A Simulation Based Evaluation of Sample Size Methods for Biomarker Studies

Kristen M. Cunanan[1] and Mei-Yin C. Polley[2,*]

[1]Memorial Sloan Kettering Cancer Center, USA

[2]Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA

Abstract: Cancer researchers are often interested in identifying biomarkers that are indicative of poor outcomes (prognostic biomarkers) or response to specific therapies (predictive biomarkers). In designing a biomarker study, the first statistical issue encountered is the sample size requirement for adequate detection of a biomarker effect. In biomarker studies, the desired effect size is typically larger than those targeted in therapeutic trials and the biomarker prevalence is rarely near the optimal 50%. In this article, we review sample size formulas that are routinely used in designing therapeutic trials. We then conduct simulation studies to evaluate the performances of these methods when applied to biomarker studies. In particular, we examine the impact that deviations from certain statistical assumptions (i.e., biomarker positive prevalence and effect size) have on statistical power and type I error. Our simulation results indicate that when the true biomarker prevalence is close to 50%, all methods perform well in terms of power regardless of the magnitude of the targeted biomarker effect. However, when the biomarker positive prevalence rate deviates from 50%, the empirical power based on some existing methods may be substantially different from the nominal power, and this discrepancy becomes more profound for large biomarker effects. The type I error is maintained close to the 5% nominal level in all scenarios we investigate, although there is a slight inflation as the targeted effect size increases. Based on these results, we delineate the range of parameters within which the use of some sample size methods may be sufficiently robust.

Keywords: Sample size methods, biomarker study, prognostic biomarker, predictive biomarker, survival data.

1. INTRODUCTION

Recent advances in biotechnology have led to the development of cancer therapeutics designed specifically to act on molecular targets. Cancer researchers are often interested in identifying molecular markers indicative of poor outcomes or response to specific therapies. These two classes of biomarkers are referred to as *prognostic* and *predictive biomarkers* [1-3]. Prognostic biomarkers are factors that identify patients at an elevated risk of relapse or death. Predictive biomarkers are molecular features that identify a subgroup of patients who are more likely to benefit from a specific treatment regimen. Statistically, the predictive value of a biomarker is exmamined through the test of a biomarker by treatment interaction. A biomarker may be both prognostic and predictive. For example, the *Oncotype DX*, an RT-PCR assay that measures 21 genes whose levels of expression are manipulated by a mathematical algorithm to calculate a recurrence score (RS), provides a prognosis for patients with estrogen receptor (ER)-positive breast cancer with stage I or II breast cancer and negative axillary lymph nodes treated with tamoxifen alone [4]. In one study, the RS also predicts chemotherapy benefit (cyclophosphamide, methotrexate, and fluorouracil or methotrexate and fluorouracil); patients with tumors who had high RS ($\geq$30) experienced a large benefit from the addition of chemotherapy to tamoxifen, whereas those with tumors that had low RS (<18) derived minimal benefit from chemotherapy treatment [5]. The American Society of Clinical Oncology recommended the use of *Oncotype DX* as a prognostic and predictive tool in ER-positive, lymph node-negative breast cancer [6].

Statistical issues in biomarker studies have been addressed by many authors [7-11]. In designing a biomarker study, the first statistical issue encountered is the sample size requirement for adequate detection of a biomarker effect. For prognostic studies with a binary biomarker, many sample size methods developed for therapeutic trials are frequently used [12-15]. The method by Hsieh and Lavori can be used for continuous biomarkers [16]. For predictive biomarker studies, the methods by Peterson and George [17] and Schmoor, Sauerbrei and Schumacher [18] provided close-form formula to calculate the sample size needed to detect a treatment by biomarker interaction. Gönen proposed a unifying framework to compute sample size necessary for an interaction effect that can accommodate normal, binary and time-to-event outcomes [19]. Recently, Polley proposed methods for estimating statistical power in biomarker studies when the clinical events of interest are already observed – a

*Address correspondence to this author at the Department of Health Sciences Research, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA; Tel: 507-538-1370; Fax: 507-266-2477; E-mail: polley.mei-yin@mayo.edu

E-ISSN: 1929-6029/18

typical scenario when biomarker studies are conducted using tissue specimens collected in a previously completed treatment trial [20].

While many sample size methods exist for designing therapeutic trials, their statistical properties have not been studied specifically in the context of biomarker studies. In biomarker studies, the desired effect size is typically larger than those targeted in therapeutic trials. While no definitive rule exists for the effect size, a biomarker that imparts a risk of two fold or greater would be considered clinically relevant. The statistical power of a biomarker study is also influenced by the prevalence of biomarker positivity. Unlike therapeutic trials where the treatment randomization ratio is pre-determined by design, the prevalence of biomarker positivity is typically unknown at the design stage but may be estimated from prior data or inferred from the literature. In most cases, the prevalence of biomarker positivity will not be near the optimal 50%.

In this article, we aim to evaluate the performances of common sample size methods in the context of biomarker studies. We assume that the biomarker assay has demonstrated satisfactory pre-analytical and analytical performances and yields a binary value for the biomarker. In the sections that follow, we first provide a brief overview of sample size methods; this review is not meant to be exhaustive of the literature but rather, is intended to provide a review of sample size methods that are frequently used for designing treatment trials or biomarker studies in practice. We then conduct simulation studies to assess their performances when applied to biomarker studies. In particular, we examine the impact that deviations from certain statistical assumptions (i.e., true biomarker effect size and biomarker positive prevalence) have on statistical power and type I error. Based on these results, we delineate the range of parameters within which the use of some sample size methods may be sufficiently robust and provide general recommendations for sample size methods suitable for prognostic and predictive biomarker studies.

2. REVIEW OF SAMPLE SIZE METHODS

2.1. Prognostic Biomarker Studies

Consider a binary biomarker M which classifies patients into one of two biomarker subgroups: $M-$ (biomarker negative) and $M+$ (biomarker positive), with the prevalence of biomarker positivity being $P(M+)=w$. Suppose the survival times for patients in the biomarker negative and positive subgroups follow an exponential distribution with hazard rates λ_0 and λ_1, respectively. For patient k in the $M-$ subgroup, let t_{0k} be the observed time at risk (i.e. time from entry to death or censoring, whichever occurs first), and let $\delta_{0k}=1$ if death is observed and $\delta_{0k}=0$ otherwise. Define t_{1k} and δ_{1k} analogously for patients in the $M+$ subgroup. Let N_0 and N_1 denote the number of patients in $M-$ and $M+$ subgroups, respectively. The likelihood function involves the parameters of interest (λ_0,λ_1) only through the product

$$\prod_{k=1}^{N_0}\lambda_0^{\delta_{0k}}e^{-\lambda_0 t_{0k}}\prod_{k=1}^{N_1}\lambda_1^{\delta_{1k}}e^{-\lambda_1 t_{1k}}. \quad (1)$$

Let $\Delta=\lambda_0/\lambda_1$ denote the prognostic biomarker effect of interest. In a prognostic study, the goal would be to determine the sample size needed to test $H_0:\Delta=1$ against $H_a:\Delta\neq 1$ at significance level α with power $1-\beta$ against the alternative $\Delta=\Delta_a$ for some pre-specified Δ_a. By applying standard likelihood theory to (1) [14, 21], it can be shown that the asymptotic distribution of the log hazard ratio is

$$\ln\hat{\Delta}\sim N\left(\ln\Delta,\frac{1}{E(d_0)}+\frac{1}{E(d_1)}\right),$$

where $\hat{\Delta}=\hat{\lambda}_0/\hat{\lambda}_1$ denotes the MLE of Δ and $E(d_0)$ and $E(d_1)$ denote the expected number of deaths for the $M-$ and $M+$ subgroups, respectively.

Since $\ln\Delta=0$ under H_0, the asymptotic 2-sided α-level test rejects H_0 in favor of H_1 if the test statistic is greater than or equal to the standard normal $(1-\alpha/2)$ quartile. It then follows that the following condition will need to be satisfied to detect a prognostic effect Δ_a with $(1-\beta)\%$ power, usinga a 2-sided α level test:

$$\frac{(\ln\Delta_a)^2}{(z_{\alpha/2}+z_\beta)^2}=\frac{1}{E(d_0)}+\frac{1}{E(d_1)}, \quad (2)$$

where $Z_{\alpha/2}$ and z_β are the standard normal $\alpha/2$ and β quantiles, respectively. For 1-sided α–level test, one can simply replace $(\alpha/2)$ in (2) with α.

2.1.1. Rubinstein Method

Rubinstein *et al*. derived a closed-form formula for the required sample size in a two-arm randomized treatment trial assuming exponential death times and equal treatment randomization [14]. Their method allows for censoring due to either loss to follow-up or end of study, i.e. administrative censoring. In particular,

by assuming: (i) that patients enter the study uniformly over $[0,a]$, and (ii) an accrual rate of n such that the total sample size *N* follows a Poisson distribution with mean na, they expressed the expected number of deaths as a function the accrual rate n, hazard rates λ_0 and λ_1, accrual time a, follow-up period f, and loss to follow-up rates ϕ_0 and ϕ_1. Using their results and assuming no loss to follow-up, the expected number of deaths in biomarker subgroup i, for $i=0,1,$ can be expressed as

$$E(d_i)=\frac{nw_i}{\lambda_i}\left\{\lambda_i a-e^{-\lambda_i f}\left(1-e^{-\lambda_i a}\right)\right\}, \tag{3}$$

where $w_1=w$ and $w_0=1-w.$ Substituting $E(d_0)$ and $E(d_1)$ in (2) with the expressions in (3) gives

$$\frac{(\ln\Delta_a)^2}{(z_{\alpha/2}+z_\beta)^2}=\sum_{i=0,1}\frac{\lambda_i}{nw_i}\left\{\lambda_i a-e^{-\lambda_i f}\left(1-e^{-\lambda_i a}\right)\right\}^{-1} \tag{4}$$

One can then solve for the required accrual rate n (or the required accrual time a if n is fixed) in (4) with other factors being fixed. The total sample size required for a prognostic biomarker study is thus $N=n\times a.$.

2.1.2. Schoenfeld Method

Schoenfeld derived the required sample size for the Cox proportional hazards model, obtaining the same formula as that for the two-sample log-rank test under the proportional hazards assumption [12, 13]. For a two-sided significance level α, power $(1-\beta)$ and prognostic effect of interest Δ_a the total number of events required is

$$D=\frac{(z_{\alpha/2}+z_\beta)^2}{w(1-w)(\ln\Delta_a)^2} \tag{5}$$

In order to calculate the actual number of patients that are required for the prognostic biomarker study, one needs to consider the probability of death π over the duration of the study, which can be expressed as the weighted probability of death in the two biomarker subgroups, i.e. $\pi=(1-w)\pi_0+w\pi_1$. Assume that patients enter the study at a constant rate during an accrual period a, followed by a follow-up period f, one can use Simpson's rule [22] to approximate the proportion of patients who will die in biomarker subgroup i as

$$\pi_i=1-\frac{1}{6}\left(\hat{S}_i(f)+4\hat{S}_i(f+0.5a)+\hat{S}_i(f+a)\right), \tag{6}$$

where $S_i(\cdot)$ is the survival distribution for patients in biomarker subgroup i, which can be estimates by assuming a parametric survival distribution such as exponential. Finally, the total number of patients required for the prognostic biomarker study is simply $N=D/\pi$ using equations (5) and (6).

2.2. Predictive Biomarker Studies

The sample size methods for prognostic biomarker studies reviewed above can be extended to allow for the calculation of sample sizes for predictive biomarker studies. Let w denote the prevalence of a biomarker positive status and p denote the probability of a patient being assigned to the experimental arm. Assume that death times for the four treatment-by-biomarker groups arise from exponential distributions with hazard rates λ_{ij}, for $i=0,1,$ and $j=0,1,$ as represented in Table **1**. Let $\Delta_0=\lambda_{10}/\lambda_{00}$ and $\Delta_1=\lambda_{11}/\lambda_{01}$ denote the hazard ratios for the treatment effects in $M-$ and $M+$ subgroups, respectively. Statistically, the predictive ability of the biomarker can be quantified by the interaction term $\Delta^*=\Delta_1/\Delta_0,$ the ratio of the treatment hazard ratios in the two biomarker subgroups. In a predictive marker study, the goal would be to determine the sample size needed to test $H_0:\Delta^*=1$ against $H_a:\Delta^*\neq 1$ at significance level α with power $1-\beta$ against the alternative $\Delta^*=\Delta_b$ for some pre-specified Δ_b.

Based on standard likelihood theory [14, 22], one can show that the asymptotic distribution of $\ln\hat{\Delta}^*$ is

$$\ln\hat{\Delta}^*\sim N\left(\ln\Delta^*,\sum_{i=0,1}\sum_{j=0,1}\frac{1}{E(d_{ij})}\right),$$

where $\ln\hat{\Delta}^*=\ln\left(\frac{\hat{\lambda}_{00}\hat{\lambda}_{11}}{\hat{\lambda}_{01}\hat{\lambda}_{10}}\right)$ is the MLE of $\ln\Delta^*$, and $E(d_{ij})$ denotes the expected number of death in the (i,j) cell. It then follows that the following condition will need to be satisfied to detect an interaction effect Δ_b with $(1-\beta)\%$ power, using a 2-sided α level test:

$$\frac{(\ln\Delta_b)^2}{(z_{\alpha/2}+z_\beta)^2}=\sum_{i=0,1}\sum_{j=0,1}\frac{1}{E(d_{ij})}. \tag{7}$$

Table 1: The 2 x 2 Design Parameter Table in a Predictive Biomarker Study

	Biomarker Status	
	Negative (1 - ω)	**Positive (ω)**
Control (1-p)	λ_{00}	λ_{01}
Experimental (p)	λ_{10}	λ_{11}
	$\Delta_0 = \lambda_{10} / \lambda_{00}$	$\Delta_1 = \lambda_{11} / \lambda_{01}$
	$\Delta^* = \Delta_1 / \Delta_0$	

2.2.1. Peterson and George Method

Peterson and George modified Equation (A2) from the Appendix of Rubinstein et al. and provided a closed-form a closed-form *al.* and provided a closed-form formula for the expected number of events $E(d_{ij})$ [17]. Assuming no loss to follow-up, this number is a function of the accrual rate (n), the accrual time (a), the follow-up period (f) and the cell-specific hazard rate (λ_{ij}):

$$E(d_{ij}) = (nap_i w_j)\left\{1 - \frac{e^{-\lambda_{ij} f}\left(1 - e^{-\lambda_{ij} a}\right)}{a\lambda_{ij}}\right\} \quad (8)$$

for $i, j = 0,1$, where $p_1 = p = (1 - p_0)$ an $w_1 = w = (1 - w_0)$. Note that the leading term $(nap_i w_j)$ in (8) represents the expected number of patients in the (i, j) cell. Substitutin $E(d_{ij})$ in (7) with the expression in (8) gives

$$\frac{(\ln \Delta_b)^2}{(z_{\alpha/2} + z_\beta)^2} = \sum_{i=0,1} \sum_{j=0,1} \frac{\lambda_i}{np_i w_i}\left\{\lambda_{ij} a - e^{-\lambda_{ij} f}\left(1 - e^{-\lambda_{ij} a}\right)\right\}^{-1}. \quad (9)$$

One can then solve for the required accrual rate n (or the required accrual time a if n is fixed) in (9) with all other factors fixed. The total sample size required is then $N = n \times a$.

2.2.2. Schmoor Method

Schmoor and colleagues derived an approximate sample size formula for detecting an interaction effect for the case of exponential failure times [18]. For a two-sided level α test with $(1 - \beta)$ power, the required number of events is

$$D = \frac{(z_{\alpha/2} + z_\beta)^2}{(\ln \Delta_b)^2}\left[\frac{1}{(1-p)(1-w)} + \frac{1}{(1-p)w} + \frac{1}{p(1-w)} \frac{1}{pw}\right]$$

The total number of patients required for the predictive biomarker study is $N = D / \pi$, where π denotes the overall probability of death over the duration of the study which could be calculated as the weighted probability of death in the four cells, i.e. $\pi = (1-p)(1-w)\pi_{00} + (1-p)w\pi_{01} + p(1-w)\pi_{10} + pw\pi_{11}$. Again, using Simpson's rule, the cell-specific probability of death can be estimated as

$$\pi_{ij} = 1 - \frac{1}{6}(\hat{S}_{ij}(f) + 4\hat{S}_{ij}(f + 0.5a) + \hat{S}_{ij}(f + a)), \quad (10)$$

where $S_{ij}(\cdot)$ is the survival distribution for patients in treatment group i and biomarker subgroup j.

2.2.3. Factor of 16 (FO16) Method

Peterson and George also presented a simple formula (Equation (3) in their paper) for determining the sample size for interaction effects [17]:

$$N = \frac{16(z_{\alpha/2} + z_\beta)^2}{\pi(\ln \Delta_b)^2}, \quad (11)$$

where π is the overall death probability that can be calculated as the sum of the weighted probabilities of deaths in the four cells as in (10). This equation was obtained by equating the expected number of events in (7) to the observed number of events (i.e. $E(d_{ij}) = d_{ij}$) and assuming that $d_{ij} = d$ all i, j. Note that the Schmoor formula also reduces to the same formula when $p = w = 0.5$ is assumed.

We refer to (11) as the "Factor of 16" (FO16) formula because of the leading constant 16 in the formula. This formula is often used by practitioners to devise the sample size for the interaction effect primarily due to its simplicity. Insofar as when this simplified formula may be adequate for practical use, Peterson and George provided that *"in practice, we recommend that sample size calculations be done without assuming an equal number of failures per cell for designs where the hypothesized ratio of maximum to minimum number of cell-specific deaths is greater than 2 or where the ratio of maximum to minimum cell-specific hazard rates is greater than 3."* While these general rules of thumbs are useful guides in selecting between the more complex Peterson and George formula and the Factor of 16 formula, the design of a predictive biomarker study typically involve direct specifications of the biomarker positive prevalence and the hypothesized treatment by biomarker interaction effect. In this article, we assess directly how the statistical power and type I error based on these formula may be impacted by these parameters.

3. SIMULATION STUDIES

3.1. Prognostic Biomarker Studies

We conduct a simulation study to compare the performances of the two common sample size methods presented for prognostic biomarker studies [12-14]. We assume an accrual period of $a = 24$ months and a follow-up period of $f = 12$ months. The median survival for the biomarker positive cohort is assumed to be 15 months, corresponding to a monthly hazard rate of $\lambda_1 = 0.046$ assuming exponential death times. Suppose patients in the biomarker negative cohort confer a worse prognosis such that $\Delta_a = \lambda_0 / \lambda_1 > 1$. We vary the

prognostic effect size Δ_a between 1.2 and 3 by increments of 0.01 and the biomarker positive prevalence *w* between 0.1 to 0.9 by increments of 0.1. For each (Δ_a, w) configuration, we first calculate the sample size required to achieve 80% power using a 2-sided $\alpha = 0.05$ level test based on the two methods. For each method, we then simulate 5000 datasets with the same sample size as determined by the formula. Note the Rubinstein formula was derived with the assumption of exponential event times and a test statistic that is based on the ratio of the maximum likelihood estimates of the event hazard rates. They showed (via simulations) that their method is approximately valid for the non-parametric log-rank test, which is nearly efficient for hazard ratios between 1/2 and 2 [14]. To ensure fair power comparisons, we use the log-rank test for comparing the prognosis between the two biomarker subgroups in both methods. Specifically, in each simulated dataset, the 2-sided log-rank test with $\alpha = 0.05$ is applied to assess statistical significance of the prognostic biomarker effect. For each method and parameter combination, the empirical power is calculated as the percentage of simulated studies that reach statistical significance for the targeted prognostic effect. Using the same sample size, we also simulate 5000 datasets under the null hypothesis (i.e. $\lambda_0 = \lambda_1 = 0.046$). The empirical type I error is then calculated as the percentage of simulated studies that reject the null based on a 2-sided 0.05 level log-rank test.

Rubinstein *et al*. demonstrated that even when the underlying distribution of the event times is not exponential, their method is valid for the log-rank statistic and retains desirable power against alternatives with constant hazard ratios for a wide range of Weibull distributed event times [14]. Here we further determine whether our observed trends regarding the approximate power of the methods with exponential death times may be extended to situations when event times are not exponential. To that end, we carried out additional simulation studies assuming that event times follow a Weibull distribution. Specifically, we generated event times from the cumulative density function $F_+(t) = 1 - \exp(-\gamma t^b)$ for the biomarker positive subgroup and $F_-(t) = 1 - \exp(-\gamma \Delta_a t^b)$ for the biomarker negative subgroup, where γ is the scale parameter, b is the shape parameter and Δ_a is the constant hazard ratio. The hazard decreases over time for $0 < b < 1$ and increases over time for $b > 1$. For a given value of b, we solve for the scale parameter γ in the biomarker positive subgroup so that the median survival in that group is 15 months to be consistent with the simulations in the exponential settings. All other design parameters are fixed to be the same as those in the exponential cases.

3.2. Predictive Biomarker Studies

We conduct a simulation study to compare the performances of the three sample size methods presented for predictive biomarker studies [17, 18]. We assume an accrual period of $a = 9$ years and a follow up period of $f = 9$ years. The median survival for the patients who are in the control arm and biomarker negative is assumed to be 3 years, corresponding to a annual hazard rate of $\lambda_{00} = 0.23$ with exponential death times. To simplify, we further assume that within the control arm, patients whose biomarker status is positive have the same prognosis as those whose biomarker status is negative; that is, the biomarker is *not* prognostic so that $\lambda_{00} = \lambda_{01} = 0.23.$ Suppose that the treatment effect among patients with biomarker negative status is represented by $\Delta_0 = \lambda_{00} / \lambda_{10} = 1.2.$ We vary the treatment effect among patients with biomarker positive status $\Delta_1 = \lambda_{01} / \lambda_{11}$ between 1.4 and 3.6 by increments of 0.01, so that the interaction effec $\Delta_b = \Delta_1 / \Delta_0$ varies between 1.17 and 3. We consider the biomarker positivity prevalence *w* to be between 0.1 and 0.9 with increments of 0.1. Then for each (Δ_b, w) configuration, we first calculate the sample size required to achieve a 80% power using a 2-sided $\alpha = 0.05$ level test based on the three methods. For each competing method, we then simulate 5000 datasets each with the same sample size as determined by the formula. In each simulated dataset, we fit a Cox proportional hazards (PH) model with a treatment main effect, a biomarker main effect, and an interaction term between treatment and biomarker status. The significance of the interaction effect is then tested using a likelihood ratio test. In situations where the Cox PH model does not converge, the simulated dataset is excluded from the power evaluation. For all methods and parameter combinations considered, the maximum number of excluded datasets is only 10 (or 0.2% of the 5,000 simulated datasets). The empirical power is calculated as the percentage of simulated studies that reach statistical significance for the targeted interaction effect. Using the same sample size, we also simulate 5000 datasets under the null hypothesis (i.e. $\lambda_{00} = \lambda_{01} = 0.23, \Delta_0 = \Delta_1 = 1.2$). The empirical type I error is calculated as the percentage of simulated studies for which the test of interaction reaches statistical significance using a 2-sided 0.05 level likelihood ratio test.

We also conducted additional simulation studies to examine whether our observed trends for the empirical power based on the three sample size methods for predictive biomarker studies may be extended to outcome data that are not exponential. Specifically, we generated event times for patients who are in the control arm and are biomarker negative from the

cumulative density function $F_{00}(t) = 1 - \exp(-\gamma t^b)$, where γ and b represent the scale and shape parameters for a Weibull distribution, respectively. For a fixed value of b, we solve for γ so that the median survival in this patient subgroup is 3 years, to be consistent with the simulations in the exponential settings. As in the exponential cases, we further assume the biomarker is not prognostic such that $F_{00}(t) = F_{01}(t)$ for a given b. Assuming proportional hazards for the treatment effect within each biomarker subgroup, we then generated event time data for patients in the experimental arm from $F_{10}(t) = 1 - \exp(-\gamma\Delta_0 t^b)$ for patients who are biomarker negative and from $F_{11}(t) = 1 - \exp(-\gamma\Delta_1 t^b)$ for patients who are biomarker positive. All other design parameters are fixed to be the same as those in the exponential cases.

4. RESULTS

4.1. Prognostic Biomarker Studies

Figure **1** presents simulated power curves based on the Schoenfeld (left plot) and Rubinstein (right plot) methods, respectively. The x-axis represents varying prognostic effect sizes. Each curve within a plot represents a specified prevalence rate for biomarker positivity. Figure **1** suggests that when the biomarker positive prevalence is 50%, the empirical power based on both methods is maintained at the 80% nominal level even for large targeted hazard ratios. However, as the biomarker positive prevalence deviates from 50%, the empirical power starts to deviate from the 80% nominal level.

For the Schoenfeld method (see Figure **1**, left plot), the empirical power is lower than the expected power when $w < 50\%$ but is higher than the nominal level when $w > 50\%$. It is interesting to note the fan shape of the power curves and their symmetric nature around the targeted 80% power. The degree of deviation in power increases as the targeted prognostic effect increases. For example, when the biomarker positive prevalence is 10%, this method yields 76% and 70% power to detect hazard ratios of $\Delta_a = 1.5$ and $\Delta_a = 3$, respectively. When the biomarker positive prevalence is 90%, this method yields 86% and 91% power to detect hazard ratios of $\Delta_a = 1.5$ and $\Delta_a = 3$, respectively. Within the context of our simulation studies, the power based on the Schoenfeld method varies between 70% and 92%.

Power curves based on the Rubinstein *et al*. method are presented in Figure **1** (right plot). To aid visualization, here we present only three power curves corresponding to biomarker positive prevalence $w = 0.1, 0.5, 0.9$, due to the proximity of the curves. Power based on the Rubinstein *et al*. method appears to be more robust to changes in w and Δ_a than the Schoenfeld method. In our simulation exercise, the empirical power based on the Rubinstein method varies between 78% and 86%. In general, as the biomarker positive prevalence deviates from 50%, the empirical power based on the Rubinstein *et al*. method appears slightly higher than the expected power specified in the formula and this power overage is magnified as the targeted hazard ratio increases.

Table **2** presents the empirical power for prognostic biomarker effect assuming Weibull death times. To conserve space, here we only present results based on four values of $b = \{1/4, 1/2, 3, 5\}$ and two values of $\Delta_a = \{2, 3\}$. For each (b, Δ_a) combination, we consider three biomarker positive prevalence rates $w = \{0.2, 0.5, 0.8\}$. Overall, two methods perform

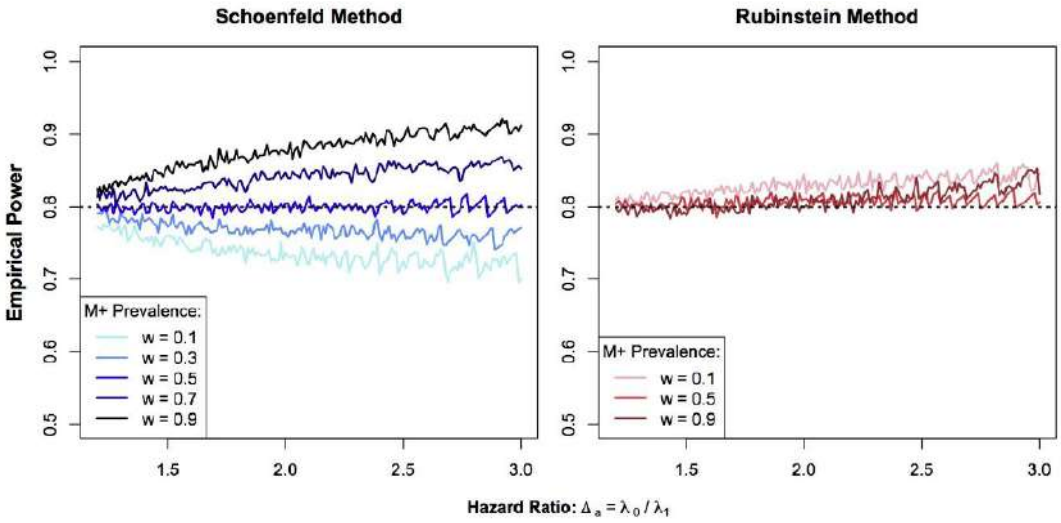

Figure 1: Empirical power for prognostic biomarker studies based on two sample size methods (5000 simulations), for varying biomarker positive (M+) prevalence rates.

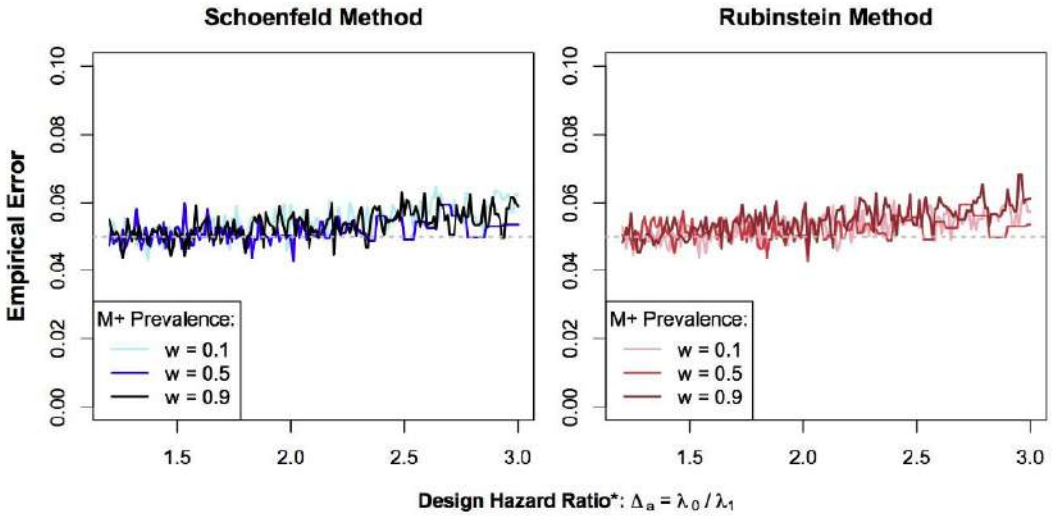

Figure 2: Empirical type I error for prognostic biomarker studies based on two sample size methods for varying biomarker positive (M+) prevalence rates. The x-axis is the targeted prognostic effect size used to calculate the sample size. Based on this sample size, 5000 datasets are then simulated under the null hypothesis ($\lambda_0 = \lambda_1$).

Table 2: Empirical Power for Prognostic Biomarker Effect Assuming Weibull Death Times Based on 5000 Simulations. Δ_a Represents the Prognostic Effect. *b* Defines the Shape Parameter for the Weibull Distribution. ω Represents the Biomarker Positive Prevalence Rate

b	Method	$\Delta_a = 2$			$\Delta_a = 3$		
		ω = 0.2	ω = 0.5	ω = 0.8	ω = 0.2	ω = 0.5	ω = 0.8
1/4	Schoenfeld	0.68	0.75	0.82	0.71	0.77	0.86
	Rubinstein	0.76	0.75	0.75	0.78	0.78	0.79
1/2	Schoenfeld	0.71	0.77	0.84	0.73	0.78	0.87
	Rubinstein	0.78	0.77	0.78	0.81	0.79	0.81
3	Schoenfeld	0.79	0.84	0.90	0.77	0.81	0.90
	Rubinstein	0.86	0.84	0.85	0.85	0.83	0.84
5	Schoenfeld	0.79	0.84	0.90	0.77	0.81	0.91
	Rubinstein	0.86	0.85	0.86	0.85	0.83	0.84

similarly when the biomarker positive prevalence is 50% as in the exponential cases. When $w \leq 50\%$, the Schoenfeld method produces power consistently lower than the Rubinstein *et al*. method. In contrast, the Rubinstein method gives lower power than the Schoenfeld method when $w > 0.5$.

Figure **2** presents simulated type I error with exponential event times based on the two sample size methods. In general, the type I error is maintained close to the nominal 5% level, although a slight inflation is noted as the targeted prognostic effect increases. Across all scenarios we investigate, the empirical type I error ranges from 0.042 to 0.065 for the Schoenfeld method and from 0.042 to 0.068 for the Rubinstein method, respectively. With Weibull event times, the type I error ranges from 0.047 to 0.065 for the Schoenfeld method and from 0.046 to 0.061 for the Rubinstein method, respectively. In summary, the type I error rate is quite robust against the violations of statistical assumptions for both methods.

4.2. Predictive Biomarker Studies

The three top plots in Figure **3** present the simulated power curves based on the three sample size methods with various biomarker positive prevalence rates. The x-axis represents varying degrees of the interaction effect. The top left plot in Figure **3** suggests as the biomarker positive prevalence deviates from 50%, the empirical power based on the FO16 method could be substantially lower than the 80% nominal level. The top middle plot in Figure **3** indicates when the biomarker positive prevalence is less than 50%, the empirical power based on the Schmoor method appears to be slightly lower than the expected power. This power loss increases with an increase in the targeted interaction effect size, although

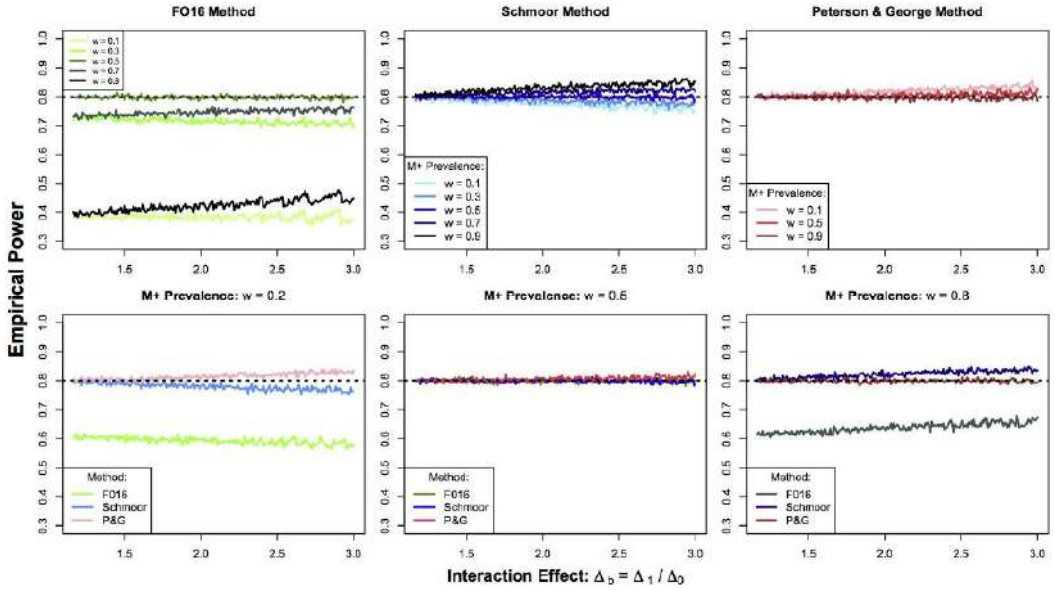

Figure 3: Empirical power for predictive biomarker studies based on three sample size methods (5000 simulations). The top three plots display, for each method, the power curves for varying biomarker positive (M+) prevalence rates. The bottom three plots compare power curves of different sample size methods for three different biomarker positive (M+) prevalence rates.

the magnitude of power loss is less severe than that of the FO16 method. As the biomarker positive prevalence increases, there appears to be a symmetric pattern of power overage around the 80% nominal level. The Peterson and George method appears to be most robust to changes in w and Δ_b in that the power curves track the nominal 80% level most closely, see the top right plot in Figure **3**.

The three bottom plots in Figure **3** display comparisons of power curves based on the three competing methods for 20% (low), 50% (medium) and 80% (high) biomarker positive prevalence rates. When the biomarker positive prevalence is near the 50% optimal level, all three methods perform well in terms of power for an interaction effect size less than 3. When the biomarker positive prevalence is low, both the FO16 and Schmoor formulas produce empirical power lower than expected. The power loss is more drastic for the FO16 method. The Peterson and George method has empirical power slightly higher than the nominal level. Finally, when biomarker positive prevalence is high, the FO16 method produces power that is lower than the nominal level while the Schmoor formula yields power that is higher than expected. The Peterson and George method again tracks the nominal level closely.

Table **3** presents results for four values of $b = \{1/2, 2/3, 5/4, 3/2\}$ and two values of $\Delta_b = \{2, 3\}$. For each (b, Δ_b) combination, we consider three biomarker positive prevalence rates $w = \{0.2, 0.5, 0.8\}$. Again, we note that the three methods perform similarly when the biomarker positive prevalence is 50%. The FO16 method produces empirical power lower than the 80% nominal level when the biomarker positive prevalence deviates from 50% as previously observed for the exponential cases. For $w < 50\%$, the Schmoor method has lower power than the Peterson and George method but this trend is reversed when $w > 50\%$.

Figure **4** presents simulated type I error with exponential event times based on the three competing methods. In general, the type I error is maintained close to the nominal 5% level, although there is a small inflation as the targeted interaction effect increases and this inflation appears to be slightly more profound for the FO16 method. The ranges for the empirical type I error across all scenarios we examine are (0.042, 0.068), (0.042, 0.06), and (0.04, 0.062) for the FO16 method, Schmoor method, and the Peterson and George method, respectively. With Weibull event times, the ranges are (0.047, 0.063), (0.047, 0.061), and (0.048, 0.062) for the three corresponding methods. Overall, the type I error rate is quite robust against the violations of statistical assumptions for all three methods.

5. DISCUSSION

The calculation of sample size in therapeutic trials usually assumes an equal treatment randomization and a modest treatment effect size. These assumptions are often unrealistic in correlative science studies involving the investigation of the prognostic or predictive value of a biomarker. Specifically, the prevalence of a positive biomarker is rarely near 50% and the desired biomarker effect is often larger than the conventionally targeted treatment effect size to be considered clinically meaningful. In this work, we investigate via

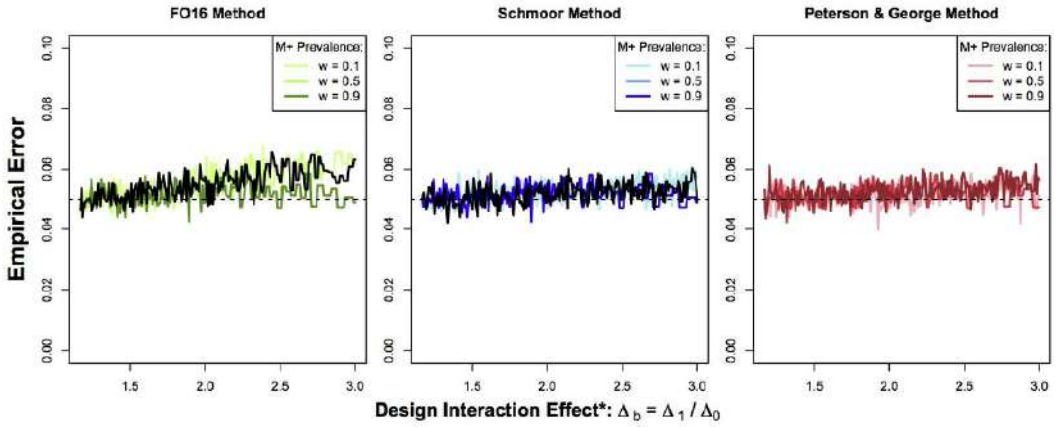

Figure 4: Empirical type I error for predictive biomarker studies based on three sample size methods for varying biomarker positive (M+) prevalence rates. The x-axis is the targeted interaction effect used to calculate the sample size. Based on this sample size, 5000 datasets are then simulated under the null hypothesis ($\Delta_0 = \Delta_1$).

Table 3: Empirical power for predictive biomarker effect assuming Weibull death times based on 5000 simulations. Δ_b represents the interaction effect. *b* defines the shape parameter for the Weibull distribution. ω represents the biomarker positive prevalence rate.

b	Method	$\Delta_b = 2$			$\Delta_b = 3$		
		ω = 0.2	ω = 0.5	ω = 0.8	ω = 0.2	ω = 0.5	ω = 0.8
1/2	Factor of 16	0.48	0.69	0.52	0.46	0.69	0.53
	Schmoor	0.68	0.69	0.70	0.61	0.67	0.72
	Peterson & George	0.69	0.70	0.70	0.69	0.69	0.67
2/3	Factor of 16	0.52	0.73	0.56	0.50	0.73	0.58
	Schmoor	0.70	0.73	0.75	0.68	0.72	0.76
	Peterson & George	0.74	0.74	0.74	0.75	0.73	0.74
5/4	Factor of 16	0.63	0.83	0.65	0.63	0.84	0.67
	Schmoor	0.80	0.82	0.84	0.80	0.83	0.85
	Peterson & George	0.83	0.82	0.81	0.87	0.83	0.82
3/2	Factor of 16	0.63	0.84	0.67	0.65	0.84	0.70
	Schmoor	0.82	0.84	0.84	0.82	0.83	0.86
	Peterson & George	0.83	0.83	0.83	0.87	0.85	0.83

simulation studies the robustness of some popular sample size methods in the context of biomarker studies. While the empirical power of the methods is jointly influenced by the targeted biomarker effect size and the prevalence of the biomarker, we found the impact of the latter is much greater. In particular, when the biomarker prevalence is close to 50%, we observed all methods perform well in terms of power regardless of the magnitude of the biomarker effect, implying the study power would be maintained at the desired level using any sample size method. However, when there is a large imbalance in biomarker prevalence, the empirical power may be profoundly different from the stated asymptotic power especially for large targeted biomarker effects. Specifically, when the biomarker prevalence is very high or very low, the FO16 method (for predictive studies) produce power substantially lower than the nominal level regardless of the effect size and hence should not be used. The Schoenfeld method (for prognostic studies) and the Schmoor method (for predictive studies) perform reasonably well so long as the desired prognostic or predictive effect size is modest, e.g. Δ_a or $\Delta_b < 2$. The Rubinstein method (for prognostic studies) and the Peterson and George method (for predictive studies) outperform other methods in their respective class. This is not necessarily surprising given the fact that these methods imposes the fewest statistical assumptions in

their derivations. The type I error is maintained close to the 5% nominal level in all scenarios we investigate, although there is a slight inflation as the targeted biomarker effect size increases.

The simulation results presented in this article are limited. The performance of the methods under investigation may vary with different design parameters. To evaluate this further, we conduct additional simulation studies. For prognostic cases, we consider an additional scenario with longer accrual and follow-up time as well as longer median survival for the biomarker positive cohort. For predictive cases, we consider an additional scenario with shorter accrual and follow-up time and a longer median survival for patients who are in the control arm with biomarker negative status. In both cases, we note that the performances of the competing methods in each class are similar to those presented in this article (data not shown). Similarly, we have only considered exponential distribution and Weibull distribution as the true data generating distributions in our simulation studies. It is worth noting that the Schoenfeld method was derived for the Cox proportional hazards model. Hence our simulations based on the Weibull distribution represents a scenario in which exponential assumption is violated whereas the proportional hazards condition holds true. In reality, the true data generating distribution is not known, and future work may focus on comparison of methods under other data generating distributions. Our observations in this article may only be generalized to clinical scenarios where exponential or Weibull distributions could be reasonably assumed.

In summary, we provide a systematic simulation-based evaluation to elucidate the adequacy of using some existing sample size methods for biomarker studies. The results of this paper demonstrate that some existing methods that may be suitable for estimating the sample size in treatment trials may yield power substantially different from what is expected in biomarker studies. An immediate implication of these results is that basing sample size estimation on some existing methods may result in a futile attempt to discover a potentially important biomarker due to the lack of statistical power even if a true prognostic or predictive effect exists. On the other hand, some methods may overestimate the sample size needed to achieve the desired power, resulting in an unnecessary waste of precious tissue specimens. It is our hope that these simulation results can serve as a useful guide to biomarker researchers when devising the sample size in their studies. The R functions to implement the sample size methods in this article are available at https://github.com/kristenmay206/BiomarkerStudySS

ACKNOWLEDGEMENT

The senior author (MYP) would like to acknowledge the support of research funding from the 2017 Fifth District Eagles through the Mayo Clinic Cancer Center. The authors are thankful to Dr. Larry Rubinstein at the National Cancer Institute and Dr. Eric Polley at Mayo Clinic for their helpful suggestions and comments.

REFERENCES

[1] Henry LN, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. The Oncologist 2006; 11(6): 541-552. https://doi.org/10.1634/theoncologist.11-6-541

[2] Polley M-YC, Freidlin B, Korn EL, Conley BA, Abrams JS, Mc-Shane LM. Statistical and practical considerations for clinical evaluation of predictive biomarkers. Journal of the National Cancer Institute 2013; 105(22): 1677-1683. https://doi.org/10.1093/jnci/djt282

[3] Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. Journal of the National Cancer Institute 2009; 101(21): 1446-1452. https://doi.org/10.1093/jnci/djp335

[4] Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, LBaehner F, Walker MG, Watson D, Park T, Hiller W, Fisher ER, Wickerham L, Bryant J, Wolmark N. A multigene assay to predict recurrence of tamoxifentreated, node-negative breast cancer. New England Journal of Medicine 2004; 10(351): 2817-2826. https://doi.org/10.1056/NEJMoa041588

[5] Paik S, Tang G, Shak S, Kim C, Baker J, Kim W, Cronin M, Baehner FL, Watson D, Bryant J, Costantino JP, Geyer Jr, CE, Wickerham DL, Wolmark N. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. Journal of Clinical Oncology 2006; 24(23): 3726-3734. https://doi.org/10.1200/JCO.2005.04.7985

[6] Harris L, Fritsche H, Mennel R, Norton L, Ravdin P, Taube S, Somerfield MR, Hayes DF, Bast RC Jr. American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. Journal of Clinical Oncology 2007; 25(33): 5287-5312. https://doi.org/10.1200/JCO.2007.14.2364

[7] Taube SE, Clark GM, Dancey JE, McShane LM, Sigman CC, Gutman SI. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. Journal of the National Cancer Institute 2009; 101(21): 1453-1463. https://doi.org/10.1093/jnci/djp334

[8] Freidlin B, McShane LM, Korn EL. Randomized clinical trials with biomarkers: Design issues. Journal of the National Cancer Institute 2010; 102(3): 152-160. https://doi.org/10.1093/jnci/djp477

[9] Subramanian J, Simon R. Gene expression{based prognostic signatures in lung cancer: Ready for clinical use? Journal of the National Cancer Institute 2010; 102(7): 464-474. https://doi.org/10.1093/jnci/djq025

[10] Sargent D, Mandrekar S. Statistical issues in the validation of prognostic, predictive, and surrogate biomarkers. Clinical Trials 2013; 10(5): 647 -652. https://doi.org/10.1177/1740774513497125

[11] Polley M-YC, Polley EC, Huang EP, Freidlin B, Simon R. Two-stage adaptive cutoff design for building and validating a prognostic biomarker signature. Statistics in Medicine 2014; 33(29): 5097-5110. https://doi.org/10.1002/sim.6310

[12] Schoenfeld DA. Sample-size formula for the proportional-hazards regression model. Biometrics 1983; 39(2): 499-503. https://doi.org/10.2307/2531021

[13] Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika 1981; 68(1): 316-319. https://doi.org/10.1093/biomet/68.1.316

[14] Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. Journal of Chronic Diseases 1981; 34(9): 469-479. https://doi.org/10.1016/0021-9681(81)90007-2

[15] George SL Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. Journal of Chronic Diseases 1974; 27(1): 15-24. https://doi.org/10.1016/0021-9681(74)90004-6

[16] Hsieh FY, Lavori PW. Sample-size calculations for the cox proportional hazards regression model with nonbinary covariates. Controlled Clinical Trials 2000; 21(6): 552-560. https://doi.org/10.1016/S0197-2456(00)00104-5

[17] Peterson B, George SL. Sample size requirements and length of study for testing interaction in a 2 xK factorial design when time-to-failure is the outcome. Controlled Clinical Trials 1993; 14(6): 511-522. https://doi.org/10.1016/0197-2456(93)90031-8

[18] Schmoor C, Sauerbrei W, Schumacher M. Sample size considerations for the evaluation of prognostic factors in survival analysis. Statistics in Medicine 2000; 19(4): 441-452. https://doi.org/10.1002/(SICI)1097-0258(20000229)19:4<441::AID-SIM349>3.0.CO;2-N

[19] Gonen M. Planning for subgroup analysis: a case study of treatment-marker interaction in metastatic colorectal cancer. Controlled Clinical Trials 2003; 24(4): 355-363. https://doi.org/10.1016/S0197-2456(03)00006-0

[20] Polley M-Y. Power estimation in biomarker studies where events are already observed. Clinical Trials 2017; 14(6): 621-628. https://doi.org/10.1177/1740774517723830

[21] Cook TD, DeMets DL. Introduction to statistical methods for clinical trials.Chapman & Hall/CRC, 2013.

[22] Collett D. Modelling survival data in medical research. Boca Raton, Fla.: Chapman & Hall/CRC, third ed., 2014.

Received on 27-07-2018 Accepted on 09-08-2018 Published on 25-10-2018

https://doi.org/10.6000/1929-6029.2018.07.04.1