Published in final edited form as:

J Biopharm Stat. 2018 ; 28(2): 256–263. doi:10.1080/10543406.2017.1372777.

Phase II cancer clinical trials for biomarker-guided treatments

Sin-Ho Jung

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

Abstract

The design and analysis of cancer clinical trials with biomarker depend on various factors, such as the phase of trials, the type of biomarker, whether the used biomarker, is validated or not, and the study objectives. In this article, we demonstrate the design and analysis of two Phase II cancer clinical trials, one with a predictive biomarker and the other with an imaging biomarker. Statistical testing methods and their sample size calculation methods are presented for each trial. We assume that the primary endpoint of these trials is a time to event variable, but this concept can be used for any type of endpoint.

Keywords

Enrichment trial; interaction; predictive biomarker; prognostic biomarker; progression-free survival; stratified randomized trial

1. Introduction

In many cancer clinical trials, different types of biomarkers are measured from the tumor, blood, or urine using molecular, biochemical, physiological, anatomical, or imaging method at the baseline or in the middle of trials. The observed biomarkers are used for various purposes during the diagnosis and treatment of the diseases. For example, cancer biomarkers are used to diagnose diseases (diagnostic biomarker), to predict the response to a specific treatment (predictive biomarker), to measure the aggressiveness of a disease for patients with no or nontargeted treatment (prognostic biomarker), to monitor the recurrence of a disease, and so on.

These biomarkers can be used to select a treatment of cancer patients. However, biomarkers should be validated before used to select a treatment in clinical trials. If a biomarker has not been validated yet, it can be used as a stratification factor of a randomized clinical trial. In such a trial, the biomarker is used for its validation rather than for treatment selection.

The design and analysis method of a clinical trial with a biomarker can be very different depending on the type of the used biomarker, the biomarker's development stage and the study objective. Freidlin, McShane, and Korn (2010) discuss various design issues of randomized clinical trials with biomarkers. Freidlin et al. (2012) propose a series of statistical testing for a randomized Phase II trial with a potentially predictive biomarker

which is not strictly validated yet. Mandrekar and Sargent (2009) compare the efficacy of enrichment trials and stratified randomization trials with a time to event variable as the primary endpoint. They assume that the treatment effect reverses between biomarker positive and negative groups and they consider subset analysis within each biomarker group.

Phase II trials are to screen out inefficacious treatments before proceeding to large-scale studies, such as a Phase III trial. As such, Phase II trials should be completed in a short time period, so that we have to choose a small sample size and a surrogate endpoint, such as tumor response or progression-free survival (PFS), as the primary endpoint, rather than a confirmatory endpoint, such as overall survival.

In this article, two Phase II cancer trials are demonstrated, one with a predictive biomarker and the other with an imaging biomarker, which the author designed and analyzed. We present analysis and sample size calculation methods for these trials. We use a survival variable as the primary endpoint in this article, but the same concept can be used for any kind of variables including a binary variable. For the purpose of sample size calculation, we assume exponential survival distributions which are most popularly used in real trial designs, although the statistical testing does not require a specific survival distribution.

2. A Phase II trial with a predictive biomarker

Predictive biomarkers help to provide information on the likelihood of response to a specific chemotherapy. For example, tumors expressing high thymidylate synthase (TS) levels were shown to be resistant to pemetrexed in a preclinical study (Ozasa et al., 2010), but this was not validated by a clinical study yet. Suppose that we want to investigate whether TS expression is a predictive marker for the clinical outcome of pemetrexed/cisplatin (PC) in patients with nonsquamous non-small-cell lung cancer (NSCLC) through a Phase II trial. The control untargeted treatment is gemcitabine/cisplatin (GC). In order to investigate this hypothesis, we want to randomize patients between the two treatment arms stratifying by TS positivity vs. negativity. When this study is completed, in order to test the study objective, we will fit a Cox (1972) proportional hazards regression model with treatment allocation z_1 , TS positivity z_2 , and their interaction term as covariates:

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2) \quad (1)$$

and test $H_0: \beta_3 = 0$.

In fact, this study was designed and published with overall response as the primary endpoint (Sun et al., 2015). In this article, we demonstrate how to design this study using progression-free survival (PFS) as the primary endpoint based on the estimates from the previous study.

2.1 Sample size calculation

For sample size calculation of this study, we need to specify the following design parameters.

- Type I error rate and power, $(\alpha, 1 - \beta)$

- Allocation proportions, p_1, p_2 between two treatment arms ($p_1 + p_2 = 1$)
- Biomarker positivity, q_1
- Survival distributions for the four patient groups defined by treatment allocation and biomarker outcome, e.g., $\lambda_{00}, \lambda_{01}, \lambda_{10}, \lambda_{11}$.
- Accrual period a (or accrual rate r) and additional follow-up period b .

Two covariates are defined as $z_1 = 0$ for GC arm and 1 for PC arm; $z_2 = 0$ for TS positivity and 1 for TS negativity. We specify the proportion of each cell of the 2×2 table, $p_{kl} = P(z_1 = k, z_2 = l)$ for $k, l = 0$ or 1. If p_1 and p_2 denote the allocation proportions between GC and PC arms, and q_1 and q_2 the prevalence of TS positivity and TS negativity, then, under the stratified randomization scheme, we have $p_{kl} = p_{k+1}q_{l+1}$. We also specify the survival distributions of these four patient groups. For simplicity, we assume exponential survival distributions with hazard rates λ_{kl} for $k, l = 0$, under which model (1) is simplified to

$$\lambda = \lambda_0 \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2)$$

with

$$\lambda_{00} = \lambda_0, \quad \lambda_{10} = \lambda_0 \exp(\beta_1), \quad \lambda_{01} = \lambda_0 \exp(\beta_2), \quad \lambda_{11} = \lambda_0 \exp(\beta_1 + \beta_2 + \beta_3).$$

By solving these equations with respect to $(\lambda_0, \beta_1, \beta_2, \beta_3)$, we have

$$\begin{aligned} \lambda_0 &= \lambda_{00}, & \beta_1 &= \log \lambda_{10} - \log \lambda_{00}, & \beta_2 &= \log \lambda_{01} - \log \lambda_{00}, \\ \beta_3 &= \log \lambda_{11} - \log \lambda_{10} - \log \lambda_{01} + \log \lambda_{00}. \end{aligned} \quad (2)$$

We derive the asymptotic distribution of the Cox regression estimates of (1) using the specified design parameters. For patient $i = 1, \dots, n$, let X_i be the minimum of censoring time and survival time, δ_i be the event indicator taking 1 if an event occurred and 0 otherwise, and $Z_i = (z_{1i}, z_{2i}, z_{1i}z_{2i})^T$ be the covariate vector. Partial score and information functions for regression coefficients $\beta = (\beta_1, \beta_2, \beta_3)^T$ are given as

$$\begin{aligned} U(\beta) &= \sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{\sum_{j=1}^n Y_j(t) Z_j e^{\beta^T Z_j}}{\sum_{j=1}^n Y_j(t) e^{\beta^T Z_j}} \right\} dN_i(t) \\ I(\beta) &= \sum_{i=1}^n \int_0^\infty \left[\frac{\sum_{j=1}^n Y_j(t) Z_j^{\otimes 2} e^{\beta^T Z_j}}{\sum_{j=1}^n Y_j(t) e^{\beta^T Z_j}} - \frac{\left\{ \sum_{j=1}^n Y_j(t) Z_j e^{\beta^T Z_j} \right\}^{\otimes 2}}{\left\{ \sum_{j=1}^n Y_j(t) e^{\beta^T Z_j} \right\}^2} \right] dN_i(t), \end{aligned}$$

respectively, where $N_i(t) = \delta_i I(X_i \leq t)$ is the event process, $Y_i(t) = I(X_i \geq t)$ is the at-risk process, and $z^{\otimes 2} = zz^T$ for a vector z . Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^T$ denote the solution of $U(\beta) = 0$. Then, by Fleming and Harrington (1991), $\hat{\beta}$ is approximately normal with mean 0, and

variance-covariance $F^{-1}(0)$ under the global null hypothesis $\beta = 0$. Hence, a one-sided α test for β_3 rejects $H_0: \beta_3 = 0$ if $\hat{\beta}_3/\hat{\sigma}_3 < -z_{1-\alpha}$, where $\hat{\sigma}_3^2$ is the (3, 3)-component of $F^{-1}(0)$.

For sample size calculation, we calculate the limit of $I(0)$ under H_1 . It is easy to show that, as $n \rightarrow \infty$,

$$\frac{\sum_{j=1}^n Y_j(t) Z_j^{\otimes 2}}{\sum_{j=1}^n Y_j(t)} \rightarrow A_1 = \begin{pmatrix} p_{1\cdot} & p_{11} & p_{11} \\ p_{11} & p_{\cdot 1} & p_{11} \\ p_{11} & p_{11} & p_{11} \end{pmatrix},$$

$$\frac{\left(\sum_{j=1}^n Y_j(t) Z_j\right)^{\otimes 2}}{\left(\sum_{j=1}^n Y_j(t)\right)^2} \rightarrow A_2 = \begin{pmatrix} p_{1\cdot}^2 & p_{1\cdot} p_{\cdot 1} & p_{1\cdot} p_{11} \\ p_{1\cdot} p_{\cdot 1} & p_{\cdot 1}^2 & p_{\cdot 1} p_{11} \\ p_{1\cdot} p_{11} & p_{\cdot 1} p_{11} & p_{11}^2 \end{pmatrix},$$

uniformly, where $p_{1\cdot} = p_{10} + p_{11}$, $p_{\cdot 1} = p_{01} + p_{11}$. Hence, $I(0)$ converges to AD , where $A = A_1 - A_2$ and $D = \sum_{i=1}^n \int_0^\infty dN_i(t)$ is the expected number of events under H_1 .

If the patients are uniformly accrued during accrual period a and followed during time period b after the completion of accrual, then the censoring distribution is $U(b, a + b)$. Assuming the uniform censoring distribution and exponential survival distribution, the probability that a patient has an event during the study period $(0, a + b)$ is calculated by $d = p_{00}d_{00} + p_{10}d_{10} + p_{01}d_{01} + p_{11}d_{11}$, where $d_{kl} = 1 - \exp(-\lambda_{kl}b)\{1 - \exp(-\lambda_{kl}a)\}$ in group $(z_1, z_2) = (k, l)$ for $k, l = 0, 1$. Then, we have $D = nd$. Hence, the asymptotic variance of $\hat{\beta}_3$ is expressed as $\sigma_l^2 = A_{(3,3)}/(nd)$, where $A_{(3,3)}$ is the (3, 3)-component of A^{-1} .

Hence, given sample size n , the power for a local alternative hypothesis $H_1: \beta_3 = \bar{\beta}_3$ is given as

$$1 - \beta = P(\hat{\beta}_3/\hat{\sigma}_3 < -z_{1-\alpha} | \beta_3 = \bar{\beta}_3)$$

By solving this equation with respect to D or n , we obtain the required number of events

$$D = A_{(3,3)} \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\bar{\beta}_3} \right)^2$$

or the required sample size

$$n = \frac{A_{(3,3)}}{d} \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\bar{\beta}_3} \right)^2. \quad (3)$$

Formula (3) requires specification of accrual period a together with α , β , $\bar{\beta}_3$, b , and p_{kl} for $k, l = 0$ or 1 . Oftentimes, however, accrual rate r is given, instead of accrual period a , based on the number of patients treated by the member sites recently or the experience from previous studies on the same patient population. Under uniform accrual during period a , we have $n = a \times r$. Noting that $d = d(a)$ is a function of a , (3) is expressed as

$$a \times r = \frac{A_{(3,3)}}{d(a)} \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\bar{\beta}_3} \right)^2. \quad (4)$$

By solving (4) with respect to a , we obtain the required accrual period, say a^* , and the required sample size $n = a^* r$.

2.2 Example 1

We consider the NSCLC trial that is introduced above. We will randomize patients between the two treatment arms with 1-to-1 fashion, i.e., $p_1 = p_2 = 1/2$. The enrolled patients were expected to be evenly stratified to the TS-negative and TS-positive groups because the median value from a previous study (Sun et al., 2011) was selected as the cutoff value for TS positivity, i.e., $q_1 = q_2 = 1/2$. Hence, we have $p_{00} = p_{10} = p_{01} = p_{11} = 1/4$. The 6-month PFS is expected to be about 35% for GC arm regardless of TS level and for PC arm with TS positivity, and 55% for PC arm with TS negativity. So, the expected hazard rates under the alternative hypothesis are $\lambda_{00} = \lambda_{10} = \lambda_{01} = 2.100$ and $\lambda_{11} = 1.196$. For these hazard rates, we have $\lambda_0 = 2.100$, $\beta_1 = \beta_2 = 0$, and $\beta_3 = -0.563$ from (2). Suppose that about 10 patients are expected to be entered to the study, i.e., an annual accrual rate of $r = 120$. We plan to follow the patients for additional $b = 1$ year after the last patient enters. Then, the one-sided $\alpha = 0.1$ test for $H_0: \beta_3 = 0$ in model (1) requires $n = 345$ patients for a power of $1 - \beta = 0.9$. The expected number of events (i.e., number of patients with a disease progression) at the analysis will be $D = 333$. As an effort to lower the sample size for this Phase II trial, we use a large α level compared to a two-sided $\alpha = 0.05$. We observe an empirical power of 0.897 from 10,000 simulation samples of size $n = 345$ that are generated at the design setting. Sun et al. (2015) recruited 321 patients using overall response as the primary endpoint.

A stratified randomized trial of predictive treatment requires a large sample size for testing on the interaction term. Sample size of a trial for testing the interaction term with 50% of biomarker positivity may be compared to that of a trial with for arm-to-arm comparison with 1-to-3 randomization. Let's consider a randomized trial to compare two treatment arms with $(\alpha, 1 - \beta, r, b) = (0.1, 0.9, 120, 1)$ as above and 6-month PFS of 35% for the control treatment and 55% for the experimental treatment. In this case, Yateman and Skene (1992), we need only $n = 122 (D = 102)$ patients by 1-to-3 randomization and $n = 130 (D = 120)$ patients by 3-to-1 randomization. Note that a trial allocating more patients to the control arm requires a larger sample size than one allocating more patients to the experimental arm because the power of survival analysis depends on the number of events, rather than the number of patients, and the former design expects more unbalanced number of events between two arms than the latter. If TS had been already validated to be a predictive

biomarker of pemetrexed before this trial, then we could have chosen an enrichment trial for TS-negative patients that would require even a much smaller sample size. Refer to Maitournam and Simon (2005) about efficiency comparison between an enrichment design and a stratified randomization design for predictive biomarker in terms of a continuous outcome.

3. A Phase II trial with an imaging biomarker

Prognostic biomarkers provide information on the overall cancer outcome in patients to facilitate cancer diagnosis regardless of selected treatments. In this section, we consider a phase trial with an imaging prognostic biomarker. Chemotherapy B has been a standard regimen for patients with non-bulky stage I and stage II Hodgkin lymphoma. In a previous study on six cycles of B, each patient had an FDG-PET (fluorodeoxyglucose positron-emission tomography) imaging after two cycles of B. It was found that the patients with a negative PET image (group 1) and those with a positive PET image (group 2) had a 3-year PFS of $S_1(3) = 0.86$ and $S_2(3) = 0.52$, respectively, and the hazard ratio, λ_2/λ_1 , was estimated as $\theta_0 = 4.3$.

In a new single-arm Phase II trial, the patients with a negative PET image after two cycles of B will be treated by additional four cycles of the standard chemotherapy B as in the previous study, whereas those with a positive PET image after two cycles of B will be treated by four cycles of a more aggressive chemotherapy C.

Let θ denote the hazard ratio between PET-positive and PET-negative groups. In this trial, we want to show that by treating PET-positive patients with a more aggressive chemotherapy C, their PFS will become closer to that of PET-negative patients who are treated by the standard chemotherapy B. To this end, we test $H_0: \theta = \theta_0$ against $H_1: \theta < \theta_0$. While the PFS of group 2 will be different between H_0 and H_1 , that of group 1 will be identical since PET – group receives the same treatment as that of the previous study.

Let n_k denote the sample size in arm k , $n = n_1 + n_2$ the total sample size, and T_{ki} the survival time for subject i in group k ($1 \leq i \leq n_k$; $k = 1, 2$). We observe (X_{ki}, δ_{ki}) , where X_{ki} is the minimum of T_{ki} and the censoring time and δ_{ki} is an event indicator taking 1 if the subject had an event and 0 otherwise.

For group k , $T_{k1}, \dots, T_{k, n_k}$ are IID with hazard rate $\lambda_k(t)$. Under the proportional hazards assumption, $\theta = \lambda_2(t)/\lambda_1(t)$ denotes the hazard ratio. By Jung et al. (2005), $W(\theta)/\sigma_W(\theta)$ is asymptotically $N(0, 1)$, where

$$W(\Delta) = \int_0^\infty \frac{Y_1(t)Y_2(t)}{Y_1(t) + \Delta Y_2(t)} \{ \Delta d\hat{\Lambda}_1(t) - d\hat{\Lambda}_2(t) \},$$

$$\sigma_n^2(\Delta) = \Delta \int_0^\infty \frac{Y_1(t)Y_2(t)}{\{Y_1(t) + \Delta Y_2(t)\}^2} dN(t),$$

$\hat{\Lambda}_k(t) = \int_0^t Y_k^{-1}(t) dN_k(t)$ is the Aalen–Nelson estimator (Aalen, 1978; Nelson, 1969) for the cumulative hazard function $\Lambda_k(t) = \int_0^t \lambda_k(s) ds$, $Y_k(t) = \sum_{i=1}^{n_k} I(X_{ki} \geq t)$ and $N_k(t) = \sum_{i=1}^{n_k} \delta_{ki} I(X_{ki} \leq t)$ are the at-risk process and the event process for group k , respectively, and $N(t) = N_1(t) + N_2(t)$. Hence, we reject $H_0: \theta = \theta_0$, in favor of $H_1: \theta < \theta_0$, if $W(\theta_0)/\sigma_n(\theta_0) > z_{1-\alpha}$ with one-sided type I error rate α , where $z_{1-\alpha}$ denotes the 100(1 - α) percentile of $N(0, 1)$. The test statistic with $\theta_0 = 1$ is the standard log-rank test (Peto and Peto, 1972).

3.1 Sample Size Calculation

We want to estimate the sample size n under a local alternative hypothesis that $H_1: \theta = \theta_1 (< \theta_0)$ with a desired power. For sample size calculation of this study, we need to specify following design parameters:

- Type I error rate and power, ($\alpha, 1 - \beta$)
- PET negativity and PET positivity, p_1, p_2
- Survival distributions for PET-negative and PET-positive patient groups: exponential distributions with hazard rates λ_1 for PET-negative group; λ_{20} under H_0 and λ_{21} under H_1 for PET-positive group
- Accrual period a and additional follow-up period b

Note that $\theta_0 = \lambda_{20}/\lambda_1$ and $\theta_1 = \lambda_{21}/\lambda_1$. Jung et al. (2005) propose a sample size formula with $\theta_0 > 1$ and $\theta_1 = 1$ for designing noninferiority trials. Jung and Chow (2012) extend their formula for general θ_0 and θ_1 with $\theta_1 < \theta_0$ as follows.

Let $p_k = n_k/n$ denote the prevalence of group k ($k = 1, 2$). By Jung et al. (2005), $\sigma_n^2(\Delta)$, with $\Delta = \lambda_2/\lambda_1$ for an arbitrary λ_2 , is asymptotically equivalent to $n\sigma^2(\Delta)$ under the design setting, where

$$\sigma^2(\Delta) = \Delta p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t) \{p_1 \lambda_1 S_1(t) + p_2 \lambda_2 S_2(t)\}}{\{p_1 S_1(t) + \Delta p_2 S_2(t)\}^2} dt, \quad (5)$$

$S_k(t) = \exp(-\lambda_k t)$ is the survivor function of exponential PFS, and $G(t)$ is the survivor function of the $U(b, a + b)$ censoring distribution

$$G(t) = \begin{cases} 1 & \text{if } t \leq b \\ -t/b + (a+b)/a & \text{if } b < t \leq a+b \\ 0 & \text{if } t > a+b \end{cases}$$

Also, $W(\theta_0) - W(\theta_1)$ is asymptotically equivalent to $n\omega$, where

$$\omega = (\Delta_0 - \Delta_1) p_1 p_2 \int_0^\infty \frac{G(t) S_1(t) S_2(t) \{p_1 f_1(t) + p_2 f_2(t)\}}{\{p_1 S_1(t) + \Delta_0 p_2 S_2(t)\} \{p_1 S_1(t) + \Delta_1 p_2 S_2(t)\}} dt. \quad (6)$$

The integrals in (5) and (6) are calculated using a numerical method. Since we have

$$\frac{W(\Delta_0)}{\sigma_n(\Delta_0)} = \frac{W(\Delta_1)}{\sigma_n(\Delta_1)} \times \frac{\sigma_n(\Delta_1)}{\sigma_n(\Delta_0)} + \frac{W(\Delta_0) - W(\Delta_1)}{\sigma_n(\Delta_0)},$$

from (5) and (6), the test statistic can be approximated by

$$\frac{W(\Delta_1)}{\sigma_1} \times \frac{\sigma_1}{\sigma_0} + \frac{\omega \sqrt{n}}{\sigma_0},$$

where $\sigma_h^2 = \sigma^2(\Delta_h)$ for $h = 0, 1$. Hence, given n , the power of the test statistic with one-sided α is

$$1 - \beta = P\left(\frac{W(\Delta_0)}{\sigma_n(\Delta_0)} > z_{1-\alpha} | H_1\right) \approx P\left(\frac{W(\Delta_1)}{\sigma_1} \times \frac{\sigma_1}{\sigma_0} + \frac{\omega \sqrt{n}}{\sigma_0} > z_{1-\alpha} | H_1\right) = \bar{\Phi}\left(\frac{\sigma_0 z_{1-\alpha}}{\sigma_1} - \frac{\omega \sqrt{n}}{\sigma_1}\right),$$

where $\bar{\Phi}(\cdot)$ is the survivor function of $N(0, 1)$. By solving this equation with respect to n , we obtain the sample size n required for a power of $1 - \beta$ as

$$n = \frac{(\sigma_0 z_{1-\alpha} + \sigma_1 z_{1-\beta})^2}{\omega^2}. \quad (7)$$

The number of events D expected at the analysis time is calculated by $D = n(p_1 d_1 + p_2 d_2)$, where $d_k = 1 + \int_0^\infty S_k(t) dG(t)$.

In this sample size calculation, we assumed that the accrual period was known. In designing a clinical trial, however, we can estimate the accrual pattern, rather than an accrual period. Suppose that patients are expected to be enrolled to the study at a rate of r during an accrual period based on the number of patients treated by the member sites recently. Given $(\lambda_1, 0, 1, \alpha, 1 - \beta, p_1, b)$, $\omega = \omega(a)$ and $\sigma_h = \sigma_h(a)$ for $h = 0, 1$ are functions of a . Under uniform accrual assumption, we have $n = a \times r$. Hence, (7) is expressed as

$$a \times r = \frac{\{\sigma_0(a) z_{1-\alpha} + \sigma_1(a) z_{1-\beta}\}^2}{\omega^2(a)}. \quad (8)$$

By solving (8) with respect to a , using a numerical method such as the bisection method, we obtain the required accrual period a^* and the required sample size $n = a^* \times r$.

3.2 Example 2

We consider the PET-guided Hodgkin lymphoma trial introduced in the beginning of this section. Under H_0 , we assume a 3-year PFS of 86% and 52% for PET-negative and PET-positive groups, respectively, which correspond to annual hazard rates of $(\lambda_1, \lambda_{20}) = (0.050, 0.218)$ under an exponential PFS model, resulting in a hazard ratio of $\lambda_0 = 4.3$. By treating the PET-positive patients with an aggressive treatment C, we expect to increase their 3-year PFS up to 74% (from 52%), resulting in an annual hazard rate of $\lambda_{21} = 0.100$ and hazard rate of $\lambda_1 = 2.0$. The previous study observed about $p_2 = 20\%$ of PET positivity. Assuming an annual accrual rate of $r = 60$ patients and $b = 3$ years of additional follow-up after completion of accrual, we need $n = 191$ patients for $1 - \beta = 90\%$ power for detecting $H_1: \lambda_1 = 2$ by the generalized log-rank test $W(\lambda_0)/\sigma_n(\lambda_0)$ with one-sided $\alpha = 10\%$ and $H_0: \lambda_0 = 4.3$. Under this specific alternative hypothesis, we expect about 46 events (progressions or deaths) at the data analysis.

Simulation studies are conducted to evaluate the calculated sample size under the above design settings of H_0 and H_1 , respectively. Using 10,000 simulation samples of size $n = 132$ under each hypothesis, the empirical type I error rate and power are observed as 0.0984 (to be compared to $\alpha = 0.1$) and 0.8749 (to be compared to $1 - \beta = 0.9$), respectively.

4. Discussion

We have presented design analysis methods of two Phase II trials for biomarker-guided treatments. The method for Section 2 can be used for Phase III trials with appropriate selection of endpoint and design parameter values.

The power of the statistical tests of the trials discussed above depends on the prevalence of the biomarker positivity. So, it is crucial to check the observed prevalence during the patient accrual, and to recalculate the sample size if the observed prevalence is very different from the one specified at sample size calculation. For both of our example trials, the initial sample size will be under-powered if the observed prevalence is farther from 1/2 than the specified one at the design stage. In this case, we often plan for up to 30% of sample size increase based on the sample size recalculation.

For sample size calculation, we have assumed exponential survival distributions and an accrual pattern with a constant accrual rate, but we can easily extend the formulas for any survival distribution and any accrual pattern, refer to Jung (2013). Further, the concept of design and analysis for biomarker-driven Phase II trials with a survival endpoint can be used for those with other type of endpoint, such as binary outcome for tumor response.

As an effort to lower the sample size of a Phase II trial from that of a Phase III trial, we use a high type I error rate, such as one-sided $\alpha = 5\%$ or 10% (compared to two-sided $\alpha = 5\%$), a surrogate endpoint, such as tumor response or PFS (compared to a confirmatory endpoint such as overall survival), a larger treatment effect, and a single-arm design (compared to a

randomized design). In spite of these efforts, we observe that a randomized trial stratified by a predictive biomarker requires a relatively large sample size for a Phase II trial. This is already well known by literature, e.g., Simon and Maitournam (2004), Maitournam and Simon (2005), and Mandrekar and Sargent (2009).

Acknowledgments

Funding

This research was supported by a grant from the National Cancer Institute (CA142538-01).

References

- Aalen OO. Nonparametric inference for a family of counting processes. *Annals of Statistics*. 1978; 6:701–726.
- Cox DR. Regression models and life tables (with discussion). *Journal of Royal Statistical Society B*. 1972; 34:187–220.
- Fleming, TR., Harrington, DP. *Counting Processes and Survival Analysis*. New York: Wiley; 1991.
- Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: Design issues. *The Journal of the National Cancer Institute*. 2010; 102:152–160. [PubMed: 20075367]
- Freidlin B, McShane LM, Polley MYC, Korn EL. Randomized phase II trial designs with biomarkers. *Journal of Clinical Oncology*. 2012; 30:3304–3309. [PubMed: 22869885]
- Jung, SH. *Randomized Phase II Cancer Clinical Trials*. Chapman & Hall/CRC; 2013.
- Jung SH, Chow SC. On sample size calculation for comparing survival curves under general hypotheses testing. *Journal of Biopharmaceutical Statistics*. 2012; 22:485–495. [PubMed: 22416836]
- Jung SH, Kang SJ, McCall L, Blumenstein B. Sample size computation for noninferiority log-rank test. *Journal of Biopharmaceutical Statistics*. 2005; 15:957–967. [PubMed: 16279354]
- Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Statistics in Medicine*. 2005; 24(3):329–339. [PubMed: 15551403]
- Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: Theoretical considerations and practical challenges. *Journal of Clinical Oncology*. 2009; 27(24):4027–4034. [PubMed: 19597023]
- Nelson W. Hazard plotting for incomplete failure data. *Journal of Quality Technology*. 1969; 1:27–52.
- Ozasa H, Oguri T, Uemura T, Miyazaki M, Maeno K, Sato S, Ueda R. Significance of thymidylate synthase for resistance to pemetrexed in lung cancer. *Cancer Science*. 2010; 101:161–166. [PubMed: 19811498]
- Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with discussion). *Journal of the Royal Statistical Society, Series A*. 1972; 135:185–206.
- Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clinical Cancer Research*. 2004; 10(20):6759–6763. [PubMed: 15501951]
- Sun JM, Ahn JS, Jung SH, Sun J, Ha SY, Han J, Park K, Ahn MJ. Pemetrexed plus cisplatin versus gemcitabine plus cisplatin according to thymidylate synthase expression in nonsquamous non-small-cell lung cancer: A biomarker-stratified randomized phase II trial. *Journal of Clinical Oncology*. 2015; 33:2450–2456. [PubMed: 26124486]
- Sun JM, Han J, Ahn JS, Park K, Ahn MJ. Significance of thymidylate synthase and thyroid transcription factor 1 expression in patients with nonsquamous non-small cell lung cancer treated with peme-trexed-based chemotherapy. *Journal of Thoracic Oncology*. 2011; 6:1392–1399. [PubMed: 21716147]
- Yateman NA, Skene AM. Sample size for proportional hazards survival studies with arbitrary patient entry and loss to follow-up distributions. *Statistics in Medicine*. 1992; 11:1103–1113. [PubMed: 1496198]