Chapter 14
Biomarker-Based Designs of Phase III Clinical Trials for Personalized Medicine

Shigeyuki Matsui, Takahiro Nonaka, and Yuki Choai

Abstract Advances in biotechnology and genomics have accelerated development of molecularly targeted treatments and prognostic and predictive biomarkers, particularly, in oncology. This chapter provides an overview of various biomarker-based designs for phase III randomized clinical trials to evaluate clinical utility of a biomarker or biomarker-based treatment, including biomarker-strategy, enrichment, and randomize-all designs. We also provide a simulation comparison of the randomize-all designs in terms of their ability to assert treatment efficacy for the correct patient population. Complex adaptive designs with development and validation of predictive biomarkers are also discussed.

14.1 Introduction

A key component to realize personalized medicine is the development of biomarkers for treatment selection. Biomarkers that are particularly important for personalized medicine can be broadly categorized as prognostic or predictive biomarkers. Prognostic biomarkers are pretreatment or baseline measurements that predict the long-term risk for untreated patients or those receiving the standard treatment, and thus can aid in the decision of whether a patient needs a more aggressive treatment (when diagnosed with high-risk) or no additional treatment (when diagnosed with low-risk). Predictive biomarkers are baseline measurements that provide information about which patients are likely or unlikely to benefit from a specific treatment.

S. Matsui (✉)
Department of Biostatistics, Graduate School of Medicine, Nagoya University, Showa-ku, Nagoya, Japan
e-mail: smatsui@med.nagoya-u.ac.jp

T. Nonaka
Pharmaceuticals and Medical Devices Agency, Chiyoda-ku, Tokyo, Japan
e-mail: nonaka-takahiro@pmda.go.jp

Y. Choai
Department of Statistical Science, School of Multidisciplinary Sciences, The Graduate University for Advanced Studies, Tachikawa, Tokyo, Japan
e-mail: choai@ism.ac.jp

K. van Montfort et al. (eds.), *Developments in Statistical Evaluation of Clinical Trials*, DOI 10.1007/978-3-642-55345-5_14

A predictive biomarker is often designated for the use of a particular new treatment, as a companion biomarker in the development of the new treatment. For example, a biomarker that captures overexpression of the growth factor receptor protein *HER-2*, which transmits growth signals to breast cancer cells, can be a companion biomarker in developing a molecularly-targeted drug for breast cancer patients, trastuzumab (Herceptin ®), which blocks the effects of *HER-2* [24].

A biomarker needs to be validated before its clinical application. Analytical validation refers to establishment of robustness and reproducibility of the assay and accuracy of measurement, such as sensitivity and specificity, relative to a gold standard assay if one is available [3, 22]. Clinical validity refers to establishment of the ability of the biomarker in predicting clinical outcomes in individual patients [22]. For a prognostic biomarker, correlation between biomarker status and a clinical endpoint, such as disease-free or overall survival, may indicate clinical validity. For reliable clinical validation of a predictive biomarker for a survival endpoint, a randomized clinical trial would be required to estimate treatment effects (of a new treatment relative to a control treatment) unbiasedly and to assess whether the treatment effects vary depending on the status of the biomarker, i.e., a treatment-by-biomarker interaction.

The establishment of clinical utility of a biomarker or a new treatment based on a biomarker is finally required as a phase III study before their clinical applications [22]. Randomized clinical trials serve as a gold standard in this phase [2, 7, 9, 13, 16, 17, 20, 23]. One category of biomarker-based designs is to establish clinical utility for the developed biomarker *itself*. The *biomarker-strategy designs* have such an objective. Another category is to establish clinical utility of a new treatment with the aid of a biomarker. The *enrichment designs* and *randomize-all designs* have such an objective. The former is to randomize a biomarker-defined subpopulation of patients, while the latter is to randomize the entire patient population, but entail a *prospective* analysis plan based on the biomarker.

In this chapter, we provide an overview of various biomarker-based designs of phase III clinical trials for personalized medicine. We emphasize again that the two categories of the biomarker-based designs hold distinct objectives, although they have often been discussed as if all of them can be options of biomarker-based designs for a particular situation. We first outline the first category, i.e., the biomarker-strategy designs, in Sect. 14.2. We then focus on the second category; we outline the enrichment designs in Sect. 14.3 and the randomize-all designs in Sect. 14.4. The randomize-all designs can be more complex, reflecting the fact that the development and clinical validation of predictive biomarkers is generally difficult before initiating a phase III clinical trial. Typically, they involve some form of *adaptive* analysis that can demonstrate treatment efficacy for either the overall population or a biomarker-defined subpopulation of patients based on the observed performance of the biomarker. We provide a simulation study to assess their ability to assert treatment efficacy for the right patient population in Sect. 14.5. More complex adaptive designs with both developing and validating a predictive biomarker or genomic signature are outlined in Sect. 14.6. We present concluding remarks in Sect. 14.7.

14.2 Biomarker-Strategy Designs

With a biomarker-strategy design, patients are randomized either to a strategy of using the biomarker in determining their treatment or to a strategy of not using the biomarker in determining treatment. The primary objective is thus to compare two strategies with and without use of the biomarker in determining treatment. An example is a randomized trial for recurrent ovarian cancer that compares the strategy of determining treatment based on tumor chemosensitivity (predictive) assays with a strategy of using physician's choice of chemotherapy based on standard practice [5] (see Fig. 14.1a). Another example is a randomized trial for non-small cell lung cancer that compares a strategy of using a standard treatment (cisplatin+docetaxel) exclusively with a biomarker-based strategy in which patients diagnosed to be resistant to the standard treatment based on the biomarker are treated with an experimental treatment (gemcitabine+docetaxel) and the rest are treated with the standard treatment [4]. In these designs, the biomarker is evaluated only for the patients assigned to the biomarker-based strategy arm.

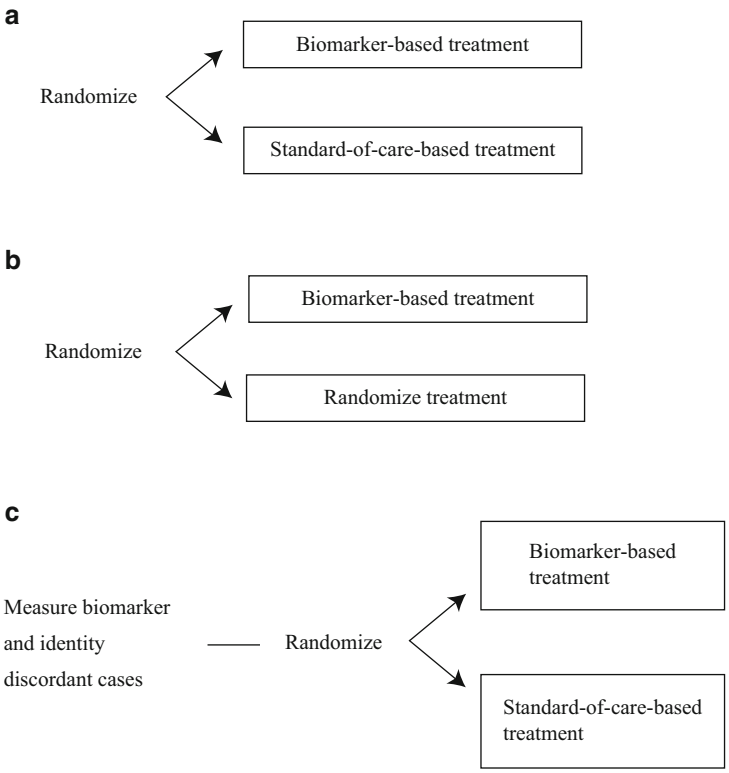

Fig. 14.1 Biomarker strategy designs

For the latter type of design with an experimental treatment, the biomarker-based arm can perform better if the experimental treatment is efficacious, regardless of whether the biomarker is predictive or not. Some authors proposed a modification in which patients in the non-biomarker-based arm undergo a second randomization to receive one of the same two treatments being used in the biomarker-based arm, i.e., the control and experimental treatments [13, 17] (see Fig. 14.1b). By measuring the biomarker status in all of the patients, the modified design would allow clinical validation of the biomarker as a predictive biomarker, through comparing treatment effects across the biomarker-based subsets of patients.

The strategy-based designs fundamentally include patients treated with the same treatment in both the biomarker-based and the non-biomarker-based arms, resulting in a large overlap in the number of patients receiving the same treatment within the two strategies being compared. Thus, a very large number of patients are required to be randomized to detect a diluted, small overall difference in the endpoint between the two arms. One modification is to randomize the two strategies to only the patients for whom the two treatments guided by the two strategies differ (see Fig. 14.1c). This modification requires measurement of the biomarker in all of the patients before randomization. The modified design is generally much more efficient than the original biomarker-strategy design. The modified design was employed in a randomized clinical trial, called the MINDACT study. In this trial, a biomarker-based strategy based on the MammaPrint prognostic signature was compared to that based on standard clinical prognostic factors for determining whether to utilize chemotherapy in women with node-negative estrogen receptor-positive breast cancer, in which discordant cases between the two strategies were subject to randomization [1].

14.3 Enrichment Designs

An enrichment or targeted design is based on a predictive biomarker and compares a new treatment and a control treatment only in biomarker-"positive" (B+) patients who are expected to be responsive to the new treatment based on the biomarker (see Fig. 14.2). Thus, the enrichment design assesses treatment efficacy only in the B+ patients, and not in the entire patient population, including biomarker-negative (B−) patients. In this design, all enrolled patients need to be screened for evaluating the biomarker status.

The efficiency of the enrichment design relative to the standard approach of randomizing all patients without using the biomarker at all depends on the prevalence of the B+ patients and on the effectiveness of the new treatment in the B− patients [12, 18]. In particular, when fewer than half of the patients are B+ and the new treatment is relatively ineffective in the B− patients, the enrichment design can be conducted with much smaller numbers of randomized patients. The enrichment design was employed in the development of trastuzumab; metastatic

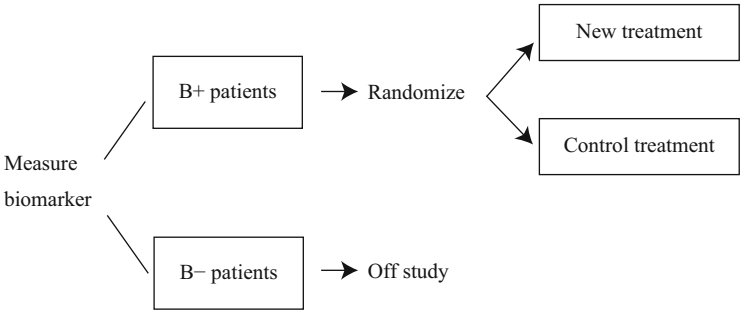

Fig. 14.2 Enrichment design

breast cancer patients whose tumors expressed *HER-2* in an immunohistochemistry test were eligible for randomization [24].

The enrichment design is appropriate for contexts where there is compelling biological evidence for believing that the B− patients will not benefit from the new treatment and that including them would raise ethical concerns [20, 23]. In addition, before initiating the trial, the biomarker used for enrichment must be analytically validated with established assay accuracy, reproducibility, and robustness.

When the biological basis is not compelling and/or assay accuracy is incomplete, assessment of clinical validity of the biomarker as a predictive biomarker would be needed. As the enrichment design does not allow it because of the absence of comparison of the new treatment with the control in the B− patients, the following designs with randomization of both B+ and B− patients, i.e., randomize-all or all-comers designs, are an alternative choice for such situations.

14.4 Randomize-All Designs

Randomization can be either unstratified or stratified on the basis of the predictive biomarker. Unstratified randomization does not diminish the validity of inference regarding treatment effects within the B+ or B− subsets of patients with moderate-to-large sizes. Under unstratified randomization, biomarker can be measured at the time of analysis. This strategy may permit such situations where an analytically validated biomarker is not available at the start of the trial but will be available by the time of analysis [20, 23]. However, careful consideration for missing biomarker data is needed for ensuring collection of sufficient numbers of patients with observed status of biomarker. On the other hand, stratified randomization requires determination and measurement of biomarker at the start of the trial, but ensures that all randomly assigned patients have biomarker status observed (see Fig. 14.3). For other practical considerations in randomized trials with biomarkers, see the references [2, 7, 9, 13, 17, 19, 20, 25, 26, 28].

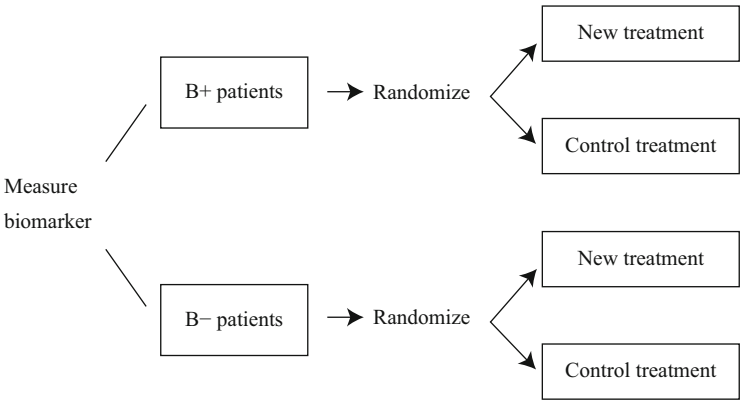

Fig. 14.3 Randomized-all design with prestratification based on the biomarker

The randomize-all designs can demonstrate the efficacy of the treatment for either the overall population or a biomarker-based subset of patients, through inspecting the predictive capability of the biomarker candidate based on the observed trial data. Various designs with a single biomarker candidate have been proposed, including fixed-sequence (FS), fallback (FB), and treatment-by-biomarker-interaction (TBBI) designs.

In what follows, we specifically consider these designs to compare a new treatment and a control treatment based on survival outcomes using a log-rank test. For a particular patient population, we assume proportional hazards between treatment arms and use the asymptotic distribution of a log-rank test statistic S under equal treatment assignment and follow-up, $S \sim N(\theta, 4/E)$ [27]. Here θ is the logarithm of the ratio of the hazard function under the new treatment relative to that under the control treatment, and E is the total number of events observed.

For a clinical trial with a given number of events, we express a standardized test statistic for testing treatment efficacy for the B+ subset of patients as

$$Z_+ = \hat{\theta}_+ / \sqrt{V_+} \,,$$

where $\hat{\theta}_+$ is an estimate of θ_+, such as a log-rank statistic S_+, and $V_+ = 4/E_+$. Similarly, we have a test statistic $Z_- = \hat{\theta}_- / \sqrt{V_-}$ for testing treatment efficacy for the B− subset, where $V_- = 4/E_-$. We consider the following standardized test statistic for testing treatment efficacy for the overall population,

$$Z_{\text{overall}} = \hat{\theta}_{\text{overall}} / \sqrt{V_{\text{overall}}} \,,$$

where $\hat{\theta}_{\text{overall}} = (E_+\hat{\theta}_+ + E_-\hat{\theta}_-)/(E_+ + E_-)$ and $V_{\text{overall}} = 4/E_{\text{overall}} = 4/(E_+ + E_-)$. We assume that the aforementioned standardized statistics follow

asymptotically normal distributions with variance 1, where the means of Z_+, Z_-, and Z_{overall} are $\theta_+/\sqrt{V_+}, \theta_-/\sqrt{V_-}$, and $\sqrt{V_{\text{overall}}}(\theta_+/V_+ + \theta_-/V_-)$, respectively.

14.4.1 FS (Fixed-Sequence) Designs

If evidence from biological or early trial data suggests the predictive ability of the biomarker, it is reasonable to consider first testing treatment efficacy for the B+ subset of patients. In such a situation, one would not expect the treatment to be effective in the B− patients unless it is effective in the B+ patients. As such, the following FS design is derived [20, 23]. In the first stage, we compare the treatment versus control in the B+ patients using the test statistic Z_+ at a significance level of 5 %. If this test is significant, we proceed to the second stage; otherwise, the analysis is stopped. In the second stage, we compare the treatment versus control in the B− patients using the test statistic Z_- at a significance level of 5 %. All tests are two-sided. This sequential approach controls the experiment-wise Type I error at 5 %. When both the first test for the B+ patients and the second test for the B− patients are significant, one may assert treatment efficacy for the overall patient population. When only the first test for the B+ patients is significant, one may assert treatment efficacy only for future patients who are biomarker positive. We refer to this method as the FS-1 design.

A simple way for determining sample size in this design is to ensure the prespecified level of power, such as 90 %, for the first test, and calculate the required number of events for the B+ patients, E_+. This coincides with the required number of events for randomized patients in the enrichment designs. In this calculation, the number of events for the B− patients, E_-, is not determined at the design stage. The B− patients are enrolled concurrently until sufficient numbers of the B+ patients with E_+ are enrolled. As such, E_- can depend on the prevalence of B+, p_+, and the event rates λ_+ and λ_- in the B+ and B− control groups, respectively, at the time that there are E_+ events in the B+ subset. Specifically,

$$E_- = E_+ \left(\frac{\lambda_-}{\lambda_+}\right)\left(\frac{1-p_+}{p_+}\right)$$

is held approximately [20]. We expect a small (large) E_-, especially when p_+ is large (small). A small E_- can lead to a lack of power for detecting clinically important treatment effects in the B− patients at the second stage. On the other hand, a large E_- can yield ethical and practical concerns about enrolling a large number of the B− patients who are unlikely to benefit from the treatment [23]. Hence, sample size determination and/or planning of an interim futility analysis for the B− patients would be warranted.

In another variation of the FS design, the second stage involves testing treatment efficacy for the overall population rather than for the subset of B− patients [13]. With this approach, when only the test for the B+ subset in the first stage is

significant, one may assert treatment efficacy for the B+ subset. When the second overall test is significant (following a significant result in the first stage), one may assert treatment efficacy for the overall population. We refer to this method as the FS-2 design.

14.4.2 FB (Fallback) Designs

When there is limited confidence in the predictive biomarker, it is generally reasonable to assess treatment efficacy for the overall patient population and prepare the subset analysis as a fallback option. Specifically, in the first stage, the treatment is compared with the control overall at a reduced significance level α_1, such as 3 %. If this test is significant, the analysis is stopped. Otherwise, in the second stage, the treatment is compared with the control for the B+ patients at a reduced significance level α_2, such as 2 %, in order to control the experiment-wise type I error rate within 5 % in testing treatment efficacy for the overall population or B+ subset [19,28]. All tests are two-sided. The significance level α_2 can be specified by taking into account the correlation between the first test in the overall population and the second test in the subset of B+ patients [25, 26, 28]. Specifically, the covariance (or correlation) between Z_+ and Z_{overall} reduces to $\sqrt{p_+}$. As the test on treatment efficacy for the overall patient population precedes the fallback test for the B+ patients, it is reasonable to set the significance values such that $\alpha_1 \geq \alpha_2$. When the first test is significant, one may assert treatment efficacy in the overall population. On the other hand, when only the second test for the B+ patients is significant (following a non-significant result of the first test for the overall population), one may assert treatment efficacy only in future B+ patients.

Sample size determination will be based on the first test on treatment efficacy for the overall population, like in the traditional randomized trials, apart from the use of the significance level $\alpha_1 (<0.05)$. Because of possible treatment effects that are clinically important in the B+ patients, it is advisable to perform sample size calculation for the second test for the B+ patients and plan for the option of delaying the second stage analysis until collection of the required number of events for the B+ patients when it is needed [23].

14.4.3 TBBI (Treatment-by-Biomarker Interaction) Designs

TBBI designs, like FB designs, are used when there is limited confidence in the predictive biomarker. This approach involves deciding whether to compare treatments overall or within the biomarker-based subsets based on a preliminary test of interaction of treatment and biomarker [17, 20, 23]. Here the test of interaction is to assess whether there is no difference in treatment effects (in term of the relative hazards ratio between the two treatment arms) between the B+ and B− subsets

of patients. Specifically, we use the following standardized statistic for testing the interaction:

$$Z_{\text{int}} = \frac{\hat{\theta}_+ - \hat{\theta}_-}{\sqrt{V_+ + V_-}}.$$

It is reasonable to consider a one-sided interaction test to detect larger treatment effects in the B+ subset [20, 23]. To be specific, we propose the following design: a preliminary test of interaction is performed as the first stage using Z_{int} at a one-sided significance level of α_{int}. If this test is not significant, the treatment is compared with the control overall using Z_{overall} at a two-sided significance level α_3. Otherwise, the treatment is compared with the control in the B+ patients using Z_+ at a two-sided significance level α_4. Here the significance levels, α_3 and α_4, are chosen such that the experiment-wise type I error rate in testing treatment efficacy for the overall population or B+ subset is less than or equal to 5 % based on an asymptotic distribution of Z_{int}, Z_{overall}, and Z_+, where the covariances between Z_{int} and Z_{overall} or Z_+ may reduce to $\text{cov}(Z_{\text{int}}, Z_{\text{overall}}) = 0$ or $\text{cov}(Z_{\text{int}}, Z_+) = \sqrt{V_+/(V_+ + V_-)} = \sqrt{E_-/(E_+ + E_-)}$. Under the null hypothesis of no treatment efficacy for the B+ and B− patients (and thus indicating no effects for the overall population), for which we will search for the significance level, α_4, for Z_+, given α_{int} for Z_{int} and α_3 for Z_{overall}, to control the experiment-wise type I error rate within 5 %, we propose to set $\text{cov}(Z_{\text{int}}, Z_+) = \sqrt{1 - p_+}$ if the hazard rate in the B+ subset can be considered to be the same as that in the B− subset. When the predictive biomarker is prognostic, a larger number of events is expected for the B+ patients, resulting in an overestimation of the correlation. This would lead to use of a stringent significance level of α_4 and thus a conservative design.

When the test for the B+ patients is significant (following a significant result of the preliminary interaction test), one may assert treatment efficacy only for B+ patients. When the overall test is significant (following a non-significant result of the preliminary interaction test), one may assert treatment efficacy for the overall population.

The TBBI designs have been discussed in the literature as a design for clinical validation of the predictive biomarker based on a test on treatment-by-biomarker interaction [17, 20, 23]. However, sizing the trial to have high power for the interaction test may require a substantially large sample size, compared to sizing trials with the other randomize-all designs. This cannot generally be justified as it requires exposing an excessive number of B− patients to a treatment from which they are unlikely to benefit [23].

On the other hand, the proposed TBBI design with strict control of the experiment-wise type I error rate described above aims to assess the clinical utility of a new treatment with the aid of the biomarker. As our simulation study indicated (see Sect. 14.5.1), it could be more efficient compared with the other randomize-all designs. An additional advantage of the proposed TBBI design is that even if the interaction test is regarded as a preliminary test, a significant interaction could

be regarded as relatively firm evidence for the clinical validity of the biomarker. Further studies on the proposed TBBI design, including determination of optimal levels of α_{int} and α_3 and sample size determination, would be worthwhile.

14.5 Probability of Asserting Treatment Efficacy

The randomize-all designs described in Sect. 14.4 can make either of two kinds of assertions regarding treatment efficacy, one for the overall population and the other for the B+ subset of patients. Which of the two assertions is considered to be valid may depend on the underlying treatment effects in the biomarker-based subsets. Specifically, let HR_+ and HR_- denote the hazard ratios of the treatment relative to the control in the B+ and B− subsets of patients, respectively. If the treatment truly has clinically meaningful effects in all of the patients, e.g., $HR_+ = HR_- = 0.7$, the assertion of treatment efficacy for the overall population would be more valid than that for the B+ patients only because the latter assertion would deprive the remaining B− patients of the chance of receiving the effective treatment. On the other hand, if the treatment can exert a clinically important effect only in the B+ patients, e.g., $HR_+ = 0.5$, and no effect in the remaining B− patients, e.g., $HR_- = 1.0$ (indicating a qualitative interaction between treatment and biomarker), the assertion of treatment efficacy for the B+ patients would be more valid than that for the overall population because the latter assertion would yield overtreatment for the remaining B− patients using the ineffective, even toxic treatment. Let P_{overall} and P_{subset} denote the probability of asserting treatment efficacy for the overall population and for the subset of B+ patients, respectively.

However, there can be other scenarios in which it is not clear which of the two assertions is valid. For example, the treatment can exert a clinically important effect for the B+ patients, e.g., $HR_+ = 0.5$, but some moderate or small effects for the remaining B− patients, e.g., $HR_- = 0.8$ (indicating a quantitative interaction between treatment and biomarker). Such a treatment effect profile could be explained by the treatment having multiple mechanisms of action, the misclassification of responsive patients into the B− subset (low sensitivity of the biomarker), and so on. Which of the two assertions is considered to be valid will be determined on a case-by-case basis incorporating many factors, including the size of the prevalence p_+, possible adverse effects, treatment costs, prognosis of the disease, availability of other treatment choices, and so on. In such situations, the probability of asserting treatment efficacy for either the overall population or the subset of B+ patients could be another meaningful criterion. From the point of view of treatment developers (e.g., pharmaceutical companies), this probability would be always important, because it can be interpreted as the *probability of success* in treatment development. Let P_{success} denote this probability. Apparently, $P_{\text{overall}} + P_{\text{subset}} = P_{\text{success}}$ for the randomize-all designs described in Sect. 14.4. As such, there is a trade-off between the two probabilities P_{overall} and P_{subset} for a given value of P_{success}.

Table 14.1 Empirical probabilities of P_{overall}, P_{subset}, and P_{success} under null effects

HR_+	HR_-	p_+	Prob.	Traditional	FS-1	FS-2	FB	TBBI	
								$\alpha_{\text{int}} = 5\,\%$	$\alpha_{\text{int}} = 10\,\%$
1.0	1.0	0.1	P_{overall}	0.051	0.003	0.005	0.032	0.030	0.029
(null effect)			P_{subset}	0.000	0.041	0.039	0.016	0.019	0.021
			P_{success}	0.051	0.044	0.044	0.047	0.049	0.050
		0.3	P_{overall}	0.050	0.002	0.010	0.031	0.029	0.028
			P_{subset}	0.000	0.048	0.040	0.020	0.020	0.022
			P_{success}	0.050	0.050	0.050	0.051	0.049	0.050
		0.5	P_{overall}	0.052	0.002	0.018	0.031	0.029	0.028
			P_{subset}	0.000	0.047	0.032	0.019	0.021	0.020
			P_{success}	0.052	0.050	0.050	0.050	0.050	0.048

14.5.1 Simulations

We provide a comparison of the randomize-all designs in Sect. 14.4 in terms of P_{overall}, P_{subset}, and P_{success}. We considered the prevalence of B+, $p_+ = 0.1, 0.3$, or 0.5. As to the underlying treatment effects within biomarker-based subsets, we considered the following scenarios: $(HR_+, HR_-) = (1.0, 1.0), (0.7, 0.7), (0.5, 1.0)$, or $(0.5, 0.8)$, i.e., null effects, constant effects, qualitative interaction, and quantitative interaction. In the FB and TBBI designs, we specified the same significance levels for the overall test, $\alpha_1 = \alpha_3 = 3\,\%$, for a fair comparison of these designs. The significance level for the one-sided interaction test, α_{int}, in the TBBI designs was specified as 5 or 10 %. The significance levels for the B+ subset tests, α_2 and α_4, in the FB and TBBI designs were determined such that the experiment-wise type I error rates were equal to 5 %. We also evaluated the traditional design without use of a biomarker as a reference, with $P_{\text{overall}} = P_{\text{success}}$ and $P_{\text{subset}} = 0$ (because there is no option for asserting treatment efficacy for the B+ subset in this design). We conducted 10,000 simulations (clinical trials) for each configuration to obtain empirical values of the probabilities. We provide the results when 400 patients with a baseline event rate of 0.2 (per year) are randomized and followed up for 5 years in each clinical trial. For larger sample sizes, P_{overall}, P_{subset}, and P_{success} became large, but similar conclusions in terms of the relative sizes of these probabilities across the designs under comparison were obtained. R codes for conducting simulations are available from author upon request. A web-based simulation program that provides estimates of required sample size for biomarker-based analysis plans for time to event or binary endpoints is also available [15].

We first confirmed control of the experiment-wise type I error rate, i.e., $P_{\text{success}} \leq 5\,\%$, for all of the designs in Table 14.1. We also confirmed control of P_{overall} as the specified significance levels for the overall tests, $\alpha_1 = \alpha_3 = 3\,\%$, for the FB and TBBI designs.

Table 14.2 summarizes the empirical values of P_{overall}, P_{subset}, and P_{success} for scenarios with non-null treatment effects. For the scenarios with constant treatment

Table 14.2 Empirical probabilities of P_{overall}, P_{subset}, and P_{success} under non-null treatment effects

								TBBI	
HR_+	HR_-	p_+	Prob.	Traditional	FS-1	FS-2	FB	$\alpha_{\text{int}} = 5\,\%$	$\alpha_{\text{int}} = 10\,\%$
0.7	0.7	0.1	P_{overall}	0.758	0.083	0.106	0.690	0.658	0.623
(constant effect)			P_{subset}	0.000	0.036	0.013	0.007	0.045	0.079
			P_{success}	0.758	0.120	0.120	0.698	0.703	0.702
		0.3	P_{overall}	0.774	0.198	0.300	0.703	0.669	0.634
			P_{subset}	0.000	0.124	0.022	0.020	0.052	0.098
			P_{success}	0.774	0.322	0.322	0.723	0.721	0.732
		0.5	P_{overall}	0.764	0.222	0.450	0.691	0.659	0.623
			P_{subset}	0.000	0.252	0.025	0.027	0.049	0.097
			P_{success}	0.764	0.474	0.474	0.717	0.708	0.720
0.5	1.0	0.1	P_{overall}	0.074	0.016	0.039	0.048	0.027	0.019
(qualitative interaction)			P_{subset}	0.000	0.301	0.279	0.178	0.296	0.304
			P_{success}	0.074	0.317	0.317	0.225	0.323	0.323
		0.3	P_{overall}	0.301	0.038	0.281	0.230	0.053	0.031
			P_{subset}	0.000	0.719	0.476	0.449	0.706	0.743
			P_{success}	0.301	0.757	0.757	0.680	0.759	0.774
		0.5	P_{overall}	0.688	0.047	0.682	0.607	0.102	0.052
			P_{subset}	0.000	0.891	0.256	0.305	0.825	0.893
			P_{success}	0.688	0.938	0.938	0.913	0.927	0.945
0.5	0.8	0.1	P_{overall}	0.519	0.115	0.205	0.432	0.326	0.266
(quantitative interaction)			P_{subset}	0.000	0.192	0.102	0.077	0.222	0.278
			P_{success}	0.519	0.307	0.307	0.509	0.548	0.544
		0.3	P_{overall}	0.762	0.232	0.644	0.692	0.369	0.270
			P_{subset}	0.000	0.532	0.119	0.124	0.455	0.584
			P_{success}	0.762	0.764	0.764	0.816	0.824	0.854
		0.5	P_{overall}	0.914	0.214	0.882	0.873	0.403	0.288
			P_{subset}	0.000	0.722	0.054	0.073	0.533	0.666
			P_{success}	0.914	0.936	0.936	0.946	0.937	0.954

effects, $(HR_+, HR_-) = (0.7, 0.7)$, where P_{overall} would be a relevant criterion, the traditional design provided the greatest values of P_{overall}, as was expected. The FB and TBBI designs provided slightly reduced values of P_{overall} than those of the traditional design. On the other hand, the FS designs, especially FS-1, provided much smaller values of P_{overall}. Similar trends were observed for P_{success}.

For the scenarios with a qualitative interaction, $(HR_+, HR_-) = (0.5, 1.0)$, where P_{subset} would be relevant, the FS-1 and TBBI designs performed best. The FS-2 and FB designs provided much smaller values of P_{subset} when $p_+ \geq 0.3$. With respect to P_{success}, all biomarker-based designs, except the FB design, generally provided comparable P_{success} values, while the traditional design provided much smaller values of P_{success}.

Lastly, for the scenarios with a quantitative interaction, $(HR_+, HR_-) = (0.5, 0.8)$, the characteristics of the respective designs became clearer. The FS-2 and FB designs tended to provide larger P_{overall}, while the FS-1 and TBBI designs tended to provide larger P_{subset} values. With respect to P_{success}, the TBBI designs provided the largest P_{success} values, followed by the FB design with slight reductions in P_{success}.

In summary, the FS-1 design would be suitable for cases with qualitative interactions between treatment and biomarker and large treatment effects in the B+ patients, but could suffer from a serious lack of power for nearly constant treatment effects in the overall population. Interestingly, the FS-2 design has quite different properties, but was not shown to be so efficient for various profiles of treatment effects. In contrast, a FB design would be suitable for cases with nearly constant treatment effects in the overall population, but could suffer from a serious lack of power for qualitative interactions between treatment and biomarker. The TBBI designs generally performed well for various patterns of treatment effects within biomarker-based subsets in terms of all the probabilities, P_{overall}, P_{subset} and P_{success}. This can be explained by the effectiveness of the preliminary interaction test in selecting the appropriate population for testing treatment efficacy.

14.6 More Complex Adaptive Designs

When the biology of the target of a new treatment is not well understood because of the complexity of disease biology, it is quite common that a completely specified predictive biomarker is not available before initiating the definitive phase III trial. One approach in such situations is to design and analyze the randomized phase III trial in such a way that both developing a predictive biomarker and testing treatment efficacy based on the developed biomarker are possible and conducted validly. Apparently, this approach works with randomize-all designs without prestratification based on any biomarkers, and careful prespecification of the analysis plan is mandatory.

Jiang et al. [10] developed the *adaptive threshold design* for settings where a single predictive biomarker candidate is available but no threshold of positivity for the biomarker is predefined. The basic idea is, for a set of candidate threshold values $(b_1, \ldots, b_K)$, to search for an optimal threshold value through maximizing a log likelihood ratio of treatment effect for the patients with biomarker value $\geq b_k$ over possible threshold values $(k = 1, \ldots, K)$. The maximum log likelihood ratio at the optimal threshold value is used as the test statistic. Its null distribution is approximated by repeating the whole analysis after randomly permuting treatment levels several thousand times. This approach can be applied to searching for a subset determined by a positive value of any single biomarker when there is a set of candidate binary biomarkers [23]. This approach can be used as the second stage analysis of the FB designs or as a stand-alone basis by incorporating the log

likelihood statistic for testing the overall treatment effects in obtaining a maximum test statistic [10].

Another adaptive design, called *adaptive signature design*, is to develop a predictor or signature using a set of covariates x, possibly high-dimensional genomic data [6,8]. As the second stage of the FB designs, the full set of patients in the clinical trial is partitioned into a training set and a validation set. A prespecified algorithmic analysis plan is applied to the training set to generate a predictor. This is a function of x and to predict, for a given patient with a particular value of x, to be responsive or not responsive to the new treatment. The predictor is used to make a prediction for each patient in the validation set. Then, the treatment efficacy is tested for the patient subset predicted as "responsive" to the treatment in the validation set.

This modified second stage analysis of the FB designs can be based on split-sample [6] or cross-validation [8]. In the latter approach, at the end of the prediction process, each of all the patients in the clinical trial is predicted as either responsive or not. Again, the treatment efficacy is tested for the patient subset predicted as "responsive" to the treatment. However, because this subset is determined by the cross-validation using the all patient data, the standard asymptotic theory does not apply. To address this issue, a permutation method that repeats the whole processes of the cross-validated prediction analysis after randomly permuting treatment levels is employed [8].

Recently, Matsui et al. [14] developed another framework designed to estimate treatment effects quantitatively as a function of a continuous cross-validated predictive score for the entire patient population, rather than qualitatively classifying patients as in or not in a responsive subset. Average absolute treatment effects for the entire population or a responsive subset of patients can be estimated based on the estimated treatment effects function and tested using a permutation method. In this framework, patient-level survival curves can be developed to predict survival distributions of individual future patients as a function of the cross-validated predictive score and a cross-validated prognostic score that is developed independently from the development of the predictive score, through correlating genomic data with survival outcomes without reference to treatment assignment.

14.7 Concluding Remarks

In this chapter, we have discussed a wide variety of biomarker-based designs of phase III clinical trials to establish the clinical utility of a biomarker or a new treatment with the aid of a biomarker. In biomarker-strategy, enrichment, and prestratified randomize-all designs, collection of specimens and biomarker assays are conducted prospectively for newly accruing patients. As these prospective designs are highly resource-intensive and time-consuming, a study using archived specimens is sometimes used as an alternative. This type of study is retrospective with regard to using archived specimens, but should prospectively specify a protocol. An unstratified randomize-all trial, possibly with the adaptive designs in

Sect. 14.6, could be categorized to this type of study because specimens archived at the beginning of the trial are analyzed. Simon et al. [21] proposed several conditions for appropriately conducting such a study with archived specimens. In summary,

1. Archived specimen, adequate for a successful assay, must be available from a sufficient large number of patients to permit appropriately powered analyses in the pivotal trial and to ensure that the patients included in the biomarker evaluation are representative of the patients in the trial.
2. Substantial data on the analytical validity of the biomarker must exist to ensure that results obtained from the archived specimens will closely resemble those that would have been obtained from analysis of specimens collected in real time. Assays should be conducted blinded to the clinical data.
3. The analysis plan for the biomarker evaluation must be completely developed before the performance of the biomarker assays. The analysis should focus on a single diagnostic biomarker that is completely defined and specified. The analysis should not be exploratory, and practices that might lead to a false-positive conclusion (e.g., multiple analyses of different candidate biomarkers based on archived specimens from the same trial) should be avoided.
4. The results must be validated in at least one or more similarly designed studies using the same assay techniques.

These conditions are also applicable to previously conducted clinical trials (with archived specimens) that evaluated the efficacy of the treatment of interest. When substantial preliminary evidence that a new biomarker predicts treatment responsiveness has been accumulated by the middle or completion of a phase III trial of the treatment, one may consider assay of the biomarker in archived specimens from this trial. As an example, an analysis based on a *KRAS* mutation in a randomized trial for the *anti-EGFR* antibody, cetuximab, which was approved for the treatment of advanced colorectal cancer, demonstrated that the treatment was not effective for patients with *KRAS* mutations [11]. Another possibility is to analyze archived specimens from a failed pivotal trial that showed no treatment effect for the entire patient population using the methods for biomarker development described in Sect. 14.6. The developed biomarker from such an analysis can provide useful information for designing a second confirmatory trial of the same treatment, possibly with an enrichment design with small sample sizes.

The recent advances in biotechnology and genomics have posed biostatisticians further important roles and challenges in various phases of biomarker development and validation, including systematic collection of specimens and measurement of biomarker/clinical data, development of an analytically and clinically-validated biomarker, and establishment of the clinical utility of the biomarker or biomarker-based treatment, through utilizing archived or prospectively-collected specimens in the context of clinical trials. Further biostatistical researches are required indeed in this important field for accelerating modern clinical studies toward personalized medicine.

Acknowledgements This research was partly supported by a Grant-in-Aid for Scientific Research (24240042) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The views expressed in this chapter are the result of independent work and do not necessarily represent the views of the Pharmaceuticals and Medical Devices Agency.

References

1. Bogaerts, J., Cardoso, F., Buyse, M., Braga, S., Loi, S., et al.: Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nature Clinical Practice Oncology* **3**(10), 540–551 (2006). doi:10.1038/ncponc0591
2. Buyse, M., Michiels, S., Sargent, D.J., Grothey, A., Matheson, A., et al.: Integrating biomarkers in clinical trials. *Expert Review of Molecular Diagnostics* **11**(2), 171–182 (2011). doi:10.1586/erm.10.120
3. Chau, C.H., Rixe, O., McLeod, H., Figg, W.D.: Validation of analytic methods for biomarkers used in drug development. *Clinical Cancer Research* **14**(19), 5967–5976 (2008). doi:10.1158/1078-0432.CCR-07-4535
4. Cobo, M., Isla, D., Massuti, B., Montes, A., Sanchez, J.M., et al.: Customizing cisplatin based on quantitative excision repair cross-complementing 1 mRNA expression: a phase III trial in non-small-cell lung cancer. *Journal of Clinical Oncology* **25**(19), 2747–2754 (2007). doi:10.1200/JCO.2006.09.7915
5. Cree, I.A., Kurbacher, C.M., Lamont, A., Hindley, A.C., Love, S.: A prospective randomized controlled trial of tumour chemosensitivity assay directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. *Anticancer Drugs* **18**(9), 1093–1101 (2007). doi:10.1097/CAD.0b013e3281de727e
6. Freidlin, B., Simon, R.: Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **11**(21), 7872–7878 (2005). doi:10.1158/1078-0432.CCR-05-0605
7. Freidlin, B., McShane, L.M., Korn, E.L.: Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Insititute* **102**(3), 152–160 (2010). doi:10.1093/jnci/djp477
8. Freidlin, B., Jiang, W., Simon, R.: The cross-validated adaptive signature design. *Clinical Cancer Research* **16**(2), 691–698 (2010). doi:10.1158/1078-0432.CCR-09-1357
9. Hoering, A., Leblanc, M., Crowley, J.J.: Randomized phase III clinical trial designs for targeted agents. *Clinical Cancer Research* **14**(14), 4358–4367 (2008). doi:10.1158/1078-0432.CCR-08-0288
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 Insititute* **99**(13), 1036–1043 (2007). doi:10.1093/jnci/djm022
11. Karapetis, C.S., Khambata-Ford, S., Jonker, D.J., O'Callaghan C.J., Tu D., et al.: K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *New England Journal of Medicine* **359**(17), 1757–1765 (2008). doi:10.1056/NEJMoa0804385.
12. Maitournam, A., Simon, R.: On the efficiency of targeted clinical trials. *Statistics in Medicine* **24**(3), 329–339 (2005). doi:10.1002/sim.1975
13. Mandrekar, S.J., Sargent, D.J.: Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology* **27**(24), 4027–4034 (2009). doi:10.1200/JCO.2009.22.3701
14. Matsui, S., Simon, R., Qu, P., Shaughnessy, J.D. Jr, Barlogie, B., et al.: Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine. *Clinical Cancer Research* **18**(21), 6065–6073 (2012). doi:10.1158/1078-0432.CCR-12-1206
15. National Institutes of Health: Sample Size Calculation for Randomized Clinical Trials. http://linus.nci.nih.gov/brb/samplesize/sdpap.html

16. Pusztai, L., Hess, K.R.: Clinical trial design for microarray predictive marker discovery and assessment. *Annals of Oncology* **15**(12), 1731–1737 (2004). doi:10.1093/annonc/mdh466
17. Sargent, D.J., Conley, B.A., Allegra, C., Collette, L.: Clinical trial designs for predictive marker validation in cancer treatment trials. *Journal of Clinical Oncology* **23**(9), 2020–2027 (2005). doi:10.1200/JCO.2005.01.112
18. Simon, R., Maitournam A.: Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research* **10**(20), 6759–6763 (2004). doi:10.1158/1078-0432.CCR-04-0496
19. Simon, R., Wang, S.J.: Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics Journal* **6**(3), 166–173 (2006). doi:10.1038/sj.tpj.6500349
20. Simon, R.: The use of genomics in clinical trial design. *Clinical Cancer Research* **14**(19), 5984–5993 (2008). doi:10.1158/1078-0432.CCR-07-4531
21. Simon, R., Paik, S., Hayes, D.F.: Use of archived specimens in evaluation of prognostic and predictive biomarkers. *Journal of the National Cancer Insititute* **101**(21), 1446–1452 (2009). doi:10.1093/jnci/djp335
22. Simon, R.: Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Medicine* **7**(1), 33–47 (2010). doi:10.2217/pme.09.49
23. Simon, R.: Clinical trials for predictive medicine. *Statistics in Medicine* **31**(25), 3031–3040 (2012) doi:10.1002/sim.5401
24. Slamon, D.J., Leyland-Jones, B., Shak, S., Fuchs, H., Paton, V., et al.: Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine* **344**(11), 783–792 (2001). doi:10.1056/NEJM200103153441101
25. Song, Y., Chi, G.Y.: A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine* **26**(19), 3535–3549 (2007). doi:10.1002/sim.2825
26. Spiessens, B., Debois, M.: Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials* **31**(6), 647–656 (2010). doi:10.1016/j.cct.2010.08.011
27. Tsiatis. A.A.: The asymptotic joint distribution of the efficient score test for the proportional hazards model calculated over time. *Biometrika* **68**(1), 311–315 (1981). doi:10.1093/biomet/68.1.311
28. Wang, S.J., O'Neill, R.T., Hung, H.M.: Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics* **6**(3), 227–244 (2007). doi:10.1002/pst.300