Sample Size Calculation for Studies with Grouped Survival Data

Zhiguo Li*, Xiaofei Wang, Yuan Wu, and Kouros Owzar

Department of Biostatistics and Bioinformatics, Duke University, North Carolina, USA

Abstract

Grouped survival data arise often in studies where the disease status is assessed at regular visits to clinic. The time to the event of interest can only be determined to be between two adjacent visits or is right censored at one visit. In data analysis, replacing the survival time with the endpoint or midpoint of the grouping interval leads to biased estimators of the effect size in group comparisons. Prentice and Gloeckler developed a maximum likelihood estimator for the proportional hazards model with grouped survival data and the method has been widely applied. Previous work on sample size calculation for designing studies with grouped data is either based on the exponential distribution assumption or approximation of variance under the alternative with variance under the null. Motivated by studies in HIV trials, cancer trials as well as in vitro experiments to study drug toxicity, we develop a sample size formula for studies with grouped survival endpoints that use Prentice and Gloeckler's method for comparing two arms under the proportional hazards assumption. We do not impose any distributional assumptions, nor do we use any approximation of variance of the test statistic. The sample size formula only requires estimates of the hazard ratio and survival probabilities of the event time of interest and the censoring time at the endpoints of the grouping intervals for one of the two arms. The formula is shown to perform well in a simulation study and its application is illustrated in the three motivating examples.

1 Introduction

In studies with survival outcomes, it is desirable to observe the exact survival times for all subjects. In practice, however, some kind of censoring is inevitable, with right censoring being the most common, which is due to the finite study period as well as early drop-out of study subjects [1]. While in right censoring, the survival times of some of the subjects are exactly observed, in some other studies the survival time can never be determined to the desired precision, which results in interval censored data [2]. Interval censored data usually arise when the event status, e.g., HIV infection or disease progression, is assessed at regular visits to the clinic. When the visit times are predetermined (nonrandom), then the resulting data are called grouped survival data (see, e.g., [3-4]). Grouped survival data are common in HIV studies where subjects are tested for HIV infection at regular visits to the clinic. The time to infection is determined to be between the first visit at which the test is positive and its previous visit. Similar data are encountered in cancer clinical trials where the time to disease progression (TTP) is the endpoint of interest. TTP is the time from treatment initiation to disease progression, in which drop-outs or deaths before progression are considered as censoring. In this kind of studies, regular assessments of disease progression are performed, and the TTP can only be observed to be between two adjacent assessment times. A further example comes from the study of drug toxicity where the LD50 (the dosage at which 50% of cells are killed) of a drug is of interest. In the experiments, only a finite number of doses are tested, and thus the LD50 of the drug can only be determined to be between two adjacent doses. See [5] for a specific example. Here LD50 is not a “survival time” itself, but the data can be analyzed using techniques for grouped survival data.

Prentice and Gloeckler [3] discussed previously existing methods for analyzing grouped survival data. Those methods are either computationally infeasible or give inconsistent estimators of the hazard ratio in a proportional hazards model. This motivated them to develop the maximum likelihood method for analyzing grouped data with a proportional hazards model. Although the proportional hazards model is a semiparametric model, the likelihood function of the grouped survival data only depends on finite number of parameters and the Newton-Raphson method can be used to solve for the score equation. This analysis method has been widely used in practice. As to the design of clinical trials with grouped survival data, Lui [6] and Lui et al. [7] derived sample size formulae for cohort studies with grouped data in the special case of exponential distribution. Inoue and Parmigiani [8] and Raab et al. [9] discussed the design of such trials but focused on the design of follow-up intervals based on parametric models for the survival time instead of sample size determination for group comparisons. Lachin [10] derived a sample size formula for grouped survival data based on the score test of the log hahard ratio [3]. In deriving the formula, the variance of the score statistic under the alternative is approximated by the variance under the null (see equation (34) in [10]).

We are mainly motivated by clinical trials to study time to HIV infection and cancer clinical trials where the time to disease progression is the primary outcome of interest. As the first example, [11] reported results of a randomized placebo-controlled phase III trial to test the efficacy of a preventive HIV-1 vaccine, and found that, compared with placebo, the vaccine did not prevent HIV-1 acquisition. In this trial, subjects were administered vaccine or placebo at months 0, 1, 6, 12, 18, 24, and 30, with a final follow-up visit at month 36. At each visit, blood sample was obtained to assess HIV-1 status. Flynn et al. used endpoints of intervals to approximate time to HIV infection and then used partial likelihood to estimate the hazard ratio in a proportional hazards model. They found similar infection rates in the two arms and the difference was not significant. As we will show in simulation, the commonly used method of approximating the grouped survival time with an endpoint or the midpoint of the corresponding interval and then using methods for right censored data results in biased estimation (also see [12]). Desirably, the consistent and efficient method in [3] should be used for data analysis. If a similar potential clinical trial is planned in which Prentice and Gloeckler's method is to be used for analysis, then the question arises as to how the necessary sample size for such a trial can be determined.

For the second example, consider cancer clinical trials in which TTP is the primary endpoint for evaluating treatment effect. There are many such trials in practice [13-15]. Here we are specifically motivated by CALGB 30607, a phase III trial studying the effect of Sunitinib, a multi-targeted receptor tyrosine kinase inhibitor, versus placebo in treating advanced non-small cell lung cancer [16]. Computerized tomography (CT) scan was used to determine whether a patient has progressed after treatment initiation and was scheduled every 6 weeks for all patients until progression. The progression-free survival (PFS), which is the time to progression or death (whichever comes first), was used as the primary endpoint in the original trial. However, PFS is a composite endpoint that includes both TTP and time to death. It is either exactly observed (for death) or is observed to be in an interval (for TTP), and thus only part of the data are grouped. Generally, one consideration for using PFS as the endpoint is to have more events in the analysis (compared to TTP), but at the same time it entails problems such as validity of the proportional hazards model and difficulty in interpretation of this model for such type of data [17]. Therefore, it is of interest to use TTP as the primary endpoint to assess treatment effect. In this case, grouped data arise and the traditional sample size formulae for survival analysis do not apply.

We derive a sample size formula for clinical trials with grouped survival data, based on the Wald test for the logarithm of the hazard ratio in a proportional hazards model and Prentice and Gloeckler's method for data analysis. The key is the determination of the formula for the asymptotic variance of the estimator for the log hazard ratio, which is the inverse of the efficient information of the log hazard ratio in the presence of nuisance parameters related to the baseline hazard function. While the Wald test is asymptotically equivalent to the score test (as well as the likelihood ratio test), we do not rely on any approximation of the variances, and thus our sample size formula is different from that in [10]. We show that the difference between the sample size using approximation of variances and the sample size without using such an approximation can be quite significant (10% to 20%). We conduct simulation to assess the performance of our sample size formula. It is shown that, while the sample size using approximation of variances may yield coverage rates far from the nominal level, our sample size formula without using approximation of variances performs much better in these cases. We also show by simulation that the common practice of replacing the grouped survival times with endpoints (or midpoints) of the corresponding intervals leads to biased estimation. An additional contribution is to show by simulation that, the power gain of the test with more frequent measurements becomes insignificant quite rapidly, thus a high frequency of measurements, i.e., fine intervals, may not be necessary. In practice an "optimal" frequency can be decided by taking into account the tradeoff between the efficiency of the inference and the cost of measurements. Finally, compared with the sample size calculation in trials with right censored data, we only need estimates of the baseline survival distribution at the finite number of visit times instead of the whole survival curve, which is due to the grouped nature of the data.

The following content is arranged as follows. Notation is introduced and the sample size formula is derived in Section 2. The performance of the sample size formula is assessed via simulation in Section 3, and Section 4 illustrates the wide applicability of the sample size formula in three distinct areas, including an HIV trial, a trial in lung cancer, and an in vitro experiment in studying LD50 of paclitaxel, a drug for cancer treatment. We then conclude with a discussion in Section 5. Details of the derivation of the sample size formula is put in an Appendix.

2 Derivation of the Sample Size Formula

Suppose the survival time, denoted by T, is observed to be in one of r intervals, denoted by $A_i = [a_{i-1}, a_i)$, for $i = 1, \cdots, r$ with $a_0 = 0$ and $a_r = \infty$. Suppose that n subjects are randomized to one of two treatment arms. Let Z to be the indicator for the two arms, where $Z = 1$ stands for the experimental treatment arm and $Z = 0$ stands for the placebo (or standard treatment) arm. Denote $p_z = P(Z = z)$ for $z = 0,\ 1$, to be the randomization probabilities. Suppose the survival time of a subject falls into the Kth interval or is right censored at time a_{K-1} (early drop-out), where K can take values $1, \cdots, r$. Let Δ be the event indicator which takes value 0 if the subject is right censored and 1 otherwise. Note that T being observed in the interval $[a_{r-1}, a_r)$ is equivalent to T being right censored at a_{r-1} and thus $\Delta = 0$. The observed data consist of (K, Δ, Z). Suppose that, given Z, T follows a proportional hazards model

$$\lambda(t|Z) = \lambda_0(t)e^{Z\beta},$$

where $\lambda(t|Z)$ is the hazard function of T given Z, $\lambda_0(t)$ is the baseline hazard function and β is the log hazard ratio which is the parameter of interest. We will derive the sample size formula for testing for $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$, based on the asymptotic distribution of the maximum likelihood estimator (MLE) for β, where usually $\beta_0 = 0$. Prentice and Gloeckler [3] gave the likelihood function of the observed data conditional on covariate Z in the case that $\Delta = 1$ holds whenever $K < r$, i.e., T is not right censored before time a_{r-1} (see display (2) therein). However, in practice this may not be true due to early drop-out and finite study period. Thus we extend this by allowing T to be right censored at one of $a_1, \cdots, a_{r-2}$. Let C be the censoring time with survival function $S_C^z(t)$ conditional on $Z = z$. Here we allow different censoring distributions in the two arms. Suppose that C is independent of T given Z. Denote

$$\alpha_j = e^{-\int_{a_{j-1}}^{a_j} \lambda_0(t)dt}, \text{ for } 1 \leq j \leq r-1.$$

Then the likelihood function of the observed data conditional on Z is given by

$$\begin{aligned} P(\Delta = 1, K = k|Z) &= P(a_{k-1} < T \leq a_k|Z)P(C > a_k|Z) \\ &= \left(1 - \alpha_k^{e^{Z\beta}}\right) \prod_{j=1}^{k-1} \alpha_j^{e^{Z\beta}} S_C^Z(a_k), \text{ for } 1 \leq k \leq r-1, \end{aligned} \qquad (1)$$

and

$$\begin{aligned} P(\Delta = 0, K = k|Z) &= P(a_{k-1} < T|Z)P(a_{k-1} \le C < a_k|Z) \\ &= \prod_{j=1}^{k-1} \alpha_j^{e^{Z\beta}} \{S_C^Z(a_{k-1}) - S_C^Z(a_k)\}, \text{ for } 1 \le k \le r, \end{aligned} \tag{2}$$

where $\Delta = 0$ and $K = k$ indicate that the subject has survived beyond a_{k-1} but the censoring time is between a_{k-1} and a_k, and the product $\Pi_{j=1}^0$ is defined as 1. The full likelihood is the above likelihood multiplied by the distribution of Z, which can be omitted because it does not depend on unknown parameters. Because of the grouped nature of the data, the likelihood function depends on the baseline hazard function only through finite number of parameters $\gamma_j = \log(-\log \alpha_j)$, where γ_j is a reparameterization of α_j to achieve better asymptotic approximation of the distribution of its MLE. Similarly, the likelihood function depends on the distribution of the censoring time only through $S_C^Z(a_j)$, for $1 \le j \le r-1$. Let $\hat{\beta}$ be the maximum likelihood estimator for β. Then by the theory of MLE, under some regularity conditions, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 = \sigma^2(\beta, \nu)$ is a function of β and ν, ν being the vector of all the γ_js and $S_C^z(a_j)$s, for $z = 0,\ 1$ and $\le j \le r-1$. The necessary regularity conditions in this case include that the α_js are bounded away from 0 and that β is bounded above. By this result, the rejection region

$$\frac{|\hat{\beta} - \beta_0|}{\hat{\sigma}(\beta_0, \hat{\nu})/\sqrt{n}} > Z_{1-\alpha/2},$$

has an (asymptotic) significance level α, where $\hat{\sigma}^2(\beta_0, \hat{\nu})$ is a consistent estimator for $\sigma^2(\beta_0, \nu)$ and a standard estimator for this is the one based on the observed information (details are omitted). Under $H_1 : \beta = \beta_1$, the power of this test is

$$\begin{aligned}
P\left\{\frac{|\hat{\beta}-\beta_0|}{\hat{\sigma}(\beta_0,\hat{\nu})/\sqrt{n}} > Z_{1-\alpha/2}\right\} &\approx P\left\{\frac{|\hat{\beta}-\beta_0|}{\sigma(\beta_0,\nu)/\sqrt{n}} > Z_{1-\alpha/2}\right\} \\
&= P\left\{\frac{\hat{\beta}-\beta_1}{\sigma(\beta_0,\nu)/\sqrt{n}} + \frac{\beta_1-\beta_0}{\sigma(\beta_0,\nu)/\sqrt{n}} > Z_{1-\alpha/2}\right\} + P\left\{\frac{\hat{\beta}-\beta_1}{\sigma(\beta_0,\nu)/\sqrt{n}} + \frac{\beta_1 -}{\sigma(\beta_0,\nu}\right. \\
&\approx P\left\{\frac{\hat{\beta}-\beta_1}{\sigma(\beta_0,\nu)/\sqrt{n}} > Z_{1-\alpha/2} - \frac{\beta_1-\beta_0}{\sigma(\beta_0,\nu)/\sqrt{n}}\right\} \\
&= P\left\{\frac{\hat{\beta}-\beta_1}{\sigma(\beta_1,\nu)/\sqrt{n}} > \frac{\sigma(\beta_0,\nu)}{\sigma(\beta_1,\nu)}\left\{Z_{1-\alpha/2} - \frac{\beta_1-\beta_0}{\sigma(\beta_0,\nu)/\sqrt{n}}\right\}\right\},
\end{aligned}$$

where we assumed $\beta_1 > \beta_0$ without loss of generality. Since $(\hat{\beta}-\beta_1)/\sigma(\beta_1,\nu)/\sqrt{n} \sim N(0,1)$ approximately under H_1, for a power $1-\eta$, we need

$$\frac{\sigma(\beta_0,\nu)}{\sigma(\beta_1,\nu)}\left\{Z_{1-\alpha/2} - \frac{\beta_1-\beta_0}{\sigma(\beta_0,\nu)/\sqrt{n}}\right\} = -Z_{1-\eta}.$$

Solving this for n, we obtain

$$n = \frac{\left\{Z_{1-\alpha/2}\sigma(\beta_0,\nu) + Z_{1-\eta}\sigma(\beta_1,\nu)\right\}^2}{(\beta_1-\beta_0)^2}.$$

Note that this sample size formula is different from the one in [10] because [10] uses the approximation $\sigma(\beta_1,\nu) \approx \sigma(\beta_0,\nu)$ (in display (34)) which results in a sample size

$$n_1 = \frac{\left(Z_{1-\alpha/2} + Z_{1-\eta}\right)^2 \sigma^2(\beta_0,\nu)}{(\beta_1-\beta_0)^2}.$$

In Figure 1 we show that there are all kinds of possibilities for the shape of the curve $\sigma^2(\beta,\nu)$ for fixed ν. Thus the approximation of variance may or may not work well, and it may result in conservative as well as anti-conservative estimates of variances. This will be further illustrated in the simulation study. Finally, the above sample sizes are based on a two-sided test with significance level α. If the test is one-sided (with the same significance level), then the $1-\alpha/2$ need to be replaced by $1-\alpha$ in the above formulae.

Next we need to derive an explicit expression for σ^2. Following [3], we denote $h_j = h_j(Z) = e^{\gamma_j + Z\beta}$, $b_j = b_j(Z) = h_j e^{-h_j}/(1-e^{-h_j})$ and $d_j = d_j(Z) = b_j(e^{-h_j} + h_j - 1)/(1-e^{-h_j})$, for $1 \le j \le r-1$. Since the distribution of the censoring distribution does not depend on $\theta = (\gamma^T, \beta)^T$, the inference on θ can be based on the following log likelihood function

$$\begin{aligned} l(\theta|K,\Delta,Z) &= \log\left\{\left(1-\alpha_K^{e^{Z\beta}}\right)^{\Delta}\prod_{j=1}^{K-1}\alpha_j^{e^{Z\beta}}\right\} \\ &= \Delta\log\left(1-e^{-e^{\gamma_K+\beta Z}}\right) - \sum_{j=1}^{K-1} e^{\gamma_j+\beta Z}. \end{aligned}$$

The information matrix is

$$I(\theta) = \begin{pmatrix} S_{\gamma\gamma} & S_{\gamma\beta} \\ S_{\gamma\beta}^T & S_{\beta\beta} \end{pmatrix},$$

where $S_{\beta\beta} = -E\partial^2 l/\partial\beta^2$, $S_{\gamma\gamma} = -\mathrm{diag}\left(E\partial^2 l/\partial\gamma_1^2, \cdots, E\partial^2 l/\partial\gamma_{r-1}^2\right)$, and $S_{\gamma\beta} = -E(\partial^2 l/\partial\beta\partial\gamma_1, \cdots, \partial^2 l/\partial\beta\partial\gamma_{r-1})^T$. By a well-known formula for block matrix inversion, the inverse of the information matrix can be written as

$$I^{-1}(\theta) = \begin{pmatrix} S_{\gamma\gamma}^{-1} + BA^{-1}B^T & -BA^{-1} \\ -A^{-1}B^T & A^{-1} \end{pmatrix},$$

where $B = S_{\gamma\gamma}^{-1}S_{\gamma\beta}$, and

$$\begin{aligned} A &= -E\frac{\partial^2 l}{\partial\beta^2} + \sum_{i=1}^{r-1}\left(E\frac{\partial^2 l}{\partial\gamma_i^2}\right)^{-1}\left(E\frac{\partial^2 l}{\partial\beta\partial\gamma_i}\right)^2 \\ &= -E\frac{\partial^2 l}{\partial\beta^2} + \sum_{i=1}^{r-1}\left(E\frac{\partial^2 l}{\partial\gamma_i^2}\right)^{-1}\left\{E\left(Z\frac{\partial^2 l}{\partial\gamma_i^2}\right)\right\}^2 \\ &= E\left\{Z\left(\Delta d_K + \sum_{i=1}^{K-1} h_i\right)\right\} - \sum_{i=1}^{r-1}\frac{\left[E\left\{Zh_i I(i<K) + Z\Delta d_i I(K=i)\right\}\right]^2}{E\left\{h_i I(i<K) + \Delta d_i I(K=i)\right\}}. \end{aligned}$$

Note that in the above we used the following results:

$$-\frac{\partial^2 l}{\partial\beta^2} = \left\{Z^2\left(\Delta d_K + \sum_{i=1}^{K-1} h_i\right)\right\},\ -\frac{\partial^2 l}{\partial\beta\partial\gamma_i} = Zh_i I(i<K) + Z\Delta d_i I(K=i),$$

and

$$-\frac{\partial^2 l}{\partial \gamma_i^2} = h_i I(i < K) + \Delta d_i I(K = i),$$

which were derived in [3]. The asymptotic variance of $\hat{\beta}$, in the presence of the nuisance parameter γ, is the right bottom component of I^{-1} which is A^{-1}. Denote $p(\delta, k|z) = P(\Delta = \delta, K = k|z)$, for $\delta = 0,\ 1$, $z = 0,\ 1$, and $1 \leq k \leq r$. It is shown in the Appendix that $\sigma^{-2} = A = \sigma_1 - \sigma_2$, where

$$\sigma_1 \quad = \quad \sum_{z=0}^{1} z \left[\sum_{k=1}^{r-1} d_k p(1, k|z) + \sum_{k=2}^{r} \left(\sum_{i=1}^{k-1} h_i \right) \{p(1, k|z) + p(0, k|z)\} \right] p_z,$$

and

$$\sigma_2 \quad = \quad \sum_{i=1}^{r-1} \frac{\left[\sum_{z=0}^{1} \left\{ z h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|z) + z d_i p(1, i|z) \right\} p_z \right]^2}{\sum_{z=0}^{1} \left\{ h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|z) + d_i p(1, i|z) \right\} p_z}.$$

By the above formula, we can determine the sample size if we have estimates of the effect size β, the distribution function of the survival time in the control group ($Z = 0$) and the distribution of the censoring time, both at the endpoints of the intervals. The R code for sample size calculation is available upon request. In practice, prior estimates of the effect size and the distribution at the endpoints of intervals may be obtained from preliminary data or from clinicians' experience. One practical issue is that different clinicians' estimates of the model parameters may be different and thus the resulting sample sizes are different. In the literature, different methods have been used to deal with this type of situations, including sequential, Bayesian and Maximin (see, e.g., [18]). For example, we can adopt the Maximin idea for our case in the following way. The sample size calculated from our formula based on each set of model parameters is the minimum sample size needed to guarantee the power. According to the Maximin idea, when there are multiple estimates of the set of model parameters, we can take the maximum of the sample sizes calculated from these estimates as our sample size to be used. This sample size may or may not be conservative but is more robust than the one based on a single estimate of the set of model parameters. An alternative is to average over the multiple estimates from different clinicians and calculate the sample size based on the average estimate of the parameters, and the resulting sample size is less likely to be conservative. As to the distribution of the censoring time, it usually contains two components that can be assessed differently. The first component is the so called administrative censoring. This happens because subjects enter the trial at different times but follow-up ends at the same time, i.e., at the end of study. Suppose that the trial stops to enroll subjects at time a after it begins, and all subjects are followed up for an additional amount of time b. Then we may assume that the administrative censoring time is uniformly distributed in the interval $(b-a,\ b)$. The second component is the early drop-out of subjects due to various reasons. We can impose a distribution for this censoring time, too. Denote the administrative censoring time as C_1 and the censoring time due to early drop-out to be C_2. Then the censoring time for a subject is $\min(C_1, C_2)$. It is reasonable to assume that C_1 and C_2 are independent because of the different mechanisms of censoring. Then it is easy to obtain the distribution of C under assumptions on C_1 and C_2. Note that only distributions of these times at the endpoints of the grouping intervals matter in all calculations. For example, we only need to know the proportion of subjects who drop out of the study in each of the intervals.

3 Simulation

In this section, we first show by simulation that if the grouped survival time is replaced by the endpoint of the interval and the partial likelihood is used for estimation, the estimator for the (log) hazard ratio is biased. We then use simulation to assess the performance of our sample size formula. We also examine the change of necessary sample size (power) with the frequency of visits (examining times), for a fixed follow-up time.

To evaluate the bias of the naive method which replaces the grouped survival time with the endpoint of the interval, we first generate the survival times from an exponential distribution or a Weibull distribution. In the exponential distribution case, the baseline hazard rate is 0.03 and the hazard ratio is either 2 or 4. The randomization probability is 0.3 and 0.7 for the experimental arm and the control arm, respectively. In the Weibull distribution case, we set the shape parameter and scale parameter to be 5 and 15, respectively, for the baseline distribution, and the hazard ratio can also take value 2 or 4. The interval [0, 30] is divided into a finite number of intervals with equal lengths, and the survival time is grouped in one of the intervals or is right censored at 30 (grouped in interval $(30, \infty)$). The number of intervals in [0,30] is chosen to be 3, 6, or 10. The Efron method [1] is used to deal with ties. In this simulation, we do not allow right censoring before the last visit time (30). To exclude biases due to small sample sizes, we choose a large sample size which is 1000. Under each scenario, we calculate the empirical bias and percent bias ($|\text{bias}/(\text{true value})| \times 100\%$) of the log hazard ratio of the naive method and the Prentice and Gloeckler method using 1000 simulation replicates. The results are shown in Table 1. These results indicate that the naive method yields significant biases ($> 10\%$ percent) in all cases and it can sometimes yield very large bias (up to 96% percent), while the bias of the Prentice and Gloeckler method is very small.

Now we assess the performance of the sample size formula by comparing the empirical power under sample sizes determined by the formula with the expected power in different scenarios. First, we assume the visit times are 0, 6, 12, 18, 24, and 30. The true underlying distributions of the data are either exponential or Weibull. In the exponential distribution case, the baseline hazard rate is 0.03, and in the Weibull distribution case the shape parameter is 1.5 and the scale parameter for the control group is 20. The hazard ratio takes values 1.3, 1.5, 1.7 or 2.0. In this simulation we assume either there is no right censoring before the last visit time, or the censoring time C has a point mass 1/3 at 30 and is otherwise uniformly distributed in the interval $(0, 30)$. Finally, the randomization probability is set to be 0.5 for both arms. We calculate the sample size using the formula when the expected power is 0.8 or 0.9, respectively, and with a type I error rate 0.05. Then we use 1000 simulation replicates under the calculated sample size to approximate the actual power and compare it with the expected power. Results in Table 2 show that the sample size calculated from the formula ranges from 129 (for a hazard ratio 2.0 and an expected power 0.8) to 962 (for a hazard ratio 1.3 and an expected power 0.9) when there is no right censoring before the last visit time, and it ranges from 169 to 1271 when there is right censoring before the last visit time. The sample size formula performs well, with the actual powers being close to the expected ones. We also assess the robustness of our sample size formula when the proportional hazards assumption is violated. In this simulation, we use the same hazard function for the control group which is exponential with hazard rate 0.03, and for the experimental group we use either a Weibull distribution or a mixture of two exponential distributions. This makes the proportional hazards assumption invalid, but the distributions are chosen such that the median of the experimental group is the same as the case in the original simulation where the two distributions are both exponential. We only consider a hazard ratio of 1.5 in this simulation. Figure 1 shows four cases in which the proportional hazards assumption is violated but the empirical power is between 0.76 and 0.85 with an expected power 0.8. These cases give a rough idea about to what extent the proportional hazards assumption may be violated while still retaining a reasonable power compared to the expected.

Table 3 lists the sample sizes calculated from our sample size formula (n) and the sample sizes obtained by using approximation of the variance (n_1) in 6 scenarios. In each of the scenarios we first choose α and β and then generate interval data using multinomial distributions with probabilities as given by (1) and (2). In all the cases there is no censoring before the last interval. For example, in the first case, $r = 4$ and we set $\alpha = (0.10, 0.18, 0.39)^T$ and $\beta = 0.4$, and in the 4th case $r = 6$, $\alpha = (0.12, 0.14, 0.77, 0.21, 0.14)^T$, and $\beta = 0.3$. In these cases, the differences between n and n_1 are significant. Also, the sample size n_1 can be seriously anti-conservative or conservative, while n results in much more reasonable powers in all these cases. This is consistent with Figure 2 which shows that the approximation of variance may not work well.

We then look at the change of sample size with the number of intervals, i.e., the frequency of visits. The last visit (measurement) time is still set to be 30, and we consider cases where the number of visit times is 3, 6, 10, or 15, and the times between two adjacent visits are the same. The same baseline distributions as above are used here, and we consider hazard ratios 1.3 and 1.5. We also consider two separate cases where there is no right censoring before the last visit time (30), or the censoring time C has a mass 1/3 at 30 and is otherwise uniformly distributed in $(0, 30)$. The randomization probability is still 0.5. The necessary sample sizes for an expected power 0.80, 0.85, and 0.90, with 0.05 type I error rate, are listed in Table 3. We can see that the sample size decreases when the number of intervals increases, i.e., the power of detecting a treatment effect increases with more frequent measurements. However, it is also observed that with the increase of the frequency of measurements the decrease of the sample size is not dramatic. Especially, when there is no right censoring before 30, the decrease of the sample size is almost negligible when the number of intervals is greater than 6. When there is unform right censoring before 30, the sample size decreases more but it is still not dramatic.

4 Examples

In this section, we illustrate the application of the sample size formula using three examples, motivated by an HIV study, a trial in lung cancer, and in vitro experiments for drug toxicity, respectively, which are briefly described in the Introduction.

4.1 HIV Trial

Flynn et al. [11] reported that the recombinant HIV-1 envelope glycoprotein subunit (rgp120) vaccine did not prevent HIV-1 infection, based on results of a double-blind and randomized trial in which 3391 subjects received vaccine and 1704 subjects received placebo. In this trial, vaccine or placebo was administered at months 0, 1, 6, 12, 18, 24, and 30, with a final follow-up visit at month 36. At each visit, blood was obtained for assessment of HIV-1 status. The infection date was estimated as the date of the earliest sample with detectable HIV-1 RNA. The partial likelihood method was used to estimate the hazard ratio in a proportional hazards model comparing the two arms. This potentially biased method for analysis can be avoided in the presence of the rigorous and efficient method of Prentice and Gloeckler. We illustrate the sample size calculation in a potential clinical trial assessing the efficacy of such a vaccine that uses the analysis method of Prentice and Glockler.

First, we use results in [11] as preliminary data for sample size calculation. By [11], the overall log hazard ratio comparing the vaccine group with the placebo group, which was estimated to be -0.06, is too small to be interesting. Thus we choose to focus on the high risk subjects, which were defined as subjects with a risk score greater than or equal to 4, where the risk score was the total number of risk factors reported from 9 categories. In this high risk subgroup, the log hazard ratio was estimated to be -0.56. In order to use the sample size formula, we need estimates of the α_js, which can be obtained from estimates of the survival function for the control (placebo) group at the visit times. Based on Figure 5(F) in [11], the estimates of the survival probabilities of the placebo group are approximately 1, 0.75, 0.63, 0.54, 0.44, 0.25, and 0.18, at month 1, 6, 12, 18, 24, 30, and 36, respectively. These values correspond to $\alpha = (1, 0.75, 0.84, 0.86, 0.81, 0.57, 0.72)^T$. Assuming the randomization probability is 0.5 for both arms, plugging these values into the sample size formula yields a sample size 143 to achieve 0.8 power with 0.05 significance level. If the desired power is 0.9, then the necessary sample size becomes 191. If, as in [11], the randomization probability is 2/3 for the vaccine group, then the necessary sample size is 154 and 206, for a desired power 0.8 and 0.9, respectively. In the above calculations, we assume that there is no drop-out of subjects during the study. If we assume 15% lost to follow-up before the last visit time (36 month), which is the case in the trial in [11], and if we assume the 15% lost to follow-up is uniformly distributed in $[0, 36)$, then the sample sizes needed are 232 and 310, for a power of 0.8 and 0.9, with equal randomization, and these numbers become 218 and 291, if the randomization probability is 2/3 for the vaccine group.

4.2 Lung Cancer Trial

Grouped survival data may arise in advanced-stage cancer trials where TTP is the primary endpoint and the tumor progression is determined by computerized tomography (CT) at fixed time points after treatment. We will consider CALGB 30607, a phase III trial studying the effect of Sunitinib, a multi-targeted receptor tyrosine kinase inhibitor, versus placebo, in treating advanced non-small cell lung cancer [16]. CT scan was scheduled every 6 weeks for all patients until progression. In this study, the primary endpoint was progression-free survival (PFS), which is a composite endpoint that includes time to progression and time to death. Unlike TTP, PFS is exactly observed for patients who died in the study period. Here for illustration purposes, we suppose that in a future trial the primary endpoint is TTP, which is never exactly observed thus leading to grouped survival data. We can calculate the sample size for this trial using preliminary data from CALGB 30607.

In CALGB 30607, the visit schedule was every 6 weeks after treatment until week 254. We first obtain estimates of survival probabilities of the control arm at the visit times. Using the estimation method in [3], the estimates are 0.96, 0.68, 0.49, 0.32, 0.29, 0.21, 0.15, 0.13, 0.06, 0.06, 0.06, 0.04, 0.04, 0.03, 0.03, 0.03, 0.02, at weeks 6, 12, 18, $\cdots$, 96, respectively, 0.01 at weeks 102 to 144, and 0 afterwards. Based on these probabilities, we combine the intervals for weeks 54, 60 and 66, weeks 72 and 78, weeks 84, 90 and 96, weeks 108 to 144, and discard the intervals after week 144. The estimates of survival probabilities for the control arm at the endpoints of the new intervals are 0.96, 0.68, 0.49, 0.32, 0.29, 0.21, 0.15, 0.13, 0.06, 0.04, 0.03, 0.02, and 0.01. Note that now the last visit time is week 144. Based on the same data, the hazard ratio comparing the treatment arm with the control arm is estimated to be 0.64. Since this is a trial in late-stage lung cancer patients, it is expected that all the patients will follow the scheduled visits and therefore we assume that there is no early drop-out of patients. Based on these calculations and our formula, if the randomization probability is 0.5 for both arms, then the necessary sample size for 0.8 power with 0.05 significance level is 168. This is much smaller than the sample size used in CALGB30607, which was 210. Finally, we observe that the survival probabilities after 50 weeks is very small ($< 6\%$), thus more visits after 50 weeks may not contribute much to the power of the comparison due to rare events. Suppose that we set the last visit time to be week 54, then the corresponding sample size is 182, which is not a big increase from 168.

4.3 Drug Toxicity Experiment

Njiaju et al. [5] used lymphoblastoid cell lines (LCL) to study in vitro drug toxicity of paclitaxel, a treatment for cancer. The goal was to identify genetic loci that are associated with drug induced toxicities. A common measure of toxicity is the dosage at which 50% of cells are killed (LD50). Five doses, i.e., 0, 6.25, 12.5, 25, 50, and 100nmol/l, were tested. At each dose, the percentage of cells surviving can be calculated, and the LD50 is estimated to fall between two of the adjacent doses. The cell lines were obtained from 247 subjects from three populations: a population with Northern and Western European ancestry from Utah, USA (CEU, n=77), a Yoruba population in Ibadan, Nigeria (YRI, n=83), and an African-American population from the Southwest of the USA (ASW, n=87). Due to the potential association between a subject's genetics and the degree of drug induced toxicity, and due to the potential difference in genetics in different populations, it may be of interest to test if the toxicity (LD50) differs between populations. In this example, we suppose the objective is to test for pairwise difference between any two of the three populations. Here the null hypothesis is that the LD50 is the same in the two populations being tested. By the notation of this article, the LD50 acts as T and the 5 doses tested are the endpoints of the grouping intervals: $a_0 = 0,\ a_1 = 6.25,\ a_2 = 12.5,\ a_3 = 25,\ a_4 = 50$ and $a_5 = 100$. If we make the proportional hazards assumption for LD50 in the two populations, then the null hypothesis is that the hazard ratio is equal to 0. We can then use the sample size formula to determine the necessary sample size to detect a certain effect. Suppose the comparisons are ASW versus CEU, CEU versus YRI, and YRI versus ASW, where

ASW, CEU and YRI is treated as the baseline, respectively. We first estimate the the α_js associated with the baseline hazard as well as the hazard ratio based on preliminary data. From these data, $\hat{\alpha}$ is (0.11, 1.06, 1.81, 2.19), (0.09, 0.97, 1.65, 1.95) and (0.08, 1.12, 1.90, 2.41), and the estimate of the hazard ratio is 0.74, 1.12 and 1.22, for the 3 comparisons above. Also, we assume the sample sizes in the three populations are approximately equal, as in the preliminary data. Under these assumptions, we need 430, 2962, and 936 subjects for the three pairwise comparisons, respectively, for 80% power and two-sided significance level 0.05. The numbers are for each of the two groups being compared. On the other hand, if we assume that 1.2 is the minimum meaningful hazard ratio, and the hazard increases from CEU to YRI and from YRI to ASW, which is indicated by the preliminary results, then the necessary sample sizes to detect the difference are 298, 1188, and 1130, respectively, for the 3 comparisons. Finally, suppose that we have 500 subjects from each population, and we use the Bonferroni method to control for multiple comparisons. Then we can detect a hazard ratio around 1.38, for the 3 comparisons above. In the above calculation, we assume that, for any subject all 5 doses are tested, which implies that there is no right censoring before the last dose level.

5 Discussion

We derived a sample size formula for clinical trials with grouped survival endpoints which has wide applications in practical fields such as HIV studies, cancer clinical trials and drug toxicity experiments. Besides the proportional hazards assumption, the formula does not make any additional distributional assumptions on the survival time. To calculate the sample size, only estimates of the hazard ratio and the distributions of the survival time and censoring time at the endpoints of the grouping intervals are needed. In our simulation study, it is noted that further increase of the frequency of examining times may not lead to much more increase in power when the frequency reaches a certain level. Therefore, in designing a study, the sample sizes under increasing frequencies of examination can be compared. And if at some point the decrease of the sample size is not significant with the increase of frequency of examination, then there is not much gain in efficiency by further increasing the frequency anymore and thus it is best to stop increasing the frequency of examination for other considerations such as cost. In this sense, the sample size formula can also be used to guide the selection of examining times.

Acknowledgements

We thank Dr. Eileen Dolan and her lab at the University of Chicago for providing the LCL data. We also thank two referees for their helpful comments. This research was supported in part by a grant with award number P01CA142538 from the National Cancer Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Kalbfleisch, JD and Prentice, RL. *The Statistical Analysis of Failure Time Data.* Wiley, New York; 2002.

2. Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer; 2006.

3. Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics.* 1978; **34**:57-67.

4. Li Z, Gilbert P, Nan B. Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics.* 2008; **64**:1247-1255.

5. Njiaju UO, Gamazon ER, Gorsic LK, et al. Whole-genome studies identify solute

carrier transporters in cellular susceptibility to paclitaxel. *Pharmacogenet Genomics.* 2012; **22**:498-507.

6. Lui KJ. Sample size determination for cohort studies under an exponential covariate model with grouped data. *Biometrics.* 1993; **49(3)**:773-778.

7. Lui KJ, Steffey D, Pugh JK. Sample size determination for grouped exponential observations: a cost function approach. *Biometrical J.* 1993; **35(6)**:677-688.

8. Inoue LYT, Parmigiani G. Designing follow-up times. *J Am Stat Assoc.* 2002; **97**:847-858.

9. Raab GM, Davies JA, Salter AB. Designing follow-up intervals. *Stat Med.* 2004; **23**:3125-3137.

10. Lachin JM. Power of the Mantel-Haenszel and other tests for discrete or grouped time-to-event data under a chained binomial model. *Stat Med.* 2014; **32(2)**:220-229.

11. Flynn NM, Forthal DN, Harro CD, et al. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *J Infect Dis.* 2005; **191**:654-665.

12. Panageas KS, L Ben-Porat, Dickler MN, Chapman PB, Schrag D.When you look matters: the effect of assessment schedule on progression-free survival. *J Natl Cancer I.* 2007; **99**:428-432.

13. Pazdur R. Endpoints for assessing drug activity in clinical trials. *The Oncologist.* 2008; **13** (Supplement 2):19-21.

14. Saad ED, Katz A. Progression-free survival and time to progression as primary end points in advanced breast cancer: often used, sometimes loosely defined. *Ann Oncol.* 2009; **20**:460-464.

15. George S, Wang XF, Pang H. *Endpoints for cancer clinical trials. In Cancer Clinical Trials: Current and Controversial Issues in Design and Analysis.* Chapman & Hall/CRC; 2016.

16. Baggstrom MQ, Socinski MA, Wang XF, et al. Maintenance Sunitinib following initial platinum-based combination chemotherapy in advancedstage IIIB/IV non-small cell lung cancer: a randomized, double-blind, placebo-controlled phase III studyCALGB 30607 (Alliance). *J Thorac Oncol.* 2017; **12(5)**:843-849.

17. Wu L, Cook RJ. Misspecification of Cox regression models with composite endpoints. *Stat Med.* 2012; **31**:3545-3562.

18. Manju MA, Candel MJJM, Berger MPF. Sample size calculation in cost-effectiveness cluster randomized trials: optimal and maximin approaches. *Stat Med.* 2014; **33**:2538-2553.

A Derivation of the Asymptotic Variance of the Estimator for the Log Hazard Ratio

We derive the asymptotic variance of the estimator for the log hazard ratio (β) by obtaining formulae for various components in the efficient information matrix, where the parameter of interest is β and the nuisance parameter is γ. First, note that K is random, and

$$Z\Delta d_K = Z\Delta \sum_{k=1}^{r-1} d_k I(K=k).$$

It follows that

$$\begin{aligned} E(Z\Delta d_K) &= E\left[Z\sum_{k=1}^{r-1} d_k E\{\Delta I(K=k)|Z\}\right] = E\left\{Z\sum_{k=1}^{r-1} d_k P(\Delta=1, K=k|Z)\right\} \\ &= E\left\{Z\sum_{k=1}^{r-1} d_k p(1,k|Z)\right\}. \end{aligned}$$

Similarly, we get the following equalities:

$$\begin{aligned} E\left(Z\sum_{i=1}^{K-1} h_i|Z\right) &= E\left\{Z\sum_{k=1}^{r}\left(\sum_{i=1}^{k-1} h_i\right) P(K=k|Z)\right\} \\ &= E\left[Z\sum_{k=2}^{r}\left(\sum_{i=1}^{k-1} h_i\right)\{p(1,k|Z)+p(0,k|Z)\}\right], \end{aligned}$$

$$\begin{aligned} &E\{Zh_i I(i<K)\} = E[E\{Zh_i I(i<K)\}|Z] \\ = \; &E\{Zh_i P(i<K|Z)\} = E\left\{Zh_i \sum_{\delta=0}^{1}\sum_{k=i+1}^{r} p(\delta,k|Z)\right\}, \end{aligned}$$

$$\begin{aligned} &E\{Z\Delta d_i I(K=i)\} = EE[\{Z\Delta d_i I(K=i)\}|Z] \\ = \; &E[Zd_i E\{\Delta I(K=i)|Z\}] = E\{Zd_i P(\Delta=1, K=i|Z)\} \\ = \; &E\{Zd_i p(1,i|Z)\}, \end{aligned}$$

$$\begin{aligned} & E\{h_i I(i < K)\} = E[E\{h_i I(i < K)\}|Z] \\ = \quad & E\{h_i P(K > i|Z)\} = E\left\{h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|Z)\right\}, \end{aligned}$$

and

$$\begin{aligned} & E\{\Delta d_i I(K = i)\} = E[E\{\Delta d_i I(K = i)\}|Z] \\ = \quad & E\{d_i P(\Delta = 1, K = i|Z)\} = E\{d_i p(1, i|Z)\}. \end{aligned}$$

Thus, denoting $p_z = P(Z = z)$, we have

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial \beta^2}\right) &= E\left\{Z\left(\Delta d_K + \sum_{i=1}^{K-1} h_i\right)\right\} \\ &= \sum_{z=0}^{1} z\left[\sum_{k=1}^{r-1} d_i p(1, k|z) + \sum_{k=2}^{r}\left(\sum_{i=1}^{k-1} h_i\right)\{p(1, k|z) + p(0, k|z)\}\right] p_z, \end{aligned}$$

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial \beta \partial \gamma_i}\right) &= E\left\{Z h_i I(i < K) + Z \Delta d_i I(K = i)\right\} \\ &= E\left\{Z h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|Z) + Z d_i p(1, i|Z)\right\} \\ &= \sum_{z=0}^{1}\left\{z h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|z) + z d_i p(1, i|Z)\right\} p_z, \ 1 \le i \le r-1, \end{aligned}$$

and

$$\begin{aligned} -E\left(\frac{\partial^2 l}{\partial \gamma_i^2}\right) &= E\left\{h_i I(i < K) + \Delta d_i I(K = i)\right\} \\ &= E\left\{h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|Z) + d_i p(1, i|Z)\right\} \\ &= \sum_{z=0}^{1}\left\{h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta, k|z) + d_i p(1, i|z)\right\} p_z, \ 1 \le i \le r-1. \end{aligned}$$

Finally, the asymptotic variance of $\hat{\beta}$, in the presence of the nuisance parameter γ, is the right bottom component of I^{-1} which is A^{-1}. In other words, $\sigma^{-2} = A_1 - A_2$, where

$$A_1 \quad = \quad \sum_{z=0}^{1} z \left[\sum_{k=1}^{r-1} d_k p(1,k|z) + \sum_{k=2}^{r} \left(\sum_{i=1}^{k-1} h_i \right) \{p(1,k|z) + p(0,k|z)\} \right] p_z,$$

and

$$A_2 \quad = \quad \sum_{i=1}^{r-1} \frac{\left[\sum_{z=0}^{1} \left\{ z h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta,k|z) + z d_i p(1,i|z) \right\} p_z \right]^2}{\sum_{z=0}^{1} \left\{ h_i \sum_{\delta=0}^{1} \sum_{k=i+1}^{r} p(\delta,k|z) + d_i p(1,i|z) \right\} p_z},$$

where the function $p(\delta,k|z)$ is defined by (1) and (2) in the main text.

List of Figures

1 Examples of empirical powers based on sample sizes calculated from our formula when the proportional hazards assumption is violated. 26

2 Variance of $\hat{\beta}$ as a function of β, showing that the approximation of variance assuming $\beta = 0$ may not be accurate. 27

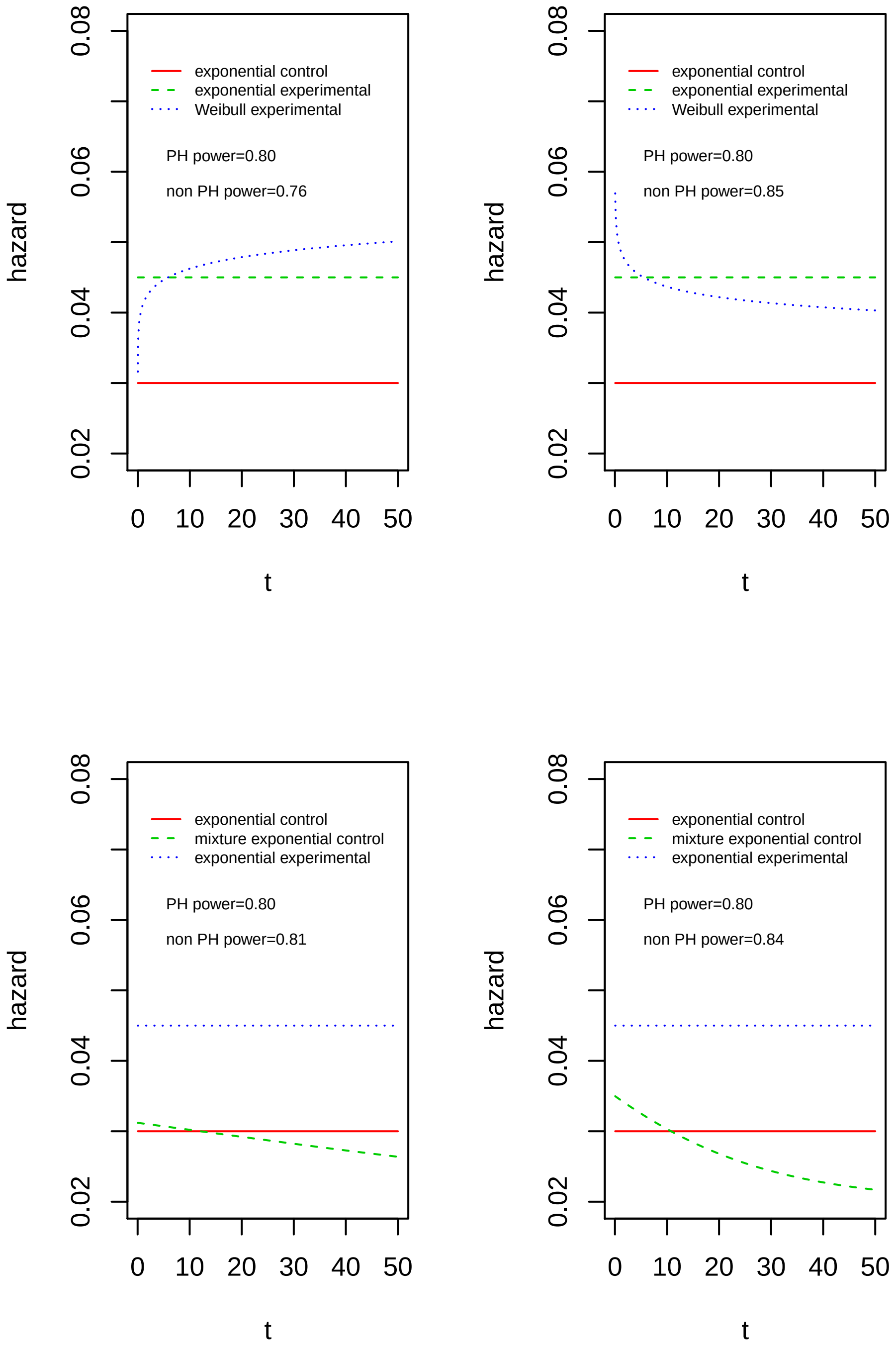

Figure 1: Examples of empirical powers based on sample sizes calculated from our formula when the proportional hazards assumption is violated.

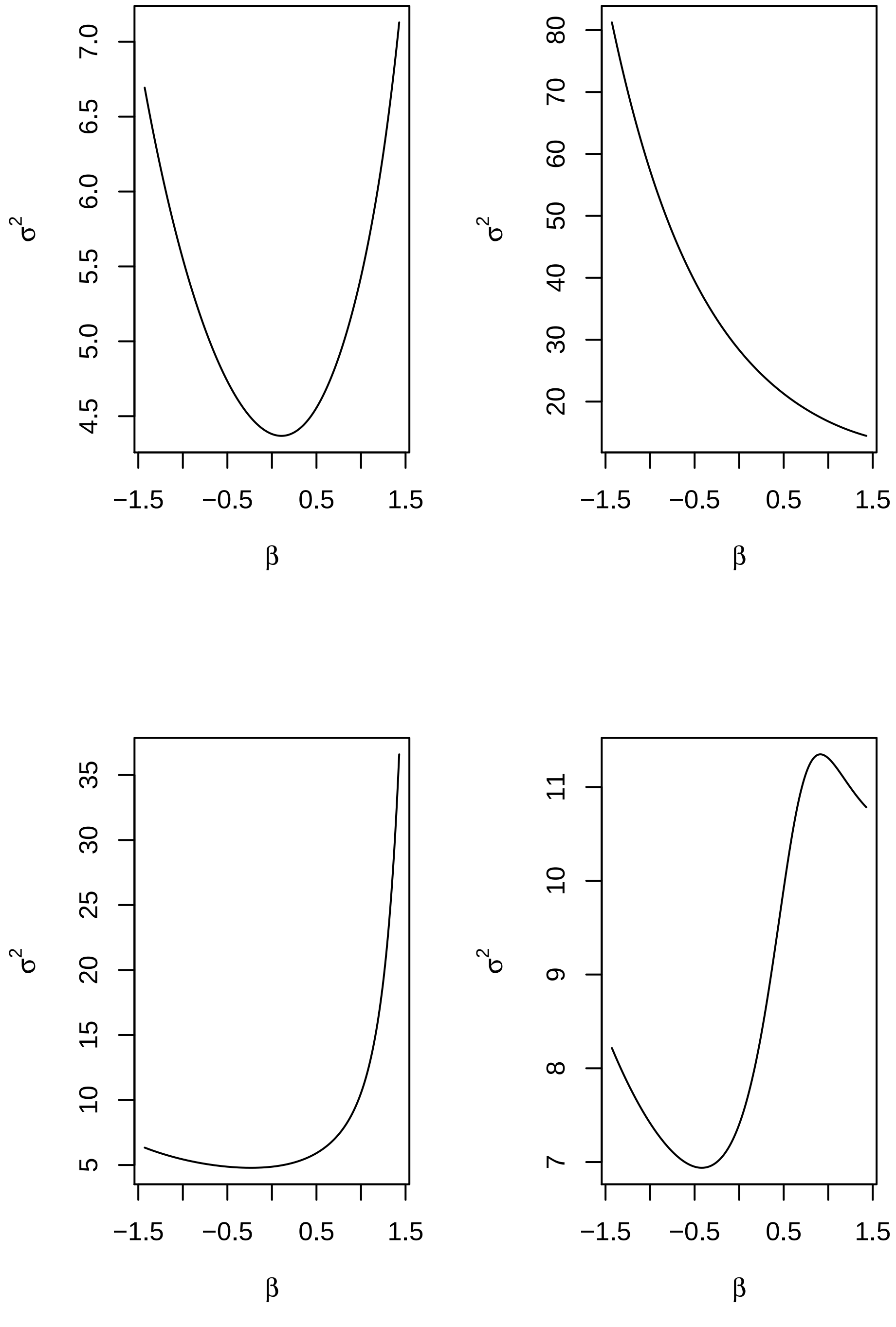

Figure 2: Variance of $\hat{\beta}$ as a function of β, showing that the approximation of variance assuming $\beta = 0$ may not be accurate.

List of Tables

1 Simulated bias(absolute percent bias) for the estimation of the log hazard ratio using the naive method and the Prentice and Gloeckler method (PG). 29

2 Sample size calculated from our formula (n), sample size based on approximation of variances (n_1), and the corresponding simulated power with sample size n. 30

3 Comparison of sample size calculated from our formula (n) and sample size based on approximation of variance (n_1) in a number of scenarios where there is considerable difference between n and n_1. Simulated powers are also included in parentheses. 31

4 Change of sample size with the number of intervals in a fixed time period [0,30]. 32

Table 1: Simulated bias(absolute percent bias) for the estimation of the log hazard ratio using the naive method and the Prentice and Gloeckler method (PG).

		true hazard ratio = 2			true hazard ratio = 4		
		$r = 3$	$r = 6$	$r = 10$	$r = 3$	$r = 6$	$r = 10$
Exponential	PG	0.004(0.0%)	0.002(0.0%)	0.004(0.0%)	0.003(0.2%)	-0.000(0.0%)	0.004(0.0%)
Exponential	Naive	-0.64(92%)	-0.30(44%)	-0.10(14%)	-1.33(96%)	-0.91(66%)	-0.42(30%)
Weibull	PG	0.003(0.5%)	0.000(0.0%)	0.002(0.4%)	0.004(0.4%)	-0.009(0.6%)	0.006(0.4%)
Weibull	Naive	-0.24(34.2%)	-0.24(35.0%)	-0.07(10.6%)	-0.53(37.9%)	-0.50(36.2%)	-0.15(11.0%)

r: $r - 1$ is the number of intervals (with equal lengths) before 30, and $(30, \infty)$ is the last interval.

Table 2: Sample size calculated from our formula (n), sample size based on approximation of variances (n_1), and the corresponding simulated power with sample size n.

hazard ratio	expected power	$n(n_1)$	simulated power	$n(n_1)$	simulated power
		no right censoring*		uniform right censoring$^\flat$	
		Exponential distribution			
1.3	80	755(771)	80.4	1008(1033)	82.4
1.5	80	314(323)	80.9	418(433)	81.9
1.7	80	182(189)	81.2	243(253)	81.5
2.0	80	107(112)	82.5	142(148)	82.3
1.3	90	1004(1032)	88.9	1339(1383)	91.7
1.5	90	416(432)	91.3	553(579)	91.4
1.7	90	241(253)	91.5	320(339)	90.4
2.0	90	141(148)	90.6	187(199)	92.5
		Weibull distribution			
1.3	80	544(548)	81.4	751(758)	80.8
1.5	80	228(230)	80.3	314(318)	82.1
1.7	80	134(135)	80.8	184(186)	82.2
2.0	80	79(79)	81.9	108(109)	82.7
1.3	90	727(735)	90.5	1002(1015)	92.7
1.5	90	304(308)	90.8	419(425)	92.5
1.7	90	178(180)	90.5	245(249)	91.6
2.0	90	105(106)	91.7	145(146)	92.3

*: no right censoring except at time 30; $^\flat$: right censoring time has a mass 1/3 at 30 and is otherwise uniformly distributed in (0, 30).

Table 3: Comparison of sample size calculated from our formula (n) and sample size based on approximation of variance (n_1) in a number of scenarios where there is considerable difference between n and n_1. Simulated powers are also included in parentheses.

n	323 (79.3)	403 (77.0)	444 (75.0)	523 (80.5)	418 (86.1)	380 (77.0)
n_1	296 (73.8)	368 (69.4)	395 (66.5)	497 (77.7)	446 (89.5)	347 (73.6)

Table 4: Change of sample size with the number of intervals in a fixed time period [0,30].

		Weibull distribution				exponential distribution			
hazard	expected	number of intervals				number of intervals			
ratio	power	3	6	10	15	3	6	10	15
						no right censoring[⋆]			
	0.80	573	532	528	526	732	719	717	716
1.3	0.85	656	610	604	601	837	822	820	819
	0.90	768	713	707	704	980	962	960	959
	0.80	242	222	220	219	298	292	291	291
1.5	0.85	277	254	252	250	341	334	333	333
	0.90	324	298	295	293	399	391	390	390
						uniform right censoring[♭]			
	0.80	1027	733	665	634	1275	950	875	838
1.3	0.85	1175	839	761	725	1458	1086	1000	959
	0.90	1375	981	891	848	1707	1271	1171	1122
	0.80	431	305	275	262	519	384	352	337
1.5	0.85	493	349	315	300	594	439	403	386
	0.90	577	409	369	351	695	514	472	451

[⋆]: no right censoring except at time 30; [♭]: right censoring time has a mass 1/3 at 30 and is otherwise uniformly distributed in (0, 30).