4-15-2016

Power and Sample Size Calculations for Interval-Censored Survival Analysis

Hae-Young Kim
New York Medical College

John M. Williamson
CDC

Hung-Mo Lin

Recommended Citation

Kim, H., Williamson, J. M., & Lin, H. (2016). Power and Sample Size Calculations for Interval-Censored Survival Analysis. *Statistics in Medicine, 35* (8), 1390-1400. https://doi.org/10.1002/sim.6832

Power and Sample Size Calculations for Interval-Censored Survival Analysis

Short title: Interval-Censored Power Calculation

Hae-Young Kim[1*], John M. Williamson[2], and Hung-Mo Lin[3]

[1] Department of Epidemiology and Community Health, New York Medical College, 40 Sunshine Cottage Rd, Valhalla, NY 10595, U.S.A.

[2] Division of Parasitic Diseases and Malaria, National Center of Global Health, Centers for Disease Control and Prevention (MS A-06), 1600 Clifton Road, Atlanta, GA 30329, U.S.A.

[3] Department of Population Health Science and Policy, Ichan School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1010, New York, NY 10029, U.S.A.

* haeyoung_kim@nymc.edu phone: 914-594-4622

SUMMARY: We propose a method for calculating power and sample size for studies involving interval-censored failure time data that only involves standard software required for fitting the appropriate parametric survival model. We use the framework of a longitudinal study where patients are assessed periodically for a response and the only resultant information available to the investigators is the failure window: the time between the last negative and first positive test results. The survival model is fit to an expanded data set using easily computed weights. We illustrate with a Weibull survival model and a two-group comparison. The investigator can specify a group difference in terms of a hazards ratio. Our simulation results demonstrate the merits of these proposed power calculations. We also explore how the number of assessments (visits), and thus the corresponding lengths of the failure intervals, affect study power. The proposed method can be easily extended to more complex study designs and a variety of survival and censoring distributions.

Keywords: Interval-censored data; power; sample size; parametric survival analysis.

1 Introduction

Interval-censored failure data often arise in longitudinal studies in which subjects are assessed only periodically for the response of interest [1]. The time when the event of interest occurs is not directly observed but is known to take place within some time interval. For example, in HIV studies, investigators cannot observe the exact moment when the virus develops; all they can determine is that the virus developed before or after the test. Interval-censored failure data often occur in observational or follow-up studies where patients are not continuously being observed. Whether or not the event occurred is ascertained at the observation times, and the failure time of the event itself is not available.

There have been numerous methods proposed for the analysis of interval-censored failure data. Peto and Peto [2] first considered the comparison of the interval-censored survival curves of two samples under the Lehman-type alternative $S_1(t) = S_2^{\theta}(t)$ where θ is the parameter of interest. They test $\theta = 0$ using the score test and describe it as the log-rank test. Finkelstein [3] proposed a semiparametric method in which the baseline distribution and regression parameters are fit simultaneously by maximizing the full likelihood of the data. Sun [4] proposed a test statistic for interval-censored failure data having the same algebraic form as the original log-rank test. Zhao and Sun [5] generalized the Sun's [4] log-rank test to include exact failure times in interval-censored data. Sun, Zhao, and Zhao [6] proposed a class of non-parametric tests for the comparison of k interval-censored survival curves that are generalizations of Peto and Peto's [2] log-rank test. Their test statistic includes Finkelstein's [3] test statistic as a special case.

Fay [7] proposed a weighted log-rank test under the proportional odds model, which gives more weight to earlier times. Satten [8] considered a marginal likelihood approach to fitting the proportional hazards model [9, 10] by maximizing a likelihood that is the sum over all rankings of the data that are consistent with the observed censoring intervals. Satten, Datta, and Williamson [11] suggested a parametric model for the baseline hazard to generate imputed failure times. In their model the usual proportional hazards model for right-censored data is used to estimate the regression parameters. Heller [12] proposed a method for estimation and inference of the regression parameters in the Cox proportional hazards model with interval censoring based on estimating equations and using an inverse probability weight to select event time pairs where the ordering is unambiguous. A Bayesian estimation approach has recently been proposed for analyzing interval-censored data under the proportional hazards model [13].

A special case of interval-censored data is current-status data, where individuals are seen only once after enrollment. Current-status data often arise in cross-sectional surveys, where the purpose is calculation of the distribution of age of onset for a disease or life event. Thus, the observations are either of the form $(0, C]$ or (C, ∞) (i.e., left- or right- censored). These data are also commonly referred to as case 1 interval-censored data [14]. Current status data are common in demography [15, 16], economics, and epidemiology [17, 18]. In the medical sciences, animal tumorigenicity and HIV studies often result in such data because the investigator cannot measure the outcome directly or accurately [19]. The proportional hazards models and tests referenced above for analyzing interval-censored data can be used for the analysis of current status data. Murphy and van der Vaart [20] considered semiparametric likelihood ratio inference and proposed a test for significance of the regression coefficient in Cox's regression model for current status data. Banerjee [21] examined the power of the test under contiguous alternatives.

Methods for calculating power and sample size for studies involving interval-censored survival data are scarce. Such calculations are important because studies without a sufficient sample size may not detect a clinically important effect. This consideration must be balanced with the high cost of recruiting and evaluating large numbers of subjects, thus making sample size and power calculations an important element in the design of medical research studies. Sample size calculations are especially important for studies using interval-censored failure data because they have less information and power than survival studies with the usual right-censored outcome data.

Williamson, Lin and Kim [22] proposed power and sample size calculations for current status survival analysis based on the Wald test assuming a Weibull survival model and various censoring distributions. As expected, the power calculations demonstrated that studies with current status data have substantially less power than studies with the usual right-censored failure time data. Marschner [23] proposed a method for designing a cross-sectional survey to estimate the age-specific incidence of an irreversible disease (resulting in current status data). It is assumed that the sample consists only of information on the current age and disease status of the individuals. Marschner focused on determining the total size of the sample and how to best choose the distribution of sampling across various age groups. Zhao, Duan, Zhao and Sun [24] proposed a new class of generalized log-rank tests and derived their asymptotic distributions under both null and alternative hypotheses. Their derivations allow power estimation under the specification of an alternative hypothesis.

In Section 2, we propose power calculations for studies comparing two groups with interval-censored failure data. We first specify the scenario (underlying survival distribution, group sizes, hazard ratio, length of study, number of study visits, dropout rate, missing data, etc.) for which we want to conduct the power calculation. We then fit the specified interval-censored failure model to an appropriate expanded data set, which is a created dataset that exemplifies the sampling distribution of the population of interest. The use of an expanded dataset has been applied to aid power and sample size calculation for fixed effect linear and generalized linear models [25-29]. Specifically, the expanded dataset comprises one record for each possible value of the outcome per combination of covariate values. In the presence of interval-censored data with no missing visits, there is only one possible outcome (i.e., drop out) before the first follow-up visit. All other subsequent intervals, except the last one, either have a failure in that interval or a drop out (right censoring) after the previous visit. The last interval also has two possible outcomes: failed in that interval or right censored at study end. Therefore, for a simple 2-group comparison, the expanded dataset will consist of ((2 × total number of scheduled follow-up visits) + 1) data lines for each subject assuming no missed visits.

In addition, we need to provide a weight for each record to reflect the probability of such occurrence, with the weight calculated from the parameters specified for the survival and censoring distributions. The weights sum across the potential failure and right-censoring intervals to 1.0 for each individual. The resulting expanded or 'exemplary' interval-censored failure data set can be easily analyzed with commonly used software (e.g., PROC LIFEREG in SAS, v 9.3 [30]) that incorporates weighting. The resulting maximum likelihood estimate of the parameters will have the same values as the assumed parameters. The variance-covariance matrix computed from the model fit is then used in conjunction with an established non-central chi-square approximation to the distribution of the Wald statistic. The same formulation of weights can be extended to allow for missed study visits. For the purpose of illustration we focus on a simple situation where the probability of a missed visit is common across visits and two visits can not be missed consecutively. As there are numerous potential failure intervals resulting from missing visits, this exercise demonstrates how one can modify the weights for the missing visit pattern applicable to one's study.

We present the details of our approach in Section 2. In Section 3 we present simulation studies to detail its performance. We also explore the relationship between the number of study visits (size of failure intervals) with power. We illustrate the proposed calculations in

Section 4 with a hypothetical example based on a breast cancer study [31]. We conclude with a short discussion on the merits of the proposed power calculations.

2 Methods

Let T_i denote the log-transformed failure time for the i^{th} observation ($i = 1, \cdots, N$, where N is the sample size). If data are interval censored, then for each individual, instead of a failure time, we observe a censoring interval $(l_i, u_i]$ that is known to contain the actual failure time. The failure indicator is defined as $\delta_i = 1$ if the i^{th} observation is of the form $(l_i, u_i]$ (interval censored, or left-censored if $l_i = 0$). If the observation is right-censored ($u_i = \infty$) then $\delta_i = 0$. We assume throughout that the censoring/dropout mechanism is independent of both the response time and the covariates. Let the survivor distribution for the failure time random variable T be denoted by $S(t; \boldsymbol{\alpha}, \boldsymbol{\beta}) = Pr(T \geq t)$, where $t \geq 0$, $\boldsymbol{\alpha}$ is a column vector of scale or shape parameters, and $\boldsymbol{\beta}$ is a $(p \times 1)$ column vector of regression parameters. The likelihood for such interval-censored failure data is

$$L(l_i, u_i, \delta_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_{i=1}^{N} [S(l_i) - S(u_i)]^{\delta_i} [S(l_i)]^{1-\delta_i}, \tag{1}$$

where $S(0) = 1$. Further assume that the log-transformed failure times follow a Weibull distribution that can be parameterized with an intercept Δ as $Pr(T_i \geq t_i) = S(t_i; \Delta, \beta) = exp(-e^{(t_i - \Delta - \boldsymbol{\beta}'\mathbf{x_i})\gamma})$, where $\mathbf{x_i}$ is a column vector of covariates and γ is a shape parameter. This model is equivalent to an accelerated failure time model and is also a member of the proportional hazards family for the Weibull distribution.

We are interested generally in the following hypothesis test:

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h_0} \quad versus \ H_A : \mathbf{H}\boldsymbol{\beta} \neq \mathbf{h_0}, \tag{2}$$

where $\mathbf{H}$ is an $(h \times p)$ matrix of full row rank and $\mathbf{h_0}$ is an $(h \times 1)$ constant vector. The Wald test statistic is given by

$$T_W = (\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h_0})'[\mathbf{H}\mathbf{v\hat{a}r}(\hat{\boldsymbol{\beta}})\mathbf{H}']^{-\mathbf{1}}(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h_0}). \tag{3}$$

Assume that our future study will be based on a two-group (e.g., new drug versus placebo) comparison, which is usually the case for most sample size calculations conducted for medical and public health studies. Thus, we have a binary covariate $x = 0, 1$ depending on group membership. Let there be n subjects in group 1 ($x_i = 0$) and rn subjects (with $r > 0$) in group 2 ($x_i = 1$). The hypothesis of interest is $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$ where β is a scalar. In particular, $exp(-\beta)$ is the hazard ratio between groups 2 ($x_i = 1$) and 1 ($x_i = 0$). Further assume that both groups have the same censoring/dropout distribution.

For this two-group comparison the Wald test statistic is given by

$$T_{\text{Wald}} = \widehat{\beta}^2/(v\widehat{a}r(\widehat{\beta})),$$

where $\widehat{\beta}$ is the maximum likelihood estimate of β. Under H_0, the test statistic is asymptotically distributed as a central chi-square random variable with 1 degree of freedom. Under H_A, and following Wald [32], T_{Wald} is asymptotically distributed as a non-central chi-square random variable with 1 degree of freedom where the non-centrality parameter ω is equal to the value of T_{Wald}, except with $\widehat{\beta}$ and $v\widehat{a}r(\widehat{\beta})$ replaced with β and $var(\widehat{\beta})$. Let α represent the specified type I error rate and $\chi^2_{1,1-\alpha}$ represent the critical value from the central χ^2_1 distribution. The power for testing H_0 with the Wald test is

$$Pr(\chi^2_{1,(\omega)} \geq \chi^2_{1,1-\alpha}) \tag{4}$$

with $\chi^2_{1,1-\alpha}$ denoting the $100(1-\alpha)$th percentile of the central chi-square distribution with one degree of freedom and $\chi^2_{1,(\omega)}$ denoting a chi-square random variable with one degree of freedom and non-centrality parameter ω.

2.1 *Study Design Consideration*

2.1.1 *No Missing Visits*

Assume our study design is a longitudinal one with regularly scheduled visits. Let ST denote the length of the study and assume there are Q scheduled follow-up visits for each individual with a visit every ST/Q units of time. For instance, we could have two years of follow-up with bi-monthly visits implying $ST = 24$ and $Q = 12$. We allow that individuals may be variable in their visit times, which we denote $t_{i1}, t_{i2}, \cdots, t_{iQ_i}$ and some visits may be missed due to dropout. We can assume that the first follow-up visit time scheduled at 2 months for each individual is uniformly varied between 1.5 and 2.5 months, and the remaining visit times are spaced out in exact increments there after (e.g., an individual having a first visit at 1.93 months has remaining visits at 3.93, 5.93, 7.93, ... months). We specify the percentage of failures occurring prior to the end of the study for the $x = 0$ group to account for censoring at study end. The corresponding percentage of failures for the $x = 1$ group will be lower or higher depending on the direction of the specified hazard ratio. We also incorporate dropout into our scenario by specifying a percentage of dropout by study end and assume that it is uniformly distributed throughout the study and is the same for both groups. For example, if we assume 20% dropout by month 24 then 10% of the subjects will not have another visit after their 12-month follow-up.

We are interested in calculating conditional power (power given any pre-specified fixed covariate design matrix) for a given sample size. In order to conduct the power calculations we need to obtain $var(\widehat{\beta})$ for computing the chi-square non-centrality parameter. We do this by following Lyles et al. [29] and create an expanded data set. We define a weight, w_{ij}, for each potential failure/censoring interval (indexed by j) for each individual i that equals the probability that such interval will occur. For our scenario failure intervals can occur between successive visits and individuals can be right censored at visit times for dropout, or censored at study end. Accordingly, potential failure intervals will be of the form $(t_{iq}, t_{i,q+1}]$, and right-censored observations will be of the form (t_{iq}, ∞). For the former intervals $(t_{iq}, t_{i,q+1}]$, weights are calculated as $Pr(t_{iq} < T_i \leq t_{i,q+1})Pr(C_i > t_{i,q+1})$. For the latter intervals (t_{iq}, ∞), weights are calculated as $Pr(t_{iq} < T_i)Pr(t_{iq} < C_i \leq t_{i,q+1})$, and $Pr(24 < T_i)Pr(24 < C_i)$ for the last interval $((24, \infty))$. The weights will sum to 1.0 for each individual ensuring that the expanded data set has total sample size N. The data lines for individual i in group j, $j = 1, 2$, in this expanded or ’exemplary’ interval-censored failure dataset resemble the following:

		interval limit			
id	group	lower	upper	failure status	weight
i	j	0	∞	0	$w_{ij,1}$
i	j	0	t_{ij1}	1	$w_{ij,2}$
i	j	t_{ij1}	∞	0	$w_{ij,3}$
i	j	t_{ij1}	t_{ij2}	1	$w_{ij,4}$
i	j	t_{ij2}	∞	0	$w_{ij,5}$
	$\ldots$				
i	j	$t_{ij(Q-1)}$	t_{ijQ}	1	$w_{ij,(2Q)}$
i	j	t_{ijQ}	∞	0	$w_{ij,(2Q+1)}$

See Table 1 for an illustration of an expanded data set for one of the data sets used in the simulation section. As shown in Table 1, the weight calculations are determined by the failure, dropout, and censoring distributions. Therefore, it is straightforward to extend the proposed method to allow for different dropout rates between groups or to have more than 2 groups in the study. Subjects may dropout before the first follow-up visit and therefore the interval $(0, \infty)$ will have a corresponding non-zero weight. These data lines will add no information when fit in the parametric survival model.

Table 1 about here

2.1.2 *Missing Visits*

Thus far we have assumed that all subjects make their scheduled visits until study end or until they have the outcome of interest. However, subjects may miss a study visit for a number of reasons, usually resulting in data with larger failure intervals. To account for potential missing visits, we make the following simplifying assumptions. First, we assume that each subject has the same probability (denoted by p) of missing a specific visit that is the same for all Q visits. For simplicity we also assume that a subject can not miss two consecutive visits, i.e., if a subject misses visit q then he or she makes visit $q + 1$. This implies that p is constrained to be ≤ 0.5. The major modification to the proposed method for missing data is to reconstruct the weights to incorporate the resulting larger intervals. As before, no information is gained when a subject drops out without any follow-up. There are five potential outcomes for the interval between visits q and $q+1$ for $1 \leq q < Q-1$: (a) Making visit q and then dropping out; (b) Missing visit q and then dropping out; (c) Making visit q and then failing; (d) Missing visit q and then failing; and (e) Failing and then missing visit $q+1$. See Table S1 of the Web Appendix for an example of one of the exemplary data sets that incorporates missing visits.

We then fit a Weibull failure model to the expanded data set in an available software package that incorporates weighting (e.g., PROC LIFEREG). The resulting maximum likelihood estimate of the parameters will have the same values as the assumed parameters. In addition, the resulting $var(\widehat{\beta})$ will equal the true variance for the specified sample size and can then be used for power calculation. In summary, power calculations for the proposed scenario proceed as follows:

1. Specify the sample sizes for each group, n and rn ($n + rn = N$), and the type I error rate (usually 0.05).

2. Specify the length of the study (ST) and how many schedule visits (Q) during follow-up. This will determine the time interval between two visits (ST/Q).

3. Specify a distribution (e.g., uniform distribution) for the first visit time (centered at time $t = ST/Q$) and assume that the remaining visit times are spaced out in exact increments thereafter. This will allow variation in visit time around any given scheduled visit.

4. Specify the regression parameter β, or the hazard ratio $e^{-\beta}$.

5. Assume the log-transformed survival time follows a Weibull distribution with intercept Δ, which can be estimated by specifying the percentage of subjects in group $x = 0$ who fail by study end assuming no dropout, and the shape parameter γ.

6. Specify a percentage of dropout by study end that is assumed to be uniformly distributed across study time and is the same for both groups.

7. Specify a probability (p) that a subject will miss a visit that is assumed to be constant across all Q visits.

8. Based on the specified failure time model, dropout model, and missing visit probability, construct the weights w_{ij} for each subject for all potential failure/censoring intervals.

9. Create an expanded dataset that has multiple lines corresponding to all possible outcomes for each of the N subjects, where each line includes a subject identifier i, group indicator (x_i), lower and upper limits (l_{ij} and u_{ij} corresponding to the appropriate visit times) of a given interval, and the corresponding weights w_{ij}. With no missing visits,

the number of lines is $2Q+1$, and with the missing visit scenario considered here, the number of lines is $5Q$.

10. Calculate the variance-covariance matrix based on the specified parameters and sample size by fitting the parametric model (e.g., in PROC LIFEREG with the WEIGHT statement).

11. Obtain the noncentrality parameter $\omega = \widehat{\beta}^2 / v\widehat{a}r(\widehat{\beta})$.

12. Use equation (4) to calculate the power.

13. Repeat steps 9-12 by increasing or decreasing the sample size (N) until the desired power is achieved.

3 Simulations

We conducted a simulation trial to assess the performance of the proposed power calculations by comparing the calculated power from the proposed method with the empirical power from the simulations. Each simulated data set consisted of two groups (exposed or unexposed) of observations. The log-transformed survival times were generated with a Weibull distribution as follows:

$$S(t_i) = exp(-e^{(t_i - \Delta - \beta x_i)\gamma}),$$

where i denoted the subject $i = 1, \cdots, N$. The covariate $x_i = 0$ (1) for the unexposed (exposed) group. The regression parameter of interest is denoted by β and γ is the shape parameter. We assumed individuals were scheduled for 6 visits every 4 months (up to 24 months) until they had the event of interest. The first visit time for each individual was uniformly varied between 3.5 and 4.5 months and the remaining visit times were spaced out in exact 4 month increments (e.g., an individual having a first visit at 3.93 months had remaining visits at 7.93, 11.93, 15.93, ... months). The probability of missing a visit (p) was specified as 0.0 and 0.4.

Censoring/dropout was generated as follows. We generated differing censoring levels by varying Δ to accommodate censoring percentages for the $x = 0$ group at 24 months of 10%, 30%, and 50%. We also incorporated dropout of 10%, 20%, and 30% uniformly by 24 months. This resulted in 3 overall censoring schemes: low, medium, and high. Data sets were generated for 18 triplets of β =log(1.3), log(1.5), and log(1.7) (corresponding to hazard ratios of 0.77, 0.67, and 0.57), $\gamma = 0.5, 1.0, 1.5$, and missing visit probabilities ($p = 0.0, 0.4$) for each of the three censoring schemes, resulting in a total of 54 scenarios. A value of $\gamma = 1.0$ corresponds to the exponential distribution.

There were an equal number of exposed and unexposed observations in each data set, and 5,000 data sets were generated for each scenario. See Table 1 for an example of one of the exemplary data sets for the first scenario when $p = 0.0$ (no missed visits) and Table S1 of the Web Appendix for the same subjects when $p = 0.4$. The data sets were analyzed with a parametric Weibull model and with Sun, Zhao and Zhao's [6] generalized log-rank test using PROC LIFEREG and PROC ICLIFETEST in SAS (v 9.4), respectively. Power was calculated for each scenario using the proposed method. Empirical power was calculated for each of the two tests as the number of data sets resulting in the rejection of $H_0 : \beta = 0$ ($\alpha = 0.05$) divided by 5,000.

The simulation results for the first trial are presented in Table 2. The expected percentage of observations failing ($\delta = 1$) was calculated for each triplet of β values, γ values, and censoring amounts (light, medium, or heavy). The calculated power was within an absolute 1% of the empirical power for 29 of the 54 scenarios with the parametric Weibull model, and with 28 of the 54 scenarios for Sun, Zhao, and Zhao's [6] test. The calculated power was within an absolute 5% of the empirical power for all but one scenario with the parametric

Weibull model and for all but two scenarios with Sun, Zhao, and Zhao's [6] test. As expected, power increases as the effect sizes, shape values, and failure percentages increase for a given sample size. Power decreases as the probability of a subject missing a visit increases due to wider failure intervals. The scenarios with a greater difference between the calculated and empirical power were those with larger β and γ values, and smaller expected numbers of failures.

As with most power calculations, accuracy of the proposed method is dependent upon correct specification of the survival distribution. We conducted a new simulation trial where we generated data with a log-logistic distribution with shape parameters 2.0 and 3.0; β values of 1.2, 1.4 and 1.6; and the same 3 censoring schemes as in the Simulation section (light, medium, and heavy), with no chance of a missing visit ($p = 0$). We calculated the interval weights based on the specified log-logistic distribution but then analyzed the 5000 simulated data sets for each scenario assuming a Weibull distribution, and calculated empirical power. As expected the power calculations were somewhat off. For the parametric model the calculated power was only within an absolute 5% of the empirical power for 8 of the 18 scenarios, although within an absolute 10% for all but 2 scenarios. For Sun, Zhao, and Zhao's [6] test the calculated power was only within an absolute 5% of the empirical power for 10 of the 18 scenarios, and within an absolute 10% for 15 scenarios. See Table S2 of the Web Appendix for the results. All simulations were conducted via SAS IML [33].

Table 2 about here

3.1 *Impact of Failure Interval Size on Power*

We chose three scenarios (amount of dropout/censoring and effect size) from the exponential distribution ($\gamma = 1.0$) with $p = 0.0$ in the simulation trial (Table 2) and conducted power calculations varying the number of visits from 1 to 24 (size of failure intervals). See Table 3. In particular, the first line of Table 3 corresponds to a current status study design (one assessment time). For each of the three scenarios power increased as the number of visits increased (smaller failure intervals), as expected, with the biggest increase occurring between one visit (current status data) and two visits. However, there is only a negligible increase in power after 3 or 4 visits. Investigators should examine the relationship between power, number of study visits, and the financial cost of each visit for their potential study as a small increase in power due to more visits may not offset the increase in expense and time due to more hospital or clinic visits and/or laboratory tests. Raab, Davies, and Salter [34] considerd the design of follow-up intervals in the context of the estimation of the median and mean survival and for covariates in parametric regression models with equally spaced examination times. Bayesian approaches [35, 36] have been proposed for planning optimal follow-up times in a sequential manner, based on accumulated data. Others have examined the loss of information due to interval censoring for various parametric distributions [37, 38].

Table 3 about here

4 Illustrative Example

Suppose one wants to design a breast cancer study where two treatments are being compared for an interval-censored failure outcome. This illustration is motivated by the data in Table 3 of Finkelstein and Wolfe [31] and presented in Guo, So, and Johnston [39]. A retrospective study of 94 women was conducted on the risk of breast cosmetic deterioration after tumorectomy. The women received either radiation therapy ($x = 0$) or radiation plus chemotherapy ($x = 1$) and visited the clinic every four to six months. No woman was seen after 48 months and 38 women never experienced the outcome. Finkelstein and Wolfe

[31] and Finkelstein [3] analyzed the data with a semiparametric regression model and a semiparametric proportional hazards model, respectively.

Assume that two groups of women of equal size (e.g., treatment groups A and B) will be examined every 6 months for breast cosmetic deterioration. Assume an exponential survival distribution ($\gamma = 1.0$) with 40% of the women in the $x = 0$ group not having the event by study end of 48 months. Also assume that 5% of the enrolled will drop out each year. See Table 4 for sample size calculation for varying effect sizes (hazard ratios= $1.50, 1.75, 2.00, 2.25, 2.50$), power values $(0.80, 0.90)$, and probabilities of missing a visit $(0.0, 0.2, 0.4)$. As expected, the required sample size decreases with an increasing hazard ratio (effect size) for given power. Sample size increases with a larger probability of missing a visit for given power. Although the investigator needs to add more subjects for smaller hazard ratios when accounting for missing visits, the percent increase of sample size compared to no missing visits remains similar across the range of hazard ratios. Assume one is interested in specifically detecting a clinically important effect corresponding to a hazard ratio of at least 2.0 ($\beta = -\log(2)$). A sample of 69 women in each group would achieve 90% power for the proposed study assuming no missed visits, but 72 women per group would be required if $p = 0.4$.

Table 4 about here

5 Discussion

Interval-censored failure data are a special case of survival data in which the only information available to the investigator is whether an event occurred before or after one or more visit (examination) times. Such data are increasing in medical studies due in part to the greater use of biomarkers that define a disease progression endpoint [12]. One loses information and thus power when analyzing such data as compared to the usual right-censored survival data due to the increased imprecision in the failure time. Here, we propose power calculations for studies with interval-censored data based on a Weibull survival model and a two-group comparison via the Wald test using an expanded data set following Lyles et al. [29]. There are few power calculations for interval-censored data analysis and most methods for the usual right-censored scenario assume the more restrictive exponential distribution.

Simulation results demonstrate that our proposed power method performs well with a Weibull survival model or Sun, Zhao, and Zhao's [6] generalized log-rank test. Our method also performs well under the simple missing visit scenario considered here. Our approach is easily extended to other parametric survival and censoring distributions, and other tests such as the likelihood ratio test. It can also be extended to study designs with more than 2 groups and designs with different dropout patterns between the groups. Moreover, the requirement that scheduled visits are equally spaced out can also be relaxed. For example, to allow for more flexible visit times one can apply a pre-established algorithm to better mimic the timing of study visits in practice. One such algorithm is similar to a split-plot design. First one can divide the number of visits into blocks. Then permutations of '+', 'o', and '-' are produced within each block. The three signs correspond to add, don't change, or subtract 5% of the interval lengths to the initially scheduled time. There are six possible permutations: each subject will start with a permutation type and then the remaining permutation types will be sequentially assigned to each of the other blocks. For example, the first subject would start with the first permutation type for the first block, the second permutation type for the second block, and so forth. The second subject would start with the second permutation type for the first block, the third permutation type for the second block, and so forth. The process of assigning permutation types can be rotated again across and within subjects if necessary. We would use the newly specified visit times to generate the weights for the follow-up intervals based on this algorithm. The proposed power calculations are dependent upon correct specification of the survival and censoring distributions regardless of how one modifies the study design.

Although we trust the accuracy of power calculations using a simulation approach, appropriately generating response data in some cases is significantly more challenging than applying the proposed technique [29]. One may encounter convergence difficulties when approximating power via simulation under more specialized models or with smaller sample sizes. Glueck and Muller [40] also hesitated to recommend simulation as a general solution for power approximation. These power and sample-size programs are written in SAS IML [33] and are available from the authors.

References

1. So Y, Johnston G. Analyzing interval-censored survival data with SAS software. *SAS Global Forum.* SAS Institute, Inc.: Cary, NC, 2010; Paper 257-2010.

2. Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society, Series A* 1972; **135**:185-207.

3. Finkelstein DM. A proportional hazards model for interval-censored failure time data. *Biometrics* 1986; **42**:845-854.

4. Sun J. A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* 1996; **15**:1387-1395.

5. Zhao Q, Sun J. Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine* 2004; **23**:1621-1629.

6. Sun J, Zhao Q, Zhao X. Generalized log-rank test for interval-censored failure time data. *Scandinavian Journal of Statistics* 2005; **32**:49-57.

7. Fay MP. Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* 1996; **52**:811-822.

8. Satten GA. Rank-based inference in the proportional hazards model for interval censored data. *Biometrika* 1996; **83**:355-370.

9. Cox DR. Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 1972; **20**:187-220.

10. Cox DR. Partial likelihood. *Biometrika* 1975; **62**:269-276.

11. Satten GA, Datta S, Williamson JM. Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association* 1998; **93**:318-327.

12. Heller G. Proportional hazards regression with interval censored data using an inverse probability weight. *Lifetime Data Analysis* 2011; **17**:373-385.

13. Lin X, Cai B, Wang L, Zhang Z. A Bayesian proportional hazards model for general interval-censored data. *Lifetime Data Analysis* 2015; **21**:470-490.

14. Huang J. Efficient estimation for the proportional type I interval censored data. *The Annals of Statistics* 1996; **24**:540-568.

15. Diamond ID, McDonald JW, Shah IH. Proportional hazards models for current-status data: application to the study of differentials in age at weaning in Pakistan. *Demography* 1986; **23**:607-620.

16. Grummer-Strawn LM. Regression analysis of current-status data: An application to breast-feeding. *Journal of the American Statistical Association* 1993; **88**:758-765.

17. Shiboski SC. Generalized additive models for current status data. *Lifetime Data Analysis* 1998; **4**:29-50.

18. Pfeiffer RM, Mbulaiteye S, Engels E. A model to estimate risk of infection with human herpes virus 8 associated with transfusion from cross-sectional data. *Biometrics* 2004; **60**:249-256.

19. Sun J. *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer Science+Business Media, Inc.: New York, 2006.

20. Murphy SA, van der Vaart AW. Semiparametric likelihood ratio inference. *Annals of Statistics* 1997; **25**:1471-1509.

21. Banerjee M. Likelihood ratio tests under local alternatives in regular semiparametric models. *Statistica Sinica* 2005; **15**:635-644.

22. Williamson JM, Lin H-M, Kim H-Y. Power and sample size calculations for

current status survival analysis. *Statistics in Medicine* 2009; **28**:1999-2011.

23. Marschner IC. Determining the size of a cross-sectional sample to estimate the age-specific incidence of an irreversible disease. *Statistics in Medicine* 1994; **13**:2369-2381.

24. Zhao X, Duan R, Zhao Q, Sun J. A new class of generalized log rank tests for interval-censored failure time data. *Computational Statistics and Data Analysis* 2013; **60**:123-131.

25. O'Brien RG. Using the SAS system to perform power analyses for log-linear models. *Proceedings of the Eleventh Annual SAS Users Group International Conference.* SAS Institute, Inc.: Cary, NC, 1986; 778-782.

26. O'Brien RG, Muller KE. Unified power analysis for t-tests through multivariate hypotheses. In *Applied Analysis of Variance in Behavioral Science.* Edwards LK (ed.). Marcel Dekker: New York, 1993; 297-344.

27. O'Brien RG. A tour of UnifyPow: a SAS module/macro for sample size analysis. *Proceedings of the Twenty-Third Annual SAS Users Group International Conference.* SAS Institute, Inc.: Cary, NC, 1998; 778-782.

28. Castelloe JM, O'Brien RG. Power and sample size determination for linear models. *Proceedings of the Twenty-Sixth Annual SAS Users Group International Conference.* SAS Institute, Inc.: Cary, NC, 1998; Paper 240-26.

29. Lyles RH, Lin H-M, Williamson JM. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine* 2007; **26**:1632-1648.

30. SAS Institute, Inc. *SAS/STAT(R) 9.3 User's Guide.* SAS Institute, Inc.: Cary, North Carolina, 2011.

31. Finkelstein DM, Wolfe RA. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* 1985; **41**:933-945.

32. Wald A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 1943; **54**:426-482.

33. SAS Institute, Inc. *SAS/IML Software: Changes and Enhancements through Release 12.1.* SAS Institute, Inc.: Cary, North Carolina, 2012.

34. Raab GM, Davies JA, Salter AB. Designing follow-up intervals. *Statistics in Medicine* 2004; **23**:3125-3137.

35. Parmigiani G. Designing observation times for interval censored data. *Sankyha A* 1998; **60**:446–458.

36. Inoue LYT, Parmigiani G. Designing follow-up times. *Journal of American Statistical Association* 2002; **97**:847–858.

37. Cheng KF, Chen CH. Estimation of the Weibull parameters with grouped data. *Communications in Statistics, Theory and Methods* 1988; **17**:325-341.

38. Rao AV, Rao AVD, Narasimham VL. Asymptotically optimal groupings for maximum likelihood estimation of Weibull parameters. *Communications in Statistics, Simulation and Computation* 1994; **23**:1077-1096.

39. Guo C, So Y, Johnston G. Analyzing interval-censored survival data with the ICLIFETEST procedure. *SAS Global Forum.* SAS Institute, Inc.: Cary, NC, 2014; Paper 279-2014.

40. Glueck DH, Muller KE. Adjusting power for a baseline covariate in linear models. *Statistics in Medicine* 2003; **22**:2535-2551.

Table 1: An illustrative example of our expanded data set for two individuals whose first follow-up visits are scheduled at 3.50 and 4.00 months after study onset, assuming that the investigator specified $N = 200$, $\beta = \log(1.3)$, $\gamma = 1$, 10% dropout at 24 months, and a 90% failure rate for the $x = 0$ group by 24 months, with 6 visits (every 4 months) and no missing visits ($p = 0.0$).

ID	x	l	u		weight (w_i)
1	0	0	∞	Dropout before 1st visit	$1 - \Pr(C_i \geq 3.5)$[a] $= 0.014$
1	0	0	3.5	Failed in 1st interval	$\Pr(T_i \leq 3.5)\Pr(C_i > 3.5) = 0.281$
1	0	3.5	∞	Dropout after 1st visit	$\Pr(T_i > 3.5)$[b] $\Pr(3.5 < C_i \leq 7.5) = 0.012$
1	0	3.5	7.5	Failed in 2nd interval	$\Pr(3.5 < T_i \leq 7.5)\Pr(C_i > 7.5) = 0.221$
1	0	7.5	∞	Dropout after 2nd visit	$\Pr(T_i > 7.5)\Pr(7.5 < C_i \leq 11.5) = 0.008$
1	0	7.5	11.5	Failed in 3rd interval	$\Pr(7.5 < T_i \leq 11.5)\Pr(C_i > 11.5) = 0.148$
1	0	11.5	∞	Dropout after 3rd visit	$\Pr(T_i > 11.5)\Pr(11.5 < C_i \leq 15.5) = 0.005$
1	0	11.5	15.5	Failed in 4th interval	$\Pr(11.5 < T_i \leq 15.5)\Pr(C_i > 15.5) = 0.099$
1	0	15.5	∞	Dropout after 4th visit	$\Pr(T_i > 15.5)\Pr(15.5 < C_i \leq 19.5) = 0.004$
1	0	15.5	19.5	Failed in 5th interval	$\Pr(15.5 < T_i \leq 19.5)\Pr(C_i > 19.5) = 0.066$
1	0	19.5	∞	Dropout after 5th visit	$\Pr(T_i > 19.5)\Pr(19.5 < C_i \leq 23.5) = 0.003$
1	0	19.5	23.5	Failed in 6th interval	$\Pr(19.5 < T_i \leq 23.5)\Pr(C_i > 23.5) = 0.044$
1	0	23.5	∞	Censored at study end	$\Pr(T_i > 23.5)\Pr(C_i > 23.5) = 0.095$
					Sum of weights = 1.000

ID	x	l	u		weight (w_i)
151	1	0	∞	Dropout before 1st visit	$1 - \Pr(C_i \geq 4.0)$[a] $= 0.016$
151	1	0	4.0	Failed in 1st interval	$\Pr(T_i \leq 4.0)\Pr(C_i > 4.0) = 0.252$
151	1	4.0	∞	Dropout after 1st visit	$\Pr(T_i > 4.0)$[b] $\Pr(4.0 < C_i \leq 8.0) = 0.012$
151	1	4.0	8.0	Failed in 2nd interval	$\Pr(4.0 < T_i \leq 8.0)\Pr(C_i > 8.0) = 0.184$
151	1	8.0	∞	Dropout after 2nd visit	$\Pr(T_i > 8.0)\Pr(8.0 < C_i \leq 12.0) = 0.009$
151	1	8.0	12.0	Failed in 3rd interval	$\Pr(8.0 < T_i \leq 12.0)\Pr(C_i > 12.0) = 0.135$
151	1	12.0	∞	Dropout after 3rd visit	$\Pr(T_i > 12.0)\Pr(12.0 < C_i \leq 16.0) = 0.007$
151	1	12.0	16.0	Failed in 4th interval	$\Pr(12.0 < T_i \leq 16.0)\Pr(C_i > 16.0) = 0.099$
151	1	16.0	∞	Dropout after 4th visit	$\Pr(T_i > 16.0)\Pr(16.0 < C_i \leq 20.0) = 0.005$
151	1	16.0	20.0	Failed in 5th interval	$\Pr(16.0 < T_i \leq 20.0)\Pr(C_i > 20.0) = 0.072$
151	1	20.0	∞	Dropout after 5th visit	$\Pr(T_i > 20.0)\Pr(20.0 < C_i \leq 24.0) = 0.004$
151	1	20.0	24.0	Failed in 6th interval	$\Pr(20.0 < T_i \leq 24.0)\Pr(C_i > 24.0) = 0.053$
151	1	24.0	∞	Censored at study end	$\Pr(T_i > 24.0)\Pr(C_i > 24.0) = 0.153$
					Sum of weights = 1.000

[a] $\Pr(C_i \geq c_i) = 1 - (c_i/24)\psi$, where ψ (dropout rate at 24 months) = 0.1

[b] $\Pr(T_i \geq t_i) = \exp(-e^{(t_i - \Delta - \beta x_i)\gamma})$, where $\Delta = \log(24) - \log(-\log(1 - 0.90)) = 2.344$, $\beta = \log(1.3)$ and $\gamma = 1$.

Table 2: Power calculations for a two-group comparison with 6 visits (every 4 months) and varying β (effect size) values, γ (shape) values, and missing visit probabilites (p). The log-transformed failure times are generated with a Weibull distribution as follows: $Pr(T_i \geq t_i) = S(t_i; \Delta, \beta) = exp(-e^{(t_i-\Delta-\beta x_i)\gamma})$ with Δ varying according to the censoring percentage for the $x = 0$ group at 24 months. Dropout varies by 10%, 20%, to 30% uniformly by 24 months for both groups. Data sets are of total size N with equal group sizes. Calculated power is in bold. Empirical power for each analysis method is presented below and is based on 5,000 simulated data sets.

			γ, Shape								
Censoring	p	Analysis	0.5			1.0			1.5		
		Method	β, Effect Size								
			$log(1.3)$	$log(1.5)$	$log(1.7)$	$log(1.3)$	$log(1.5)$	$log(1.7)$	$log(1.3)$	$log(1.5)$	$log(1.7)$
Light[a]	Ave. % of failures		85.9%	84.9%	83.9%	82.9%	80.6%	78.5%	79.9%	76.4%	73.1%
	$p = 0.00$		**0.306**	**0.605**	**0.824**	**0.390**	**0.725**	**0.909**	**0.510**	**0.842**	**0.962**
		Weibull[b]	0.300	0.611	0.826	0.402	0.736	0.909	0.514	0.853	0.973
		SZZ[c]	0.301	0.610	0.828	0.399	0.732	0.906	0.507	0.851	0.971
	$p = 0.40$		**0.295**	**0.585**	**0.805**	**0.374**	**0.704**	**0.895**	**0.483**	**0.814**	**0.947**
		Weibull	0.305	0.602	0.828	0.377	0.709	0.900	0.505	0.834	0.961
		SZZ	0.301	0.601	0.824	0.369	0.707	0.894	0.499	0.830	0.959
Medium[d]	Ave. % of failures		63.0%	61.7%	60.6%	58.8%	56.4%	54.4%	55.4%	52.0%	49.3%
	$p = 0.00$		**0.277**	**0.548**	**0.766**	**0.354**	**0.665**	**0.862**	**0.467**	**0.783**	**0.923**
		Weibull	0.279	0.552	0.774	0.359	0.676	0.879	0.460	0.800	0.948
		SZZ	0.283	0.556	0.778	0.362	0.675	0.878	0.472	0.807	0.950
	$p = 0.40$		**0.268**	**0.531**	**0.747**	**0.338**	**0.641**	**0.842**	**0.436**	**0.745**	**0.896**
		Weibull	0.279	0.548	0.768	0.335	0.650	0.846	0.443	0.772	0.924
		SZZ	0.280	0.547	0.770	0.341	0.657	0.850	0.449	0.781	0.928
Heavy[e]	Ave. % of failures		42.4%	41.4%	40.5%	38.5%	36.7%	35.2%	35.5%	33.1%	31.2%
	$p = 0.00$		**0.225**	**0.446**	**0.650**	**0.289**	**0.556**	**0.760**	**0.397**	**0.688**	**0.849**
		Weibull	0.227	0.461	0.644	0.289	0.551	0.766	0.390	0.712	0.891
		SZZ	0.228	0.468	0.648	0.290	0.553	0.766	0.406	0.731	0.900
	$p = 0.40$		**0.217**	**0.429**	**0.628**	**0.274**	**0.528**	**0.732**	**0.367**	**0.643**	**0.808**
		Weibull	0.220	0.446	0.645	0.271	0.526	0.734	0.363	0.666	0.861
		SZZ	0.227	0.453	0.652	0.283	0.544	0.752	0.381	0.687	0.872

[a] Data sets are of size $N = 600$, $N = 200$, and $N = 130$ for $\gamma = 0.5$, $\gamma = 1.0$ and $\gamma = 1.5$, respectively.

[b] Weibull refers to a parametric Weibull survival model.

[c] SZZ refers to Sun, Zhao and Zhao's (2005) generalized log-rank test.

[d] Data sets are of size $N = 700$, $N = 250$, and $N = 170$ for $\gamma = 0.5$, $\gamma = 1.0$ and $\gamma = 1.5$, respectively.

[e] Data sets are of size $N = 800$, $N = 300$, and $N = 220$ for $\gamma = 0.5$, $\gamma = 1.0$ and $\gamma = 1.5$, respectively.

Table 3: Power calculations for a two-group comparison varying the number of visits and effect size (β) with a total sample of size N and equal group sizes. The study is assumed to be 24 months long with no missing visits ($p = 0.0$). The log-transformed failure times are generated with an exponential distribution as follows: $Pr(T_i \geq t_i) = S(t_i; \Delta, \beta) = exp(-e^{(t_i - \Delta - \beta x_i)})$ with Δ and study dropout percentages varying.

		Effect Size (β)		
		$log(1.3)$	$log(1.5)$	$log(1.7)$
			Censoring	
Number of visits	Interval length (months)	Light[a]	Medium[b]	Heavy[c]
1[d]	24	0.282	0.582	0.676
2	12	0.359	0.640	0.731
3	8	0.377	0.654	0.747
4	6	0.384	0.660	0.754
6	4	0.390	0.665	0.760
8	3	0.392	0.668	0.764
12	2	0.393	0.670	0.767
24	1	0.395	0.672	0.770

[a] Light censoring refers to 10% dropout at 24 months, and a 90% failure rate for the $x = 0$ group at 24 months ($\Delta = \log(24) - \log(-\log(1 - 0.90)) = 2.344$). Data sets are of size $N = 200$.

[b] Medium censoring refers to 20% dropout at 24 months, and a 70% failure rate for the $x = 0$ group at 24 months ($\Delta = \log(24) - \log(-\log(1 - 0.70)) = 2.992$). Data sets are of size $N = 250$.

[c] Heavy censoring refers to 30% dropout at 24 months, and a 50% failure rate for the $x = 0$ group at 24 months ($\Delta = \log(24) - \log(-\log(1 - 0.50)) = 3.545$). Data sets are of size $N = 300$.

[d] Corresponds to current-status data.

Table 4: Total sample sizes for a two-group comparison varying power, missing visit probability and effect size (hazard ratio) with equal group sizes. Assume an exponential survival distribution with 40% of the women in the $x = 0$ group not having the event by study end of 48 months, with scheduled visits every 6 months, and 5% of the enrolled dropping out each year.

	Missing visit probability					
	0.0		0.2		0.4	
	Power					
Hazard Ratio	80%	90%	80%	90%	80%	90%
1.50	318	426	324	434	332	442
1.75	162	218	166	222	170	226
2.00	104	138	106	142	108	144
2.25	74	100	76	102	78	104
2.50	58	78	60	80	60	80