1	d_2	d_3	d_4	d_5	d_6
Caucasian						
1	0.01	0.05	0.10	**0.30**	0.50	0.60
2	0.01	0.02	0.03	0.05	0.10	**0.30**
Asian						
1	0.01	0.05	0.10	**0.30**	0.50	0.60
2	0.05	0.14	**0.36**	0.65	0.86	0.95
3	0.10	**0.30**	0.50	0.60	0.70	0.80
4	**0.41**	0.58	0.82	0.94	0.98	0.99
5	0.01	0.02	0.03	0.05	0.10	**0.30**
6	0.03	0.06	0.12	0.21	**0.36**	0.53

Maximum tolerated doses are shown in boldface.

and second highest (d_{J-1}), from preclinical study data. These two probabilities give the prior means μ_α and μ_β.[7] Then, we assume the common prior variance for α and β (ie, $\gamma_\alpha = \gamma_\beta$) that is specified as having an appropriate amount of prior information (prior ESS = 3) so that the priors never dominate the posterior inference.[9,10]

Power and Noninformative Priors

In this study, we use the most basic version of power prior, that is, the power prior with a fixed $a_0 \in [0, 1]$, rather than expressing uncertainty about a_0 by using an additional prior distribution.[13] With $\mathcal{D}_C$ as historical data, we denote the historical likelihood by $L(\alpha, \beta|\mathcal{D}_C)$. This likelihood is specified by the two-parameter logistic model (1). We use the following conditional power prior for the parameters α and β in the Asian trial,

$$p(\alpha, \beta|D_C, a_0) \propto L(\alpha, \beta|\mathcal{D}_C)^{a_0} p(\alpha, \beta). \quad (5)$$

In this paper we define a_0 as $a_0 = n_C/N_C$, where N_C is the total number of patients treated in the previous Caucasian trial and n_C is an integer $\in [1, N_C]$. Note that n_C in this power prior plays the same role of m in the ESS prior.[9] In the simulation study, we similarly vary the values of n_C from 1 to an appropriate number $< N_C$ as with m. With respect to $p(\alpha, \beta)$, we assume a noninformative normal prior $N(0, 10000)$ for α and a noninformative gamma prior $Ga(0.0001, 0.0001)$ with mean 1 and variance 10,000 for β. We also use the same noninformative priors of α and β in the third method that is based on noninformative priors.

Trial Conduct

Recall that we suppose that the Caucasian phase I trial was conducted with the traditional "3 + 3" cohort design. The Caucasian trial started the dose escalation at the lowest dose d_1. After the MTD in the Caucasian patients was determined according to the "3 + 3" design, 12 patients were added on the MTD level as an expansion cohort.

We carry out an Asian phase I dose-finding trial using the continual reassessment method. That is, we have a continual reassessment method–type goal of finding the "optimal" dose x^*. Optimal is defined as the posterior mean of $\pi(x^*)$ being closest to some fixed target π^*. The maximum sample size is 30, with the cohort size of 3, starting at the lowest dose d_1 and not skipping a dose level when escalating, with target toxicity probability $\pi^* = 0.33$. Dose assignment is based on the posterior distribution conditional on all data available at the time of the decision. This allows for a precise estimation of the dose level with expected toxicity closest to the desired target toxicity.

Simulation Studies

Simulation Study Design

We studied the performance of the proposed study design embedded with the ESS prior (ESS design) by comparing it to the two other designs with the power and noninformative priors (power design and noninformative design) in the setting of a new phase I trial in Asian patients. As summarized in Table 1, we constructed 2 and 6 different toxicity scenarios specifying dose-toxicity relations in the Caucasian and Asian patient groups, respectively. Under all 12 combinations of the 2 and 6 scenarios, we simulated the Caucasian and Asian trials 3000 times. That is, in each of the 3000 simulations, we first generated one set of Caucasian data, analyzed the data for the prior derivation, and then simulated one subsequent Asian trial. The same basic setup for the Asian trial simulations was used in all 3 designs for a fair comparison with respect to the dose levels (= 6 levels; 100, 200, 300, 400, 500, 600 mg), the

maximum number of patients per trial (= 30), cohort size (= 3), starting dose (= d_1), and target $\pi^* = .33$. We investigated the operating characteristics of the ESS design under $m = 1, 3$, and 10, and those of the power design similarly under $n_C = 1, 3$, and 10, as described above. As reference, in the simulations of Caucasian trials, the average number of patients per trial was around 30. Thus, for example, $n_C = 3$ on average corresponds to $a_0 = 0.1$ in the simulations. We carried out the simulation study using the integer values $m = 1$ to 10. Although we drew the figures using all the values of m from 1 to 10, we described the simulation results limited to the values of $m = 1$, 3, and 10 in the tables.

Simulation Results

The operating characteristics for the 3 designs are summarized in Table 2, which is organized into scenario sections. The results are summarized in terms of the percentage of times that each design selected each dose level as the final MTD and the percentages of patients who were treated at each dose level. Correct selection percentages are given in boldface. We also report the average number of patients experiencing toxicity in the trial. The simulation results with the 6 Asian scenarios under Caucasian scenario 1 are shown in Table 2. For each scenario section, the first rows represent the true toxicity probabilities at the 6 dose levels in Asian population of patients.

Under Caucasian scenario 1 and Asian scenario 1, both patient groups have the same MTD (= d_4). The ESS and power designs more correctly selected d_4 as the MTD than the noninformative design, obviously due to the prior information derived from the preceding phase I trial. With increasing m and n_C (incorporating more prior information), the percentage of correct final recommendations gradually increased in both the ESS and power designs.

Under Asian scenario 2, the ESS and power designs more correctly selected the MTD than the noninformative design. The correct selection percentages were higher than those obtained under Asian scenario 1, even for the noninformative design. These high percentages may be in part due to the setup of the relatively high true toxicity probability d_4 (= .65), which may lead to decreasing the selection of d_4 and increasing the selection of d_3.

Under Asian scenario 3, the ESS design appeared to perform better than the power design in terms of selecting d_2 as the MTD for Asian patients. The difference in the performance between those two designs may be partly due to the formulations of the embedded priors. The power prior (5) in a sense directly incorporated toxicity data observed at each of the 6 doses. Thus, it seemed that the power design more intensely reflected the Caucasian data, especially that observed at upper dose levels (ie, d_4 and d_5) than the ESS design. In the simulations of the Caucasian trial, 28.9% and 9.3% of patients were treated at d_4 and d_5, respectively. In contrast, the ESS design, in this paper, constructed the two separate priors for the intercept and slope parameters by analyzing the preceding trial data. This formulation might lead to more desirable performance of the ESS design. In addition, and more interesting, it seemed that the ESS design might have an optimal range of prior informativeness (ie, prior ESS, m) that provides the best performance under several conditions of the study design. Figure 1 shows the percentages of final MTD recommendation at each dose level with respect to prior ESS values ($m = 0$ to 10) under Asian scenario 3. The correct MTD selection (= d_2) percentages got the highest mark in between $m = 1$ and 3, perhaps because the ESS priors with such prior informativeness played an important role as a useful guide for dose escalation/de-escalation decisions early in the trial, and after enrolling 3 patients, the information from the likelihood started to dominate the prior, as desired. The results under the other scenarios are shown in Appendix Figure S1.

Under Asian scenarios 4 and 5, even the noninformative design worked sufficiently well. As expected, the frequency of correct MTD selection gradually decreased in both the ESS and power designs as m and n_C went up. Under Asian scenario 6, the ESS design seemed to perform somewhat better than the power design.

Under Caucasian scenario 2, results and findings were similar to those under Caucasian scenario 1 (Appendix Table S1).

Discussion

We have proposed an approach to quantifying prior information from a previous dose-finding trial to design a subsequent trial in a different population of patients via specified prior distributions. Our proposal is to calibrate the derived priors according to a prior ESS. It is motivated by the idea that one may avoid the use of an overly informative prior in the sense that inference is dominated by the prior rather than the data. Our simulations show that our proposed method has more advantages over the other two methods based on the power and noninformative priors in terms of their performance at estimating MTD in a subsequent Asian dose-finding study.

Several limitations to our proposed approach should be kept in mind. Our approach heavily depends on the model assumption—that is, the two-parameter logistic model for the dose-toxicity relationship. As always, the robustness of our approach to the model assumption should be evaluated before being recommended for practical use. Furthermore, the essential disadvantage of our approach may be in using the information obtained from one single previous study to derive priors for a subsequent trial in a different patient population. To deal with this issue, an extension of our method to combine multiple

Table 2. Simulation results using designs based on the effective sample size prior, power prior, and noninformative prior for a subsequent phase I trial in Asian patients under Caucasian scenario 1.

		Dose Level						Allocation %		
Method		d_1	d_2	d_3	d_4	d_5	d_6	MTD	>MTD	Average Toxicity
Caucasian scenario 1	True toxicity prob.	0.01	0.05	0.10	**0.30**	0.50	0.60			
Asian scenario 1	True toxicity prob.	0.01	0.05	0.10	**0.30**	0.50	0.60			
Noninformative prior		0.1	0.5	9.6	**54.4**	27.6	7.9	34.2	25.2	7.8
Effective sample size prior										
$m = 1$		0.0	0.0	6.3	**63.0**	27.3	3.4	36.6	26.7	8.1
$m = 3$		0.0	0.0	5.9	**63.5**	28.2	2.5	38.0	25.6	8.0
$m = 10$	% Recommendation	0.0	0.0	5.0	**67.3**	26.4	1.4	42.8	22.5	7.9
Power prior										
$n_c = 1$		0.2	0.5	7.9	**58.1**	27.4	5.8	36.5	26.3	8.1
$n_c = 3$		0.2	0.2	7.3	**61.2**	25.8	5.3	37.7	26.2	8.2
$n_c = 10$		0.1	0.3	8.0	**64.5**	22.5	4.6	40.0	23.6	8.0
Asian scenario 2	True toxicity prob.	0.05	0.14	**0.36**	0.65	0.86	0.95			
Noninformative prior		0.7	20.8	**68.5**	8.8	1.2	0.0	45.9	13.6	9.1
Effective sample size prior										
$m = 1$		0.2	18.4	**76.9**	4.4	0.1	0.0	50.6	13.1	9.3
$m = 3$		0.0	15.8	**79.2**	5.0	0.0	0.0	53.8	13.2	9.6
$m = 10$	% Recommendation	0.0	8.7	**85.2**	6.0	0.0	0.0	58.3	16.9	10.4
Power prior										
$n_c = 1$		1.0	18.5	**72.0**	7.9	0.6	0.1	47.0	17.2	9.8
$n_c = 3$		1.8	12.7	**75.9**	9.0	0.5	0.0	49.2	19.7	10.3
$n_c = 10$		1.8	8.4	**77.6**	12.0	0.2	0.0	48.2	25.2	11.2
Asian scenario 3	True toxicity prob.	0.10	**0.30**	0.50	0.60	0.70	0.80			
Noninformative prior		12.3	**56.8**	25.3	4.4	0.9	0.3	46.4	29.0	9.5
Effective sample size prior										
$m = 1$		5.1	**68.9**	24.1	1.7	0.1	0.0	51.4	31.6	10.0
$m = 3$		2.2	**68.4**	27.9	1.5	0.0	0.0	51.6	35.0	10.5
$m = 10$	% Recommendation	0.2	**53.4**	44.2	2.2	0.0	0.0	36.5	53.3	11.9
Power prior										
$n_c = 1$		9.9	**54.5**	30.5	4.2	0.7	0.2	40.1	37.8	10.4
$n_c = 3$		10.0	**48.1**	35.9	5.3	0.5	0.2	34.0	45.1	10.9
$n_c = 10$		9.7	**33.9**	47.8	8.0	0.5	0.1	25.4	57.4	12.2
Asian scenario 4	True toxicity prob.	**0.41**	0.58	0.82	0.94	0.98	0.99			
Noninformative prior		**96.0**	3.9	0.1	0.0	0.0	0.0	88.2	11.8	13.1
Effective sample size prior										
$m = 1$		**96.1**	3.9	0.0	0.0	0.0	0.0	83.1	16.9	13.3
$m = 3$		**92.7**	7.3	0.0	0.0	0.0	0.0	73.6	26.5	13.8
$m = 10$	% Recommendation	**67.1**	32.9	0.0	0.0	0.0	0.0	36.7	63.3	16.2
Power prior										
$n_c = 1$		**92.6**	6.8	0.6	0.0	0.0	0.0	82.4	17.6	13.5
$n_c = 3$		**92.1**	7.2	0.6	0.0	0.0	0.0	78.9	21.1	13.8
$n_c = 10$		**88.6**	10.1	1.1	0.2	0.0	0.0	67.1	32.9	14.9
Asian scenario 5	True toxicity prob.	0.01	0.02	0.03	0.05	0.10	**0.30**			
Noninformative prior		0.0	0.0	0.0	0.4	9.1	**90.4**	40.6	—	4.5
Effective sample size prior										
$m = 1$		0.0	0.0	0.0	0.1	7.6	**92.4**	45.1	—	4.8
$m = 3$		0.0	0.0	0.0	0.0	8.3	**91.6**	44.4	—	4.8
$m = 10$	% Recommendation	0.0	0.0	0.0	0.2	11.2	**88.6**	39.3	—	4.4
Power prior										
$n_c = 1$		0.0	0.1	0.1	0.3	9.6	**89.9**	43.3	—	4.7
$n_c = 3$		0.1	0.0	0.1	0.4	10.6	**88.7**	43.0	—	4.7
$n_c = 10$		0.0	0.2	1.7	4.2	12.0	**81.9**	38.2	—	4.3

(continued)

Table 2. (continued)

Method		Dose Level d_1	d_2	d_3	d_4	d_5	d_6	Allocation % MTD	>MTD	Average Toxicity
Asian scenario 6	True toxicity prob.	0.03	0.06	0.12	0.21	**0.36**	0.53			
Noninformative prior		0.0	0.5	4.5	27.9	**41.5**	25.6	23.6	13.7	7.0
Effective sample size prior										
$m = 1$		0.0	0.0	2.0	31.3	**49.1**	17.7	26.6	13.6	7.4
$m = 3$		0.0	0.0	1.6	32.4	**50.3**	15.7	27.1	11.8	7.2
$m = 10$	% Recommendation	0.0	0.0	1.0	35.9	**52.2**	10.9	28.8	7.6	6.9
Power prior										
$n_c = 1$		0.4	0.5	3.8	29.0	**43.5**	22.8	23.4	16.6	7.5
$n_c = 3$		0.4	0.5	3.2	31.2	**43.1**	21.6	23.6	16.7	7.5
$n_c = 10$		0.2	0.4	3.9	35.5	**41.5**	18.6	22.6	14.8	7.3

Correct selection percentages are given in boldface. MTD, maximum tolerated dose.

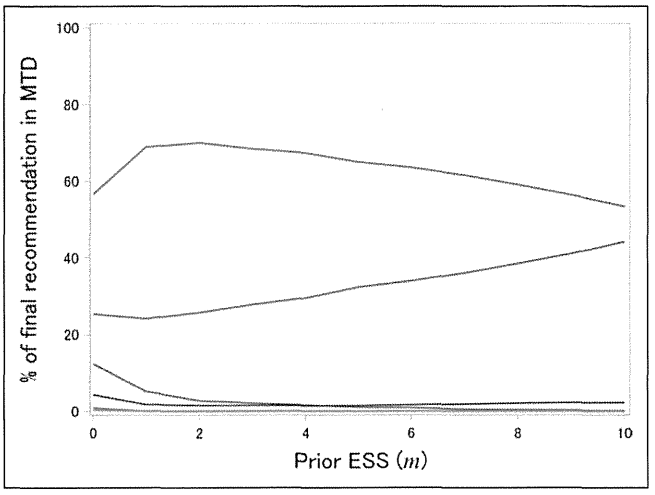

Figure 1. Percentages of final recommended MTDs at each dose level (d_1: blue, d_2: red, d_3: green, d_4: brown, d_5: purple, d_6: pale green) with respect to prior ESS values ($m = 0$ to 10) under Caucasian scenario 1 and Asian scenario 3. ESS, effective sample size; MTD, maximum tolerated dose.

previous trials would be useful. It may be possible to improve the robustness of our method by evaluating the interstudy variability of parameters of interest. We could use Bayesian hierarchical models for these purposes.[14] The prior ESS extended to determine the prior informativeness in a conditionally independent hierarchical model[15] may be useful in this setting.

So far, several Bayesian methods have been proposed for evaluating the similarity of treatment effects among patient subgroups in a randomized clinical trial setting.[16,17] Schoenfeld et al[18] proposed a Bayesian approach to pediatric trial design, which allows borrowing strength from previous or simultaneous adult trials. Taking into consideration that pediatric clinicians often rely on evidence from clinical trials in adults, our proposed method can be applied to a dose-finding study for pediatric cancer by regarding adult patients as in the previous trial. In addition, our proposed method can be extended to all phases of a dose-finding study to incorporate historical data—for example, Asian to Caucasian, preclinical to clinical, monotherapy to combination therapy, and previous regimen to current regimen.

Acknowledgments

We thank Drs Peter Thall and Peter Müller for their helpful comments and useful suggestions. We also thank the associate editor and the referees for their thoughtful and constructive comments and suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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 the nonprofit organization Epidemiological and by the nonprofit organization Epidemiological and Clinical Research Information Network.

References

1. Morita S. Application of the continual reassessment method to a phase I dose-finding trial in Japanese patients: East meets West. *Stat Med.* 2011;30:2090-2097.
2. Ogura T, Morita S, Yonemori K, et al. Exploring ethnic differences in toxicity in early-phase clinical trials for oncology drugs [published online March 3, 2014]. *Therapeutic Innovation & Regulatory Science.*
3. Budman DR, Meropol NJ, Reigner B, et al. Preliminary studies of a novel oral fluoropyrimidine carbamate: capecitabine. *J Clin Oncol.* 1998;16:1795-1802.
4. Mackean M, Planting A, Twelves C, et al. Phase I and pharmacologic study of intermittent twice-daily oral therapy with capecitabine in patients with advanced and/or metastatic cancer. *J Clin Oncol.* 1998;16:2977-2985.

5. Pharmaceuticals Medical Devices Agency. Review reports (capecitabine) [in Japanese]. http://www.info.pmda.go.jp/shinyaku/P200700068/45004500_21500AMZ00400_A100_1.pdf. Accessed March 25, 2014.
6. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 1990;46:33-48.
7. Thall P, Lee S-J. Practical model-based dose-finding in phase I clinical trials: method based on toxicity. *Int J Gynecol Cancer*. 2003;13:251-261.
8. Carter SK. Study design principles for the clinical evaluation of new drugs as developed by the chemotherapy program of the National Cancer Institute. In: Staquet MJ, ed. *The Design of Clinical Trials in Cancer Therapy*. Brussels, Belgium: Editions Scientifiques Europrennes; 1973:242-289.
9. Morita S, Thall PF, Mueller P. Determining the effective sample size of a parametric prior. *Biometrics*. 2008;64:595-602.
10. Morita S, Thall PF, Mueller P. Evaluating the impact of prior assumptions in Bayesian biostatistics. *Stat Biosci*. 2010;2:1-17.
11. Ibrahim JG, Chen MH. Power prior distributions for regression models. *Stat Sci*. 2000;15:46-60.
12. Gilks WR, Richardson S, Spiegelhalter DJ. *Markov Chain Monte Carlo in Practice*. London, England: Chapman & Hall; 1996.
13. Neuenschwander B, Branson M, Spiegelhalter DJ. A note on the power prior. *Stat Med*. 2009;28:3562-3566.
14. Viele K, Berry S, Neuenschwander B, et al. Use of historical control data for assessing treatment effects in clinical trials. *Pharm Stat*. 2014;13:41-54.
15. Morita S, Thall PF, Mueller P. Prior effective sample size in conditionally independent hierarchical models. *Bayesian Anal*. 2012; 7:591-614.
16. Liu JP, Hsiao CF, Hsueh H. Bayesian approach to evaluation of bridging studies. *J Biopharm Stat*. 2002;12:401-408.
17. Goodman SN, Sladky JT. A Bayesian approaches to randomized controlled trials in children utilizing information from adults: the case of Guillain-Barre syndrome. *Clin Trials*. 2005;2:305-310.
18. Schoenfeld DA. Bayesian design using adult data to augment pediatric trials. *Clin Trials*. 2009;6:297-304.

Research Article

Statistics in Medicine

Received 16 August 2013, Accepted 23 April 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6209

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

Satoshi Morita,[a*†] Hideharu Yamamoto[b] and Yasuo Sugitani[b]

The benefits and challenges of incorporating biomarkers into the development of anticancer agents have been increasingly discussed. In many cases, a sensitive subpopulation of patients is determined based on preclinical 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. Extensive 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. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: biomarker; molecular-targeted agent; Bayesian statistics; randomized phase II trial; time-to-event data

1. Introduction

Recently, the benefits and challenges of incorporating biomarkers into the development of anticancer 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 [2–5].

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 antitumor activity in patients with HER2-overexpressing metastatic breast cancer [6–8]. Based on preclinical and clinical data that strongly supported the existence of a sensitive subpopulation of patients, the clinical development of

[a]*Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, Kyoto, Japan*
[b]*Clinical Research Planning Department, Chugai Pharmaceutical Co., Ltd., Tokyo, Japan*
**Correspondence to: Satoshi Morita, Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan.*
†*E-mail: smorita@kuhp.kyoto-u.ac.jp*

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, 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 nonsensitive 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 because of 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], and 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 preceding 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. Preclinical and clinical works suggested that antitumor 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+, and 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, a longer PFS time tended to be observed in patients with a 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 carcinoma, 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 the '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 with 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 $g \geqslant \kappa_0$, 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 that 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 progression-free survival 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 he or she receives 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

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

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 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,

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

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 progression-free survival 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 subgroups 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, $p(\lambda_1 < \eta^* \mid \mathcal{D}_1)$, $\ldots$, $p(\lambda_G < \eta^* \mid \mathcal{D}_G)$ are lower than π^*, we determine $\kappa_0 = G + 1$. We assume a noninformative 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 $\beta_1 > \beta_2 > \cdots > \beta_G$. 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 $\beta_1 = \beta_1$, $\beta_2 = \beta_1 - \gamma_1$, $\ldots$, $\beta_G = \beta_{G-1} - \gamma_{G-1} = \beta_1 - \gamma_1 - \gamma_2 - \cdots - \gamma_{G-1}$, where $\gamma_1 > 0, \gamma_2 > 0, \ldots, \gamma_{G-1} > 0$. Assuming a noninformative normal prior N(0, 1000) for β_1 and a noninformative 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:

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

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 has more flexibility, it may perform more poorly at identifying a sensitive subpopulation because of its S-A 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 nonselected 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 $p(\lambda_g < \eta^* \mid \mathcal{D}) > \pi^*_{stop,Eff}$. Owing to the same reasons mentioned earlier, 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 5000 times using extensively varying situations. We used Markov chain Monte Carlo 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 in the following.

We first performed a series of simulation studies with all 12 combinations of the three fixed upper limits ($\eta^* = 0.70, 0.80, 0.85$), the two upper probability cutoffs ($\pi^* = 0.70, 0.80$), and the two lower probability cutoffs ($\pi^*_{stop} = 0.10, 0.20$) 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 supporting information). After determining the best combination of η^*, π^*, and π^*_{stop}, we evaluated the operating characteristics using six sample size values ($N = 250, 300, 350, 400, 450$, and 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 I, 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 I. 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 $HR_g \leqslant$ 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 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 an efficacy threshold of 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.

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 1–4 for the simulation study.

		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 and 4	0.35	0.15	0.15	0.35
3	Higher in subgroups 2 and 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 and 4	1.0	1.0	**0.5**	**0.3**
(5)	High efficacy only in subgroup 4	1.0	1.0	1.0	**0.5**

The hazard ratio values in the sensitive subpopulation under each scenario are indicated in boldface.

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 independent and identically distributed exp(ν), exponential with parameter ν, which has a PDF of $f(t \mid \nu) = \nu \exp(-\nu t)$, 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 supporting information (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 the following: (i) none of the four subgroups; (ii) subgroup 4 only; (iii) subgroups 3 and 4; (iv) subgroups 2–4; 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 $\eta^* = 0.80$, $\pi^* = 0.70$, 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 the supplementary tables in the supporting information). Table II 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 0.70–0.80. 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, which 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 a 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 (nonsensitive 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 ($N = 250, 300, 350, 400, 450,$ and 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.

Table II. Probabilities of a sensitive subpopulation finding with the fixed upper limit $\eta^* = 0.80$ and the upper and lower probability cutoffs $\pi^* = 0.70$ and $\pi^*_{stop} = 0.20$ when the total sample size $N = 500$.

			Early stopping						
Scenario	Pattern	Method	First	Second	$\mathcal{P}_{none}$	$\mathcal{P}_4$	$\mathcal{P}_{3-4}$	$\mathcal{P}_{2-4}$	$\mathcal{P}_{all}$
(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
(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
(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

The probabilities of identifying (i) none of the four subgroups, (ii) subgroup 4 only, (iii) subgroups 3 and 4, (iv) subgroups 2–4, and (v) all the four subgroups are shown in $\mathcal{P}_{none}$, $\mathcal{P}_4$, $\mathcal{P}_{3-4}$, $\mathcal{P}_{2-4}$, and $\mathcal{P}_{all}$, 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 boldface.

R-M, regression model; S-A, subgroup analysis.

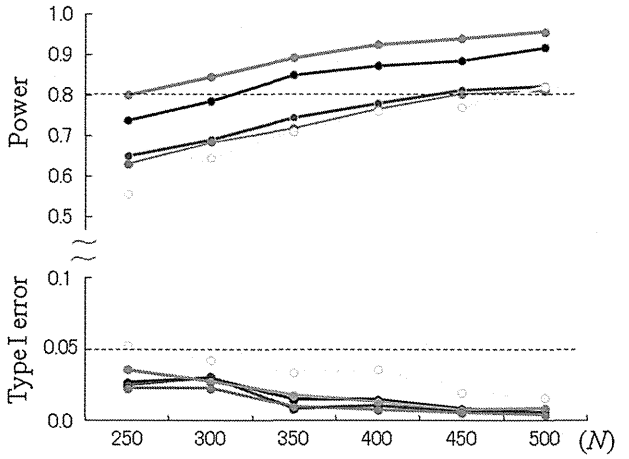

Figure 1. Type I error rates (lower circles) and power values (upper circles) provided by regression model method for the six sample sizes ($N = 250, 300, 350, 400, 450,$ and 500) under the five patterns of subpopulation proportions; patterns 1: black, 2: blue, 3: red, 4: green, and 5: yellow. In this investigation, the power is evaluated by the probability of correctly identifying subgroups 3 and 4 under scenario 4. The fixed design parameters $\eta^* = 0.80$, $\pi^* = 0.70$, and $\pi^*_{stop} = 0.20$ are used.

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 the following:

(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 multinational trial setting. In some cases, however, it may be unrealistic to enroll such a large number of patients into a phase II trial because of 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 [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.

Acknowledgements

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 the nonprofit 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. *Clinical Cancer Research* 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. *Clinical Cancer Research* 2010; **16**:1764–1769.
3. McShane LM, Hunsberger S, Adjei AA. Effective incorporation of biomarkers into phase II trials. *Clinical Cancer Research* 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. *Clinical Cancer Research* 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. *Journal of the National Cancer Institute* 2008; **100**:1204–1214.
6. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: one size does not fit all. *Journal Of Biopharmaceutical Statistics* 2009; **19**:530–542.
7. Buyse M, Michiels S, Sargent DJ, Grothey A, Matheson A, de Gramont A. Integrating biomarkers in clinical trials. *Expert Review of Molecular Diagnostics* 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. *Journal of Clinical Oncology* 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. *Journal of the National Cancer Institute* 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. *Pharmaceutical Statistics* 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. *Statistics in Medicine* 2009; **28**:1445–1463.
14. Eickhoff JC, Kim K, Beach J, Kolesar JM, Gee JR. A Bayesian adaptive design with biomarkers for targeted therapies. *Clinical 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. *Pharmaceutical Statistics* 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? *Journal of Clinical Oncology* 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*, Armitage P, Colton T (eds). John Wiley and Sons: Chichester, 2005; 352–366.
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. *Clinical Cancer Research* 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. *New England Journal of Medicine* 2008; **359**:378–390.
23. Yuan Y, Yin G. Dose–response curve estimation: a semiparametric mixture approach. *Biometrics* 2011; **67**:1543–1554.

Supporting Information

Additional supporting information may be found in the online version of this article at the publisher's web site.

International Regulatory Science

DIA DEVELOP INNOVATE ADVANCE

Therapeutic Innovation & Regulatory Science
2014, Vol. 48(5) 644-650

DOI: 10.1177/2168479014524582
tirs.sagepub.com

Exploring Ethnic Differences in Toxicity in Early-Phase Clinical Trials for Oncology Drugs

Takashi Ogura, MD[1,2], Satoshi Morita, PhD[2], Kan Yonemori, MD[3], Takahiro Nonaka, PhD[1], and Tsutomu Urano, PhD[4,5]

Abstract

During oncology drug development, it is important that ethnic differences are evaluated to determine the optimal dose and administration schedule in a new region based on the clinical data from other regions. The objective of this study was to explore the possibility of detecting ethnic differences in toxicity during early-phase clinical trials. Data were reviewed from phase I clinical trials for new drug applications conducted in Japan and Western countries. The maximum tolerated doses (MTDs), recommended phase II doses (RP2Ds), and approved doses in Japan were compared with those in Western countries. There were 4 of 28 drugs eligible for analysis that showed differences in MTDs or RP2Ds between Japanese and Western patients. Differences in MTDs or RP2Ds in 2 phase I trials were associated with ethnic differences in toxicity. It may be worthwhile to evaluate ethnic differences in toxicity during early-phase clinical trials for oncology drugs.

Keywords

ethnic differences, maximum tolerated dose, oncology drugs, phase I trials

Introduction

Differences in the dosage and dose regimen of some drugs among regions have been pointed out, although they cannot definitely be attributed to ethnic differences.[1,2] Examination of ethnic differences is important while planning and conducting global clinical trials and determining whether clinical data from other countries or regions are applicable for clinical development in new countries or regions.

In the evaluation of ethnic differences during drug development, endogenous factors such as race, sex, and genetic polymorphisms and exogenous factors including socioeconomic factors and health care environments should be considered.[3] Examples of ethnic differences with known causes include those due to genetic polymorphisms in enzymes involved in drug metabolism and the ethnic differences in the distribution of these polymorphisms. In the development of S-1, differences in the distribution of the CYP2A6 polymorphism between Japanese and Western individuals caused different toxicity profiles, leading to differences in the maximum tolerated dose (MTD) and the recommended dose for subsequent clinical trials.[4] For irinotecan, variations in the distribution of the UGT1A1*6 and *28 polymorphisms by ethnicity resulted in different metabolism profiles, which resulted in different levels of toxicity.[5] An example of ethnic differences of unknown cause is the difference in the incidence of interstitial lung disease (ILD) with the use of gefitinib and bortezomib. The incidence of ILD is higher in Japanese patients than in Western counterparts.[6–8] Ethnic differences in safety often pose a serious problem in the development of oncology drugs with narrow therapeutic windows.

If ethnic differences in the incidence of serious adverse events can be predicted early in drug development in a new

[1] Office of New Drug V, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan
[2] Department of Biostatistics and Epidemiology, Yokohama City University Graduate School of Medicine, Yokohama, Japan
[3] Breast and Medical Oncology Division, National Cancer Center Hospital, Tokyo, Japan
[4] Office of Vaccines and Blood Products, Pharmaceuticals and Medical Devices Agency, Tokyo, Japan
[5] Yokohama City University Graduate School of Medicine, Yokohama, Japan

Submitted 12-Dec-2013; accepted 28-Jan-2014

Corresponding Author:
Takashi Ogura, Office of New Drug V, Pharmaceuticals and Medical Devices Agency, Shin-Kasumigaseki Building, 3-3-2, Kasumigaseki, chiyoda-ku, Tokyo, 100-0013 Japan.
Email: ogura-takashi@pmda.go.jp

region or country, it could be determined early on whether clinical data in other regions or countries can be used or whether additional data are necessary, and then clinical development would proceed more appropriately. For example, the development of erlotinib, which targets EGFR in the same manner as gefitinib, was based on information of an ethnic difference with a similar drug—that is, a higher incidence of ILD in Japanese patients with the use of gefitinib. Since this higher incidence was recognized in Japan, studies evaluating safety in Japanese persons were conducted during the development of erlotinib.[9,10] In addition, postmarketing data collection for erlotinib focused on the occurrence of ILD.[11] The clinical development of drugs in new countries or regions will proceed more appropriately if the extent of ethnic differences can be evaluated in an exploratory manner during phase I clinical trials that are first conducted in the residents of the new country or region, in addition to referring to the data on similar drugs.

In the present study, we examined the MTD in phase I clinical trials and the recommended phase II doses (RP2D) and approved doses of new oncology drugs to evaluate whether or not ethnic differences in toxicity can be detected in early-phase clinical trials in new countries or regions.

Methods

We reviewed the data from phase I clinical trials for new drug applications conducted in Japan and Western countries that had been reviewed by the Pharmaceutical and Medical Devices Agency (PMDA) and approved by the Japanese Ministry of Health, Labour, and Welfare between September 1999 and March 2011. Specifically, we examined the PMDA review reports—the documents submitted by the application sponsors, which have been publicly released on the websites of the PMDA[12]—and the published study reports to compare the MTD (or the maximum administered dose, if MTD was not reached) and the RP2D for the Japanese population and that in the US and Europe. The definitions of the terms in this study were as follows: MTD was the lowest dose level at which more than 33% of patients experience dose-limiting toxicity (DLT). RP2D was one dose level below the MTD.

To evaluate ethnic differences between Japanese and Western populations, we compared the approved doses of all drugs according to the prescribe information shown on the website of the regulatory agencies in each region,[13–15] and we retrospectively analyzed the safety profile and frequency of adverse events of all drugs based on the published study reports when differences in MTD or RP2D were identified.

To assess the adequacy of phase I clinical trial design for detecting any differences in toxicity, we compared the dose escalation methods and reasons for stopping dose escalation in the Japanese trials with those conducted in the US and Europe.

No statistical comparisons were made because of the retrospective nature of this analysis.

Results

Between 1999 and 2011, a total of 97 oncology drugs were approved in Japan. Among them, 39 drugs with novel active ingredients were approved. The following drugs were excluded from this study: 4 drugs that had not been approved in the US and Europe (miriplatin, tamibarotene, talaporfin, amrubicin); 3 hormonal drugs (letrozole, exemestane, anastrozole); 2 drugs for which phase I clinical trials were not conducted in Japan (thalidomide, nelarabine); 1 drug for which dose escalation studies were not conducted in the US and Europe (azacitidine); and 1 drug used with different supportive therapies between Japan and the US and Europe (pemetrexed). Thus, 28 drugs were examined in this study.

Drugs With Differences in MTD, RP2D, and Approved Doses Between Japanese and Western Populations

Differences in MTD or RP2D between Japanese and Western populations were observed for 4 of 28 drugs: temsirolimus (with differences only in MTD) and capecitabine, fludarabine, and topotecan (with differences in both MTD and RP2D). Among the drugs with differences in MTD or RP2D, fludarabine and topotecan had different approved dosages and dose regimens. These differences and details of DLT are shown in Table 1. For the drugs without differences in MTD or RP2D, there was also no differences in the approved dosage and dose regimen.

Safety Profiles of the Drugs With Differences in MTD, RP2D, and Approved Doses

The incidence of adverse events with capecitabine—including pigmentation, diarrhea, increased aspartate aminotransferase level, and elevated bilirubin level—was different between Japanese and non-Japanese patients (Table 2). For temsirolimus, a higher incidence of stomatitis and ILD was observed in Japanese persons than in non-Japanese persons (Table 3). The safety profile of topotecan and fludarabine could not be compared owing to the lack of studies conducted using the same dose regimens in Japan as in the US or Europe. However, the occurrence rate of hematologic toxicity with topotecan in Japanese patients is the same as in European patients despite using their different doses, suggesting that there are differences in the occurrence rate of hematologic toxicity between Japanese and European patients (Table 4). Besides, a higher incidence of hematologic toxicity was observed with fludarabine at lower doses in Japanese people than in US people. The

Table 1. MTD, RP2D, approved dose, and DLT of drugs with different toxicity profiles between Japanese and Western populations found in phase I trials.

Drug: Region	MTD or MAD[a]	RP2D	Approved Dose	DLT
Capecitabine				
US	1657 mg/m^2/d; daily	1331 mg/m^2/d; daily	2500 mg/m^2/d; days 1-14 every 3 wk	Hand-foot syndrome, diarrhea, nausea, vomiting, vertigo, dehydration, abdominal pain, dyspnea, venous thrombosis, thrombocytopenia
Europe (UK, NLD)	1657 mg/m^2/d; days 1-14 every 3 wk	2510 mg/m^2/d; days 1-14 every 3 wk	2500 mg/m^2/d; days 1-14 every 3 wk	Hand-foot syndrome, diarrhea, nausea, vomiting, stomatitis, abdominal pain, neutropenia, leucopenia, thrombocytopenia, neutropenia with sepsis
Japan	2510 mg/m^2/d; daily	1657 mg/m^2/d; days 1-21 every 4 wk	2500 mg/m^2/d; days 1-14 every 3 wk[b]	Hemorrhagic gastric ulcer, skin toxicity
Fludarabine				
US	40 mg/m^2/d; days 1-5 every 4 wk	25 mg/m^2/d; days 1-5 every 4 wk for patients without prior therapy[c]	25 mg/m^2/d; days 1-5 every 4 wk	Granulocytopenia, thrombocytopenia
Japan	25 mg/m^2/d	20 mg/m^2/d	20 mg/m^2/d; days 1-5 every 4 wk	Neutropenia, thrombocytopenia
Topotecan				
US	2.5 mg/m^2/d; days 1-5 every 3 wk	*Initial dose:* 1.5 mg/m^2/d; days 1-5 every 3 wk *2nd dose:* 2.0 mg/m^2/d; days 1-5 every 3 wk	1.5 mg/m^2/d; days 1-5 every 3 wk	Neutropenia, febrile neutropenia
Europe (NLD, DNK)	1.5 mg/m^2/d; days 1-5 every 3 wk	1.5 mg/m^2/d; days 1-5 every 3 wk	1.5 mg/m^2/d; days 1-5 every 3 wk	Neutropenia, leukopenia
Japan	1.5 mg/m^2/d; days 1-5 every 3 wk	1.2 mg/m^2/d; days 1-5 every 3 wk	1.0 mg/m^2/d; days 1-5 every 3 wk (maximum dose: 1.5 mg/m^2/d)	Neutropenia, leukopenia
Temsirolimus				
Europe	Not reached (220 mg/m^2)	Not determined	25 mg	Stomatitis, asthenia
Japan	45 mg/m^2	15 mg/m^2	25 mg	Diarrhea, stomatitis

DLT, dose-limiting toxicity; DNK; Denmark; MAD, maximum administered dose; MTD, maximum tolerated dose; NLD, Netherlands; RP2D, recommended phase 2 dose.

[a]If MTD was not reached, MAD was given.

[b]There was a difference in dosage and dose regimen at the time of the first approval application in Japan (2 wk of administration followed by 1 wk without administration in the US and Europe and 3 wk of administration followed by 1 wk without administration in Japan). However, additional clinical studies were conducted in Japan, resulting in the approval of the same dosage and dose regimens as those approved in the US and Europe.

[c]The RP2D was 18 mg/m^2/d for patients with prior chemotherapy or radiotherapy.

incidence of neutropenia was 69% in Japanese people and 18% in US people (Table 5).

Dose Escalation Methods and Reasons for Stopping Dose Escalation

According to the PMDA review reports, for the 28 drugs examined, 78 dose escalation studies were conducted, which consisted of 32 studies in Japanese patients and 46 in European and American patients.

The dose was increased in a 3 + 3 design in 31 of 32 studies in Japanese patients and another design in the remaining study. In the 46 studies in Western persons, the dose was increased in a 3 + 3 design in 37 studies, with a continual reassessment method in 2 studies and other designs in 7 studies (Table 6).

In the 32 studies with Japanese participants, the reason for discontinuation of dose escalation was toxicity in 8 studies, confirmation of the tolerability of the overseas recommended dose in 20 studies, and other in 4 studies. In the 46 studies with Western participants, the reason was toxicity in 24 studies, consideration of pharmacokinetics in 3 studies, achievement of the dose expected to block the target in 3 studies, and other in 16 studies (Table 7).