Published in final edited form as:

Epidemiol Methods. 2019 December ; 8(1): . doi:10.1515/em-2018-0011.

Modeling of Clinical Phenotypes Assessed at Discrete Study Visits

Emily J. Huang¹, Ravi Varadhan², Michelle C. Carlson³

¹Department of Mathematics and Statistics, Wake Forest University

²Division of Biostatistics and Bioinformatics, Johns Hopkins Sidney Kimmel Comprehensive Cancer Center

³Department of Mental Health, Johns Hopkins University, Bloomberg School of Public Health

Abstract

In studies of clinical phenotypes, such as dementia, disability, and frailty, participants are typically assessed at in-person clinic visits. Thus, the precise time of onset for the phenotype is unknown. The discreteness of the clinic visits yields grouped event time data. We investigate how to perform a risk factor analysis in the case of grouped data. Since visits can be months to years apart, numbers of ties can be large, causing the exact tie-handling method of the Cox model to be computationally infeasible. We propose two, new, computationally efficient approximations to the exact method: Laplace approximation and an analytic approximation. Through extensive simulation studies, we compare these new methods to the Prentice-Gloeckler model and the Cox model using Efron's and Breslow's tie-handling methods. In addition, we compare the methods in an application to a large cohort study ($N = 3,605$) on the development of clinical frailty in older adults. In our simulations, the Laplace approximation has low bias in all settings, and the analytic approximation has low bias in settings where the regression coefficient is not large in magnitude. Their corresponding confidence intervals also have approximately the nominal coverage probability. In the data application, the results from the approximations are nearly identical to that of the Prentice-Gloeckler model.

1 Introduction

Older adults are at increased risk of experiencing poor health outcomes, such as frailty, disability, and dementia. Identifying risk factors for these outcomes is a major goal of prospectively conducted epidemiological studies of aging. Many aging-related phenotypes can only be assessed at in-person study visits. For these phenotypes, a participant's time of onset cannot be known precisely. It is known up to the time interval between the study visit in which the outcome was first observed and the previous visit. This type of data in which the true time of onset is only known to be in a certain group is called grouped event time

Corresponding Author: Ravi Varadhan (ravi.varadhan@jhu.edu).

Supplementary Materials

Web Figure 1, referenced in Section 5, Web Figures 2–7, referenced in Section 6, and Web Tables 1–4, referenced in Section 6, are provided in the Supplementary Materials. The R code for the proposed methods, simulation studies, and CHS application is provided online at <https://github.com/emhuang1/groupedData>.

data, and is the focus of our paper. Grouped data arises whenever the outcome of interest is assessed at discrete study visits, so it is common not only in aging research but also in many other fields. Some examples of epidemiological studies that have collected grouped event data include the Cardiovascular Health Study, the Multiethnic Study of Atherosclerosis, and the Women's Health and Aging Study (Fried, Borhani, Enright, Furberg, Gardin, Kronmal et al., 1991, Bild, Bluemke, Burke, Detrano, Diez Roux, Folsom et al., 2002, Onder, Penninx, Ferrucci, Fried, Guralnik, and Pahor, 2005).

The structure of event times for grouped data differs from that of interval censored event times. In the grouped data setting, the intervals are fixed by the number and timing of visits given by the study design, while in the interval-censored data setting the intervals are random. In the grouped data setting, all the observed event times will cluster into a small number of groups, which would correspond to the study visits, whereas in the case of interval censored data the number of intervals can be quite large.

We are interested in examining how the grouped nature of the data affects risk factor analysis in prospective cohort studies. One modeling approach is the Prentice-Gloeckler (PG) model, which accounts for the grouped nature of the data (Prentice and Gloeckler, 1978). A straightforward, but less principled approach is to use the Cox proportional hazards (PH) model, the workhorse of survival analysis, substituting the unknown precise onset of the outcome with the visit time at which the outcome is first observed. The Cox model estimates the impact of covariates on the hazard function, which is based on continuous event time, whereas the observed event times in the grouped data setting are discrete. Due to the grouped nature of the event time, there will be a substantial number of ties in event times, and this needs to be addressed while using the Cox PH model. There are three well-known tie-handling techniques: the exact method, the method by Breslow, and the method by Efron. A challenge of the exact method is that it can be computationally intractable when there is a large number of ties in the data. We propose two new approximations to the exact method that are computationally efficient, even for a very large number of tied failure times. The first approximation is the Laplace approximation to an integral (Erdelyi, 1956, p. 36-37), and the second is an analytic approximation to the integral.

Through simulation, we compare our two approximations to the PG model, and to the Cox PH model using Breslow's and Efron's tie-handling schemes. In general, the proposed methods yield estimators with low bias, and confidence intervals with the desired coverage probability. Also, we apply the proposed methods to the Cardiovascular Health Study (Fried et al., 1991), a cohort study on the development of clinical frailty in older adults. The methods give comparable results to the Prentice-Gloeckler model, in our data application. The R code for the proposed methods is provided online. The link can be found in the Supplementary Materials section at the end of this paper. Issues of competing events, although relevant given the type of events considered in our motivating example, are not addressed in this paper.

Currently, Breslow's approximation is the default method for handling tied data in most software packages, such as Stata and SAS. Hertz-Picciotto and Rockhill (1997) and Scheike and Sun (2007) found that Breslow's approximation tends to underestimate the coefficients

in the Cox model and that Efron's approximation performs better than Breslow's approximation. The simulation results of this paper are consistent with those findings.

This paper is organized as follows. The assumptions are stated in Section 2. The PG model and the Cox PH Model, including the exact tie-handling method and the tie-handling methods by Breslow and Efron, are introduced in Section 3. We present our proposed approximations to the exact method in Section 4. The data application to the Cardiovascular Health Study and the simulation studies are described in Sections 5 and 6, respectively. Concluding remarks are given in Section 7.

2 Context and Assumptions

We consider a population whose members have not experienced the event of interest at baseline ($t = 0$), which is entry into the study. For each member, we assume her/his hazard function $\lambda_{\beta}(t|Z)$ satisfies:

$$\lambda_{\beta}(t|Z) = \lambda_0(t)\exp(Z_1\beta_1 + Z_2\beta_2 + \cdots + Z_p\beta_p), \quad (1)$$

which is the Cox proportional hazards assumption. In (1), $Z = (Z_1, Z_2, \dots, Z_p)$ is a vector of covariates collected at baseline, and $\beta = (\beta_1, \dots, \beta_p)$ is a vector of unknown coefficients corresponding to Z . The function of time, $\lambda_0(t)$, is the hazard function corresponding to $Z = (0, \dots, 0)$. The Cox model assumes that the event time T is absolutely continuous. Our objective is to estimate the coefficient vector β , using data from a prospective cohort study where the participants are assessed at in-person visits. During the baseline visit ($t = 0$), the covariates Z are collected for each participant. Let Z_i denote the vector of baseline covariates $(Z_{i1}, Z_{i2}, \dots, Z_{ip})$ for participant i . After the baseline visit, follow-up visits are used to determine whether or not participants have experienced the event of interest. We assume that the spacing between visits is uniform across participants, though the visits themselves need not occur at regular intervals. Let m denote the number of follow-up visits. We label the baseline visit as Visit 0, and the follow-up visits as Visits 1 through m . Let t_k denote the length of time between Visit 0 and Visit k . Missing data can result due to early drop-out. However, we assume no intermittent missing data, i.e., no cases where a participant missed Visit v but attended Visit v' , where $v' > v$.

For any given participant i , the observed data vector is (Z_i, δ_i, V_i) , where Z_i is the vector of covariates, δ_i is an indicator of censoring, and V_i is an integer between 0 and m . A participant is censored if the study ended or the participant dropped out before she/he experienced the event. The censoring indicator δ_i takes the value 0 if participant i is censored, and 1 otherwise. If $\delta_i = 0$, V_i is the last visit that participant i attended. If $\delta_i = 1$, V_i is the first visit where participant i was observed to have experienced the event, and we refer to V_i as the "observed event time." From the observed data, the true (i.e., precise) event time is unknown, but it is known to occur between Visits $V_i - 1$ and V_i . We make two assumptions about the observed data. First, the data vectors (Z_i, δ_i, V_i) , $i = 1, \dots, n$, are independent and identically distributed from a true, unknown joint distribution P on (Z, δ, V) . Second, for participants who are censored due to dropout, their dropout time is independent of their true event time, conditional on the observed covariates.

A common attribute of grouped data is heavy ties in the observed event times. This occurs because adjacent visits can be months to years apart. For example, in the Cardiovascular Health Study, the first follow-up assessment of the clinical frailty outcome occurred three years after the baseline visit. There were 307 participants whose observed event time was the first follow-up visit, and 187 participants whose observed event time was the second follow-up visit. Define I_k as the set of participants who are tied at Visit k , i.e., $I_k = \{i: \delta_i = 1, V_i = k\}$. Let d_k denote the number of participants tied at Visit k , i.e., the number of elements in the set I_k .

We compare the performance of various approaches to estimate β . Existing approaches include the Prentice-Gloeckler model and the Cox model with tie-handling using the exact method, Breslow's method, or Efron's method. We propose two new, computationally efficient approximations to the exact method.

3 Existing Approaches

3.1 Prentice-Gloeckler Model

Prentice and Gloeckler (1978) apply maximum likelihood to estimate β , while accounting for the grouped nature of the data. When deriving the likelihood function, they make the Cox proportional hazards assumption. Let $t_0 = 0$ and $t_{m+1} = \infty$. For any given $i = 1, \dots, n$, let T_i be the unobserved continuous event time of participant i . Let T'_i be the first visit where participant i would be observed to have experienced the event under no censoring, i.e.,

$$T'_i = v \text{ if and only if } T \in (t_{v-1}, t_v].$$

We refer to T'_i as the “grouped event time.” The grouped event time T'_i is defined for every participant, but is only observed for uncensored participants. Throughout the rest of the paper, we use the dot symbol $\cdot$ to represent dot products, i.e., $a \cdot b$ is the dot product of vectors a and b . The likelihood function derived by Prentice and Gloeckler (1978) is:

$$L(\beta) = \prod_{i=1}^n [P(T'_i = v_i | Z_i)]^{\delta_i} [P(T'_i > v_i | Z_i)]^{1 - \delta_i}, \quad (2)$$

where

$$P(T'_i = v_i | Z_i) = P(t_{v_i-1} < T_i \leq t_{v_i} | Z_i) = (1 - \alpha_{v_i}^{\exp\{Z_i\beta\}}) \prod_{j=1}^{v_i-1} \alpha_j^{\exp\{Z_i\beta\}}, \quad (3)$$

$$\alpha_j = \exp \left\{ - \int_{t_{j-1}}^{t_j} \lambda_0(u) du \right\}, \quad (4)$$

$$P(T_i' = v_i | Z_i) = P(T_i > t_{v_i} | Z_i) = \prod_{j=1}^{v_i} \alpha_j^{\exp\{Z_i \beta\}}. \quad (5)$$

The PG model is equivalent to using a generalized linear model of the binomial family with a complementary log-log link (Allison, 1982). The Cox model and PG model both make the proportional hazards assumption and they do not require the user to assume a specific form for the baseline hazard λ_0 . A distinction between the models is that the former does not parametrically specify the baseline hazard function, while the latter does through the α_j terms defined in (4). In the PG model, the α_j terms are estimated from data without requiring assumptions about the functional form of $\lambda_0(t)$. The PG model is not appropriate if the number of parameters is close to the number of uncensored subjects, which occurs if the visits are frequent resulting in small tie sizes (Therneau and Grambsch, 2010).

3.2 Cox Model with Existing Tie-handling Methods

Let $R_i(\beta) = \exp(Z_i \cdot \beta)$ be the risk score of participant i . Let the indicator function $1(E)$ equal 1 if event E is true and 0 otherwise. In the case of no ties in the observed event times, the vector β can be estimated as the maximizer of the partial likelihood function, which is

$$PL(\beta) = \prod_{i: \delta_i = 1} \frac{R_i(\beta)}{\sum_{j=1}^n 1(V_j \geq V_i) R_j(\beta)}. \quad (6)$$

However, ties are common in our context of grouped data, as discussed in Section 2. We present the existing methods for tie-handling using a simple example. Suppose that participants 1 and 2 are tied at Visit k . Their contributions to the partial likelihood equal either:

$$\frac{R_1}{R_1 + R_2 + \sum_{j=3}^n 1(V_j \geq k) R_j} \cdot \frac{R_2}{R_2 + \sum_{j=3}^n 1(V_j \geq k) R_j}, \quad (7)$$

if participant 1 experienced the event before participant 2, or

$$\frac{R_2}{R_1 + R_2 + \sum_{j=3}^n 1(V_j \geq k) R_j} \cdot \frac{R_1}{R_1 + \sum_{j=3}^n 1(V_j \geq k) R_j}, \quad (8)$$

if participant 2 experienced the event before participant 1. Both (7) and (8) have the term $\sum_{j=3}^n 1(V_j \geq k) R_j$ in the first and second denominators; the subjects contributing to this term are those who are censored at or after Visit k or have an observed event time after Visit k . The difference between (7) and (8) arises from the presence of R_1 or R_2 in the second denominator. Since participants 1 and 2 have the same observed event times, we do not know which is correct.

In the exact approach, one takes the average of (7) and (8) (Kalbfleisch and Prentice, 2002). The exact approach is computationally feasible when the number of tied subjects d_k is small,

as in this case. However, it is computationally prohibitive for even moderately large d_k since the number of possibilities to average over, $d_k!$, becomes massive. For example, when the tie size d_k is 10, the number of possibilities is $10! = 3,628,800$. For the frailty outcome in the Cardiovascular Health Study, we encounter 307 and 187 ties in the two follow-up visits. The exact method is computationally impossible. Motivated by this, we propose two computationally efficient approximations of the exact approach in Section 4.

Commonly used tie-handling methods in the Cox PH model are those proposed by Breslow and Efron (Therneau and Grambsch, 2010). Rather than taking the average of the $d_k!$ possibilities, Breslow's and Efron's methods each involve taking a product of two terms, as in (7) and (8). In Breslow's method, the sum of the risk scores of participants 1 and 2 appears in both denominators, i.e.,

$$\frac{R_1}{R_1 + R_2 + \sum_{j=3}^n 1(V_j \geq k)R_j} \cdot \frac{R_2}{R_1 + R_2 + \sum_{j=3}^n 1(V_j \geq k)R_j}. \quad (9)$$

In Efron's method, the full sum is in the first denominator, while the second denominator uses the average of R_1 and R_2 , i.e.,

$$\frac{R_1}{R_1 + R_2 + \sum_{j=3}^n 1(V_j \geq k)R_j} \cdot \frac{R_2}{0.5(R_1 + R_2) + \sum_{j=3}^n 1(V_j \geq k)R_j}. \quad (10)$$

More generally, if there are d_k participants tied at Visit k , the exact method will take the average of all the $d_k!$ possibilities, whereas Breslow's and Efron's methods only involve taking a product of d_k terms. Without loss of generality, assume that the d_k participants tied at Visit k are participants 1, 2, ..., d_k . Then the contribution to the partial likelihood at Visit k , for Breslow's and Efron's methods, respectively, can be written as follows:

$$\prod_{k=1}^{d_k} \frac{R_k}{\sum_{j=1}^{d_k} R_j + \sum_{j=d_k+1}^n 1(V_j \geq k)R_j}, \quad (11)$$

$$\prod_{k=1}^{d_k} \frac{R_k}{\left(\frac{d_k - k + 1}{d_k}\right) \sum_{j=1}^{d_k} R_j + \sum_{j=d_k+1}^n 1(V_j \geq k)R_j}. \quad (12)$$

4 Proposed Approximations to the Exact Tie-Handling Method

As discussed in Section 3, computing the exact likelihood is difficult when the tie sizes are not small. For any given participant i and Visit k , where $k \in \{1, \dots, m\}$, define

$$a_{i,k}(\beta) = \frac{R_i(\beta)}{\sum_{j \notin I_k} 1(V_j \geq k)R_j(\beta)}. \quad (13)$$

For conciseness, we refer to this as $a_{i,k}$ throughout the rest of the paper. DeLong, Guirguis, and So (1994) showed that the contribution to the exact likelihood at Visit k (i.e., the average of the $d_k!$ possibilities for Visit k) can be represented as $d_k!$ multiplied by

$$\int_{t=0}^{\infty} \left[\prod_{i \in I_k} (1 - e^{-a_{i,k}t}) \right] e^{-t} dt. \quad (14)$$

Our proposed method is to approximate (14), using two new approaches: the Laplace approximation and a simple analytic approximation. We also considered a Gaussian quadrature approach, the Gauss-Laguerre quadrature (Davis and Rabinowitz, 1984, p. 222-223). However, it did not work as well as the other approaches discussed here, mainly due to the difficulty in choosing the appropriate number of quadrature points (results not shown).

4.1 Laplace Approximation

Our first approximation is based on the Laplace approximation to an integral (Erdelyi, 1956) and will be referred to as “Laplace” throughout the paper. Here, we present the procedure for Laplace. The mathematical derivations are presented in Appendix A. Denote the integrand of (14) as $P^*(t)$, i.e.,

$$P^*(t) = \left[\prod_{i \in I_k} (1 - e^{-a_{i,k}t}) \right] e^{-t}.$$

Let t_k^* be the value of t such that

$$\sum_{i \in I_k} \frac{a_{i,k}}{e^{a_{i,k}t} - 1} = 1.$$

Define

$$\sigma_k^* = \sum_{i \in I_k} \frac{a_{i,k}^2 e^{a_{i,k}t_k^*}}{(e^{a_{i,k}t_k^*} - 1)^2}.$$

Let X_k be a Normally distributed random variable with mean t_k^* and variance $1 / \sigma_k^*$. We estimate the unknown coefficient vector β as the value of β that maximizes the function

$$\prod_{k=1}^m \left[P^*(t_k^*) \sqrt{2\pi / \sigma_k^*} \Pr(X_k > 0) \right]. \quad (15)$$

This is equivalent to finding the maximizer of the function

$$f(\beta) = \sum_{k=1}^m [\log P^*(t_k^*) - 1/2 \log \sigma_k^* + \log \Pr(X_k > 0)],$$

which is obtained by taking the logarithm of (15) and omitting the terms that do not depend on β . We denote the Laplace estimator of β as $\hat{\beta}_{Laplace}$. For any given matrix M , let $[M]_{ij}$ denote the element in the i -th row and j -th column of M , and let $(M)^{-1}$ denote the inverse of M . Let $H_f(\beta)$ denote the Hessian of $f(\beta)$, i.e., for any given i, j , $[H_f(\beta)]_{ij} = \partial^2 f(\beta) / \partial \beta_i \partial \beta_j$. The standard error of the i -th element in $\hat{\beta}_{Laplace}$ is estimated by: $\sqrt{[-H_f(\hat{\beta}_{Laplace})]^{-1}}_{ii}$.

4.2 Analytic Approximation

We propose another approximation that will be referred to as “Analytic,” in which (14) is approximated as

$$\sqrt{2\pi d_k} e^{-(d_k+1)} \prod_{i \in I_k} (1 - e^{-(d_k+1)a_{i,k}}), \quad (16)$$

which is a function of β due to the term $a_{i,k}$ that was defined in (13). The details of the derivation are provided in Appendix B. The unknown coefficient vector β is estimated by the value of β that maximizes the function

$$\prod_{k=1}^m \left[\sqrt{2\pi d_k} e^{-(d_k+1)} \prod_{i \in I_k} (1 - e^{-(d_k+1)a_{i,k}}) \right], \quad (17)$$

which is the product of (16) over $k = 1, \dots, m$. This is equivalent to maximizing the function

$$g(\beta) = \sum_{k=1}^m \sum_{i \in I_k} \log (1 - e^{-(d_k+1)a_{i,k}}), \quad (18)$$

which results from taking the logarithm of (17) and omitting the terms that do not depend on β . We denote the Analytic estimator of β as $\hat{\beta}_{Analytic}$. Let $H_g(\beta)$ denote the Hessian of $g(\beta)$. Analogous to the Laplace method, we estimate the standard error of the i -th element in $\hat{\beta}_{Analytic}$ by $\sqrt{[-H_g(\hat{\beta}_{Analytic})]^{-1}}_{ii}$.

5 Application to the Cardiovascular Health Study

The Cardiovascular Health Study (CHS) was a study of 5,201 older adults enrolled between 1989 and 1990, and followed up at annual clinic visits until 1999 (Fried et al., 1991). One of the outcomes collected in CHS was clinical frailty, which is a syndrome defined as the presence of three or more of the following: unintentional weight loss, exhaustion, low grip strength, slow walking speed, and low physical activity (Fried, Tangen, Walston, Newman, Hirsch, Gottdiener et al., 2001). Frailty status was assessed at the baseline visit and at two follow-up visits, occurring three and seven years after the baseline visit, respectively. We

fitted the Cox and PG models to time-to-frailty data collected in CHS. It should be noted that, in addition to the cohort of 5,201 participants, a second cohort of 687 participants was recruited in 1992. We focus on the first cohort in our analysis. This is because the follow-up schedule for frailty differs between the two cohorts. The second cohort had only a single follow-up assessment of frailty.

In our analysis of time-to-frailty, five covariates were used: age, gender (0 = female, 1 = male), whether the person had ever smoked (0 = no, 1 = yes), whether the person had received education beyond high school (0 = no, 1 = yes), and C-reactive protein (CRP) level. These covariates were collected at the baseline visit. We standardized both age and CRP level. Age, in units of years, was standardized by subtracting 65 and dividing by 10. CRP level, in units of mg/L, was standardized by dividing by the interquartile range (2.15 mg/L). The fitted models included main terms for the covariates and no interaction terms. We let δ be an indicator of whether the participant was assessed as frail during her/his follow-up (0 if no, 1 if yes). If $\delta = 0$, we let V be the last follow-up visit at which the participant's frailty status was assessed. If $\delta = 1$, we let V be the first follow-up visit at which the participant was assessed as frail.

We included only participants who satisfied the following conditions: (i) were non-frail at the baseline visit, (ii) frailty status was assessed at the first follow-up only, or at both follow-ups, (iii) covariates were non-missing. The resulting sample size was $n = 3,605$. The purpose for (i) was, since we sought to study the time-to-frailty development after the baseline visit, we focused on those who were not yet frail at the baseline visit. Criterion (ii) ensured that there was some follow-up on the frailty status after the baseline visit because those without any such follow-up would not affect the model fits. Also, criterion (ii) excluded those whose frailty status was observed at the second follow-up but not the first follow-up. If the status at the second follow-up was frail, then $\delta = 1$ but it could not be deduced whether V should be 2 or 1 (i.e., the person was actually frail by the first follow-up). If the status at the second follow-up was non-frail, there was also some ambiguity since the person might have been non-frail through the second follow-up, or have been frail at the first follow-up and transitioned to being non-frail by the second follow-up. A potential impact of excluding participants who did not attend the first follow-up visit is that the sample used in our analysis may under-represent a specific subtype of participants. For example, the participants who missed the first follow-up visit may be more likely to have severe health problems, e.g., some of them might have missed their study visit due to their health problems. Criterion (iii) was required since participants without the full set of covariates would not impact the model fits.

Web Figure 1 in the Supplementary Materials presents the empirical marginal distributions of age, gender, smoking status, education level, and CRP level, after excluding participants as described above. The mean age is 72 (range 64-95) and the proportion of female participants is 0.58. Table 1 presents the empirical joint distribution on (δ, V) , after excluding participants as described above. Each cell gives the number of participants, followed by the corresponding sample proportion in parentheses. We evaluated the validity of the proportional hazards assumption by testing the null hypothesis that the correlation between the scaled Schoenfeld residuals and time is zero, separately for each covariate. The

results supported the proportional hazards assumption; the p -value was 0.10 for age, 0.77 for education level, 0.48 for gender, 0.33 for smoking status, and 0.80 for CRP level.

Table 2 presents the regression coefficient estimates in the top panel, with the corresponding standard error estimates in the middle panel. The values are shown up to four decimal places. The results in this table and those throughout the paper were computed using R. When fitting the Cox model, we applied our proposed tie-handling methods Laplace and Analytic, as well as the existing methods by Efron and Breslow. When assessing each of these methods, we use PG as a reference since it accounts for the discreteness of the data directly. All of the methods were computationally efficient; in this data application, the running time was 0.01 seconds for Efron's method, 0.02 seconds for Breslow's method, 0.3 seconds for PG, 1.2 seconds for Analytic, and 1.4 seconds for Laplace.

In the upper two panels of Table 2, the results for Laplace perfectly match those for PG. For all covariates, the absolute difference between the Laplace estimate and the corresponding PG estimate was on the order of 10^{-6} or less. This applied to both the regression coefficient estimates and standard error estimates. For each covariate, the magnitude of the percent difference of Laplace compared to PG, i.e., $|(Laplace - PG)/PG \times 100|$, was at most 0.01%. The results for Analytic and Efron's method were also close to those of PG. For both Efron's method and Analytic, the absolute difference from PG was at most 6×10^{-3} in magnitude. For each covariate, the magnitude of the percent difference from PG was at most 0.8% for Efron's method and 0.6% for Analytic. Compared to PG and Laplace, the regression coefficients for Efron's method and Analytic were closer to the null. Those for Breslow's method were further in the direction of the null. Breslow's method yielded the most different results from PG. For the regression coefficient estimates, the magnitude of the percent difference of Breslow's method compared to PG was as much as 8%.

The bottom panel of Table 2 shows the 95% confidence intervals for the exponentiated regression coefficients computed using each approach. For interpretability, we present the confidence intervals for the exponentiated regression coefficients rather than for the regression coefficients themselves. Each confidence interval is computed by taking the corresponding regression coefficient estimate, adding and subtracting the quantity 1.96 multiplied by the standard error estimate, and exponentiating both values. Beside each 95% confidence interval, we also present the corresponding regression coefficient estimate (after exponentiation). Although there are some differences between the confidence intervals computed by the different approaches, the qualitative conclusions are the same. The various approaches all find that a person's hazard for frailty increases with higher age, holding the other factors constant. Also, the hazard for frailty is lower for those who have education beyond high school than for those who have a high school education or less. Females have a higher hazard than males, holding other factors constant. All of the approaches concluded that the hazard for frailty is not statistically different between those who have ever smoked versus those who have never smoked, but the point estimate suggests that smoking raises the hazard. The hazard for frailty increases with higher CRP level, holding other factors constant.

6 Simulation Studies

Through simulation, we evaluate the Cox and PG models in the grouped data setting. For the Cox model, the Laplace, Analytic, Efron's, and Breslow's tiehandling methods are separately applied. We compare the regression coefficient estimators with respect to bias, standard error, and root-mean-square error. Also, we evaluate the actual coverage probability of the nominal 95% confidence interval for each method.

Each simulation study includes a large number of simulations, $n_{sim} = 10,000$. For any given simulation s , a study with sample size n is simulated. The contents of the study data set are described in the following paragraph. Using the data set, the parameter β is estimated using the Cox model and the PG model, respectively. For a given estimator $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$, we approximate the bias, standard error, and root-mean-square error of the i -th element $\hat{\beta}_i$ as:

$$\text{bias} \approx \left(\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} \hat{\beta}_{s,i} \right) - \beta_i, \quad (19)$$

$$\text{standard error} \approx \sqrt{\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} \left(\hat{\beta}_{s,i} - \frac{1}{n_{sim}} \sum_{s'=1}^{n_{sim}} \hat{\beta}_{s',i} \right)^2}, \quad (20)$$

$$\text{root-mean-square error} \approx \sqrt{\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} (\hat{\beta}_{s,i} - \beta_i)^2}, \quad (21)$$

where $\hat{\beta}_{s,i}$ is the estimate of β_i in the s -th simulation. Also, we estimate the coverage probability of the 95% confidence interval for β_i by

$$\frac{1}{n_{sim}} \sum_{s=1}^{n_{sim}} 1(\hat{\beta}_{s,i} - 1.96\widehat{SE}_{s,i} \leq \beta_i \leq \hat{\beta}_{s,i} + 1.96\widehat{SE}_{s,i}),$$

where $\widehat{SE}_{s,i}$ is the estimate of the standard error of $\hat{\beta}_i$ from the s -th simulation.

Each simulation follows n participants according to a schedule which includes a baseline visit (Visit 0) and m follow-up visits (Visits 1, ..., m), occurring at fixed times after the baseline visit. The collected data set includes covariate, censoring, and time information of the n participants, designated as (Z_i, δ_i, V_i) , $i = 1, \dots, n$. These random variables were previously defined in Section 2. For any given i , participant i does not miss any visits preceding visit V_i . The simulation studies were organized into two sets, based on the structure of the data generating distribution. We refer to the two sets as the Single Covariate Simulations and the Data-Based Simulations.

6.1 Single Covariate Simulations

In these simulation studies, each simulation is a randomized controlled trial with $m = 3$ follow-up visits. The single covariate Z represents the treatment assignment (1 if control, 0 if treatment). Treatment assignment is random, with equal probability of being assigned to treatment or control, i.e., $P(Z = z) = 0.5$ for $z \in \{0, 1\}$.

The simulations also generate the censoring time C and the grouped event time T' for each participant, from which δ and V are derived. The censoring time C represents a participant's last visit before dropping out or being administratively censored. Each participant's censoring time is an independent draw from the discrete distribution having point masses at 0, 1, 2, and 3, with probabilities of 0.08, 0.1, 0.1, and 0.72, respectively. A participant with $C = 0$ attends only the baseline visit, while one with $C = 3$ continues till the end of the trial. Censoring time is independent of event time and treatment assignment.

The grouped event time T' is the first visit following the true continuous event time, which is unobserved and denoted as T . The value T' can be 1, 2, 3 or ∞ , where $T' = \infty$ indicates that the event occurs after the trial has ended and thus would not be observed at any of the visits. We generate the true event time T as a random draw from the continuous failure time distribution induced by the hazard function

$$\lambda(t|Z) = \left(\frac{t}{50}\right) \exp\{Z\beta\}. \quad (22)$$

We formulated (22) so that there would be approximately 25% with observed event times, 25% dropout, and 50% administrative censoring in each randomized trial, if $\beta = 0$. The baseline hazard function, i.e., (22) under $Z = 0$, is $\lambda_0(t) = t/50$. We let T have units of years, and define Visits 1, 2, and 3 to occur 2, 4, and 6 years after the baseline visit, respectively. Based on T , the grouped event time T' is derived as follows: 1 if $0 < T \leq 2$, 2 if $2 < T \leq 4$, 3 if $4 < T \leq 6$, and ∞ if $T > 6$.

By definition, the censoring indicator δ and visit number V are as follows:

$$\delta = 1(T' \leq C), \quad (23)$$

$$V = T'\delta + C(1 - \delta). \quad (24)$$

Note that, although T' can take on the value ∞ , the visit number V is at most 3 since $\delta = 0$ when $T' = \infty$. Administrative censoring occurs if $\delta = 0$ and $V = 3$, i.e., $C = 3$ and $T' = \infty$. We vary β between 0, 0.2, 0.4, 0.6, 0.8, 1, and 2. Also, we vary the sample size n between 200, 500, and 2000. For each β and sample size n , Table 3 shows the average number of ties at each visit. The average is taken across the 10,000 simulations for the given pair of β and n . For any given sample size and visit, the average number of ties grows with higher β (there are a few exceptions, namely for each sample size at Visit 3 when changing from $\beta = 1$ to 2). For instance, when $n = 200$, the average number of ties for Visit 1 is 7 if $\beta = 0$, 13 if $\beta = 1$, and 27 if $\beta = 2$. For a given β , the average number of ties grows with higher sample size n .

For instance, when $\beta = 0.2$, the average number of ties at Visit 3 is 24 for $n = 200$, 60 for $n = 500$, and 240 at $n = 2000$.

Table 4 presents the absolute bias, standard error, root-mean-square error, and coverage probability at $n = 500$. Web Tables 1 and 2 of the Supplementary Materials present the results for sample sizes $n = 200$ and 2000. For a visual representation, the tabulated results are also plotted in Web Figures 2–4 in the Supplementary Materials. These figures exclude the results for Breslow's method since it has larger bias than the other methods when $\beta = 0$.

Laplace and PG have absolute bias close to zero that also improves with higher sample size. At $n = 200$, the magnitude of the absolute bias is 0.02 for Laplace and 0.03 for PG, for each choice of β . At $n = 500$, it is 0.006 for Laplace and 0.01 for PG. Also, their coverage probabilities are approximately the nominal coverage probability of 0.95. For every pair of β and n , their coverage probability is within 0.006 of 0.95. Laplace and PG have very comparable standard error and root-mean-square error. Efron's method also performs well, which is consistent with the simulation results of Hertz-Picciotto and Rockhill (1997) and Scheike and Sun (2007) on tied data in the Cox model. Generally, Efron's method has bias close to zero and coverage probability close to 0.95. In addition, its standard error can be less than Laplace and PG. However, it does not perform as well as Laplace and PG in the case of $\beta = 2$. At $\beta = 2$, the bias of Efron's method is -0.03 at $n = 200, 500$, and 2000, while the bias of Laplace and PG becomes negligible with larger sample size. At $n = 2000$, Efron's method has a coverage probability of 0.92, while Laplace and PG have the correct coverage probability of 0.95. Analytic has the lowest standard error among Laplace, Efron's method, and PG. Its bias is close to zero for $\beta = 1$. For β in this range, the magnitude of the absolute bias for Analytic is 0.01 at $n = 200$, 0.02 at $n = 500$, and 0.02 at $n = 2000$. However, for $\beta = 2$, its bias is approximately -0.17 at $n = 200, 500$, and 2000.

Although it has the best standard error and negligible bias when $\beta = 0$, Breslow's method has the worst bias of all the methods when $\beta = 0$. In addition, its coverage probability can be low. For instance, at $n = 2000$, the coverage probability of Breslow's method is 0.85 when $\beta = 0.8$, 0.71 when $\beta = 1$, and almost 0 when $\beta = 2$. The results show that Breslow's method can be unsuitable when the number of ties is large. Intuitively, this is because the approximation used in Breslow's method, shown in (11), puts the risk score of each tied subject in every denominator, while in truth risk scores are omitted one at a time with each subsequent denominator. When the number of ties is large, this method may not yield a good approximation of the true partial likelihood.

This first set of simulation studies mimics a randomized controlled trial setting. To mimic a typical epidemiological setting without randomization to the exposures, we perform simulations based on the Cardiovascular Health Study.

6.2 Data-Based Simulations

In the data-based simulation studies, we generate each data set by sampling n participants with replacement from the Cardiovascular Health Study data set, discussed in Section 5. We focus on the outcome frailty and the covariates age, gender, education, CRP level, and smoking. The true β is taken to be the regression coefficient estimates when PG is applied to

the CHS data set (i.e., the coefficients in the first column of Table 2). We vary n between 500, 1000, and 2000. For each sample size, Table 5 shows the average number of ties, across the 10,000 simulations, for each visit. The average number of ties grows with higher sample size, as expected. For each sample size, there are on average 9% of subjects tied at the first visit and 5% of subjects tied at the second visit. For even the smallest number in this table (26), the exact method of tie-handling is computationally challenging since we have to compute $26! = 4 \times 10^{26}$ terms.

Table 6 presents the absolute bias, standard error, root-mean-square error, and coverage probability for $n = 500$. Web Tables 3 and 4 of the Supplementary Materials present the results for sample sizes $n = 1000$ and 2000. The tabulated results are plotted in Web Figures 5–7 in the Supplementary Materials. For all three sample sizes, Laplace, Analytic, Efron's method, and PG have comparable results for bias, standard error, root-mean-square error, and coverage probability. Their bias is small. For all four methods, at $n = 500$, the magnitude of the bias is 0.01 for each covariate. At $n = 1000$, the magnitude of the bias is 0.006 for Laplace and PG, and 0.003 for Analytic and Efron's method. At $n = 2000$, it is 0.002 for Laplace and PG, and 0.007 for Analytic and Efron's method. Their coverage probabilities are about 0.95, except for the covariate CRP at $n = 500$. In this case, Laplace, Analytic, Efron's method, and PG each have coverage probability between 0.93 and 0.94, which is slightly lower than the nominal coverage probability of 0.95. At the larger sample sizes, $n = 1000$ and $n = 2000$, these methods have good coverage probability of 0.945 or higher for all covariates.

7 Discussion

The exact method of tie-handling is computationally intractable when there is a large number of ties in the observed event times. Hence, we propose two, new and computationally efficient approximations, Laplace and Analytic. Both approximations performed well in the simulation studies and a real data application to the Cardiovascular Health Study. The Laplace approximation yielded results that were almost identical to that of the PG model for all scenarios. Laplace and PG had low bias and generated confidence intervals with approximately the nominal coverage probability. Our results provide strong evidence that the Cox model, using the Laplace approximation for tie-handling, can serve as a substitute for the PG model in the grouped data setting. The Analytic approximation, given its remarkable simplicity, performed reasonably well in most scenarios, although it tended to have substantial bias for large β , as shown in the Single Covariate Simulations when $\beta = 2$. Hence, Analytic may be an attractive option in grouped event time data where the relative risks are not expected to be large. In fact, in the CHS application, the Analytic approximation provided results very similar to PG and Laplace. Our new approximations use a different approach than Breslow's and Efron's approximations because the former approximate the exact likelihood with ties, while the latter approximate the actual likelihood. A potential advantage of our approach compared to using the PG model is that fitting the PG model for high-dimensional data might be computationally intensive, relative to our simple approximations.

Both PG and Cox models make the proportional hazards assumption. Although the Cox model estimates the baseline hazard nonparametrically, and the PG model estimates it as the intercept terms, these two are essentially equivalent (with a complementary log-log link in the PG model). Thus, the essential comparison is in terms of different ways of handling the regression part, i.e. the covariate part of the model.

A potential area of future research is to examine the accuracy of our methods under cases with more covariates. Another potential area of future research is to determine the impact of a large number of visits on the results. In this paper, we considered the case of a small number of follow-up visits, e.g., up to 3 follow-up visits in the simulation studies and 2 in the CHS data example.

We would not recommend the use of Breslow's tie-handling method in the grouped data setting. In simulation, it was observed that Breslow's method had larger bias than the other methods and can have poor coverage probability, as low as 0.5. In the CHS application, the results of Breslow's method were also quite different from the other methods.

For those who use the R programming language, we note that the exact method implemented in the "survival" package of R is not the exact method for tiehandling discussed in this paper. The exact method in the "survival" package corresponds to the discrete logistic model proposed by Cox (1972). The discrete logistic model has a different parameter of interest than the models considered in this paper. This is because, rather than making the proportional hazards assumption given in (1), the discrete logistic model assumes that the odds ratios of the hazards are proportional, i.e., $\text{logit}[\lambda\beta(t|Z)] = \text{logit}[k_0(t)] \exp(Z \cdot \beta)$. Thus, the term $\exp(Z \cdot \beta)$ represents the increase in the odds ratio, rather than the increase in the hazard. This model simplifies to the proportional hazards model in (1), if the event times are continuously distributed (Cox, 1972).

The strong performance of the Cox PH model, using Laplace, Analytic, and Efron's tie-handling methods, was an impressive and highly surprising finding, given the extremely discrete nature of the event times that we considered. There were 307 and 187 tied event times, respectively, at the first and second follow-up visits of frailty in the CHS study. One would not think that the Cox PH model would be appropriate for estimating relative risks in such a situation. To our knowledge, this type of result has not been reported in the literature before.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Appendix A:: Derivation of the Laplace Approximation

Denote the integrand of (14) as $P^*(t)$, i.e.,

$$P^*(t) = \left[\prod_{i \in I_k} (1 - e^{-a_i, k^t}) \right] e^{-t}.$$

We have

$$\log P^*(t) = \left[\sum_{i \in I_k} \log(1 - e^{-a_i, k^t}) \right] - t;$$

$$\frac{\partial \log P^*(t)}{\partial t} = \sum_{i \in I_k} \frac{a_{i, k}}{e^{a_{i, k}t} - 1} - 1;$$

$$\frac{\partial^2 \log P^*(t)}{\partial^2 t} = - \sum_{i \in I_k} \frac{a_{i, k}^2 e^{a_{i, k}t}}{(e^{a_{i, k}t} - 1)^2}.$$

Let t_k^* be the solution to $\frac{\partial \log P^*(t)}{\partial t} = 0$, i.e.,

$$\sum_{i \in I_k} \frac{a_{i, k}}{e^{a_{i, k}t_k^*} - 1} = 1.$$

Define

$$\sigma_k^* = - \frac{\partial^2 \log P^*(t_k^*)}{\partial^2 t} = \sum_{i \in I_k} \frac{a_{i, k}^2 e^{a_{i, k}t_k^*}}{(e^{a_{i, k}t_k^*} - 1)^2}.$$

Using a Taylor expansion of $\log P^*(t)$ about the point $t = t_k^*$, we have

$$\log P^*(t) \approx \log P^*(t_k^*) - \frac{\sigma_k^*}{2} (t - t_k^*)^2.$$

Thus, (14) is approximately equal to

$$\int_{t=0}^{\infty} P^*(t_k^*) e^{\sigma_k^* (t - t_k^*)^2 / 2} dt = P^*(t_k^*) \sqrt{2\pi / \sigma_k^*} \Pr(X_k > 0), \quad (25)$$

where X_k has a Normal distribution with mean t_k^* and variance $1 / \sigma_k^*$. Note that (25) is a function of β since t_k^* and σ_k^* depend on β . We estimate the unknown coefficient vector β as the value of β that maximizes the function

$$\prod_{k=1}^m \left[P^*(t_k^*) \sqrt{2\pi / \sigma_k^*} \Pr(X_k > 0) \right], \quad (26)$$

which is the product of (25) over $k = 1, \dots, m$. This is equivalent to finding the maximizer of the function

$$f(\beta) = \sum_{k=1}^m [\log P^*(t_k^*) - 1/2 \log \sigma_k^* + \log \Pr(X_k > 0)],$$

which is obtained by taking the logarithm of (26) and omitting the terms that do not depend on β .

Appendix B:: Derivation of the “Analytic” Approximation

Without loss of generality, we assume that $I_k = \{1, 2, \dots, d_k\}$ to simplify the notation in this proof. We have that

$$\begin{aligned} & \int_{t=0}^{\infty} \left[\prod_{i \in I_k} (1 - e^{-a_i, k t}) \right] e^{-t} dt = \int_{t=0}^{\infty} \left[\prod_{i=1}^{d_k} \left(a_{i, k} t \int_{x_i=0}^1 e^{-a_{i, k} t x_i} dx_i \right) \right] e^{-t} dt \\ &= \left(\prod_{i=1}^{d_k} a_{i, k} \right) \int_{t=0}^{\infty} t^{d_k} e^{-t} \left(\prod_{i=1}^{d_k} \int_{x_i=0}^1 e^{-a_{i, k} t x_i} dx_i \right) dt \\ &= \left(\prod_{i=1}^{d_k} a_{i, k} \right) \int_{t=0}^{\infty} t^{d_k} e^{-t} \left(\int_{x_{d_k}=0}^1 \cdots \int_{x_1=0}^1 e^{-t \sum_{i=1}^{d_k} a_{i, k} x_i} \prod_{i=1}^{d_k} dx_i \right) dt \\ &= \left(\prod_{i=1}^{d_k} a_{i, k} \right) \left(\int_{x_{d_k}=0}^1 \cdots \int_{x_1=0}^1 \int_{t=0}^{\infty} t^{d_k} e^{-t(1 + \sum_{i=1}^{d_k} a_{i, k} x_i)} dt \prod_{i=1}^{d_k} dx_i \right) \\ &= \left(\prod_{i=1}^{d_k} a_{i, k} \right) \int_{x_{d_k}=0}^1 \cdots \int_{x_1=0}^1 \frac{1}{(\sum_{i=1}^{d_k} a_{i, k} x_i + 1)^{d_k+1}} \int_{u=0}^{\infty} u^{d_k} e^{-u} du \prod_{i=1}^{d_k} dx_i \\ &= \left(\prod_{i=1}^{d_k} a_{i, k} \right) \Gamma(d_k + 1) \left[\int_{x_{d_k}=0}^1 \cdots \int_{x_1=0}^1 \frac{1}{(\sum_{i=1}^{d_k} a_{i, k} x_i + 1)^{d_k+1}} \prod_{i=1}^{d_k} dx_i \right]. \end{aligned}$$

We approximate the term in the square brackets by

$$\int_{x_{d_k}=0}^1 \cdots \int_{x_1=0}^1 e^{-(d_k+1) \sum_{i=1}^{d_k} a_{i, k} x_i} \prod_{i=1}^{d_k} dx_i = \frac{1}{(d_k+1)^{d_k}} \prod_{i=1}^{d_k} \frac{1 - e^{-(d_k+1)a_{i, k}}}{a_{i, k}}.$$

Using this approximation, we have

$$\begin{aligned}
& \int_{t=0}^{\infty} \left[\prod_{i \in I_k} (1 - e^{-a_i, k^t}) \right] e^{-t} dt \approx \left(\prod_{i=1}^{d_k} a_{i, k} \right) \Gamma(d_k + 1) \frac{1}{(d_k + 1)^{d_k}} \prod_{i=1}^{d_k} \frac{1 - e^{-(d_k + 1)a_{i, k}}}{a_{i, k}} \\
& \approx \sqrt{2\pi d_k} d_k^{d_k + 1/2} e^{-d_k} \left(\prod_{i=1}^{d_k} a_{i, k} \right) \frac{1}{(d_k + 1)^{d_k}} \prod_{i=1}^{d_k} \frac{1 - e^{-(d_k + 1)a_{i, k}}}{a_{i, k}} \\
& \text{by stirling's formula} \\
& = \sqrt{2\pi d_k} e^{-d_k} [(1 + 1/d_k)^{d_k}]^{-1} \prod_{i=1}^{d_k} [1 - e^{-(d_k + 1)a_{i, k}}] \\
& \approx \sqrt{2\pi d_k} e^{-d_k} e^{-1} \prod_{i=1}^{d_k} [1 - e^{-(d_k + 1)a_{i, k}}] \\
& \text{using the approximation } (1 + 1/d_k)^{d_k} \approx e \\
& = \sqrt{2\pi d_k} e^{-(d_k + 1)} \prod_{i=1}^{d_k} [1 - e^{-(d_k + 1)a_{i, k}}] \\
& = \sqrt{2\pi d_k} e^{-(d_k + 1)} \prod_{i \in I_k} [1 - e^{-(d_k + 1)a_{i, k}}].
\end{aligned}$$

References

- Allison PD (1982): "Discrete-time methods for the analysis of event histories," *Sociological Methodology*, 13, 61–98.
- Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. (2002): "Multi-ethnic study of atherosclerosis: objectives and design," *American journal of epidemiology*, 156, 871–881. [PubMed: 12397006]
- Cox DR (1972): "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202.
- Davis P and Rabinowitz P (1984): *Methods of Numerical Integration*, Dover Publications Mineola, New York.
- DeLong D, Guirguis G, and So Y (1994): "Efficient computation of subset selection probabilities with application to Cox regression," *Biometrika*, 81, 607–611.
- Erdelyi A (1956): *Asymptotic Expansions*, Dover Publications New York, New York.
- Fried L, Borhani N, Enright P, Furberg C, Gardin J, Kronmal R, et al. (1991): "The Cardiovascular Health Study: Design and rationale," *Annals of Epidemiology*, 1, 263–276. [PubMed: 1669507]
- Fried L, Tangen C, Walston J, Newman A, Hirsch C, Gottdiener J, et al. (2001): "Frailty in older adults: Evidence for a phenotype," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 56, M146–M157.
- Hertz-Picciotto I and Rockhill B (1997): "Validity and efficiency of approximation methods for tied survival times in Cox regression," *Biometrics*, 53, 1151–1156. [PubMed: 9333345]
- Kalbfleisch JD and Prentice RL (2002): *The Statistical Analysis of Failure Time Data*, John Wiley & Sons.
- Onder G, Penninx BW, Ferrucci L, Fried LP, Guralnik JM, and Pahor M (2005): "Measures of physical performance and risk for progressive and catastrophic disability: results from the women's health and aging study," *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 60, 74–79.
- Prentice R and Gloeckler L (1978): "Regression analysis of grouped survival data with application to breast cancer data," *Biometrics*, 34, 57–67. [PubMed: 630037]
- Scheike T and Sun Y (2007): "Maximum likelihood estimation for tied survival data under cox regression model via em-algorithm," *Lifetime Data Analysis*, 13, 399–420. [PubMed: 17682942]
- Therneau T and Grambsch P (2010): *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag New York.

Table 1:Distribution of (δ, V) in Cardiovascular Health Study Application.

		V	
		1	2
δ	0	884 (0.25)	2227 (0.62)
	1	307 (0.09)	187 (0.05)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Regression Coefficient Estimates, Standard Error Estimates, and 95% Confidence Intervals in the Cardiovascular Health Study Application. Panel (a) presents the regression coefficient estimates and Panel (b) presents the standard error estimates. Panel (c) presents both the estimates of and the 95% confidence intervals for the exponentiated regression coefficients. Each confidence interval is obtained by taking the corresponding regression coefficient estimate (in Panel A), adding and subtracting the quantity of 1.96 multiplied by the standard error estimate (in Panel B), and then exponentiating the results. The covariate age, in units of years, has been standardized by subtracting 65 and dividing by 10. The covariate CRP level, in units of mg/L, has been standardized by dividing by the interquartile range (2.15 mg/L).

(a) Regression Coefficient Estimates					
PG	Cox PH				
	Laplace	Analytic	Efron	Breslow	
age	1.1127	1.1069	1.1063	1.0263	
education	-0.2874	-0.2860	-0.2859	-0.2669	
gender	-0.5384	-0.5361	-0.5361	-0.5045	
smoke	0.1856	0.1845	0.1849	0.1708	
CRP	0.0418	0.0417	0.0418	0.0405	

(b) Standard Error Estimates					
PG	Cox PH				
	Laplace	Analytic	Efron	Breslow	
age	0.0793	0.0792	0.0786	0.0775	
education	0.0924	0.0924	0.0922	0.0922	
gender	0.1003	0.1003	0.1002	0.1000	
smoke	0.0949	0.0949	0.0948	0.0947	
CRP	0.0123	0.0123	0.0123	0.0125	

(c) Point Estimates and 95% Confidence Intervals for Exponentiated Regression Coefficients									
PG	Cox PH								
	Laplace	Analytic	Efron	Breslow					
age	3.04 [2.60, 3.55]	3.04 [2.60, 3.55]	3.02 [2.59, 3.53]	2.79 [2.40, 3.25]					
education	0.75 [0.63, 0.90]	0.75 [0.63, 0.90]	0.75 [0.63, 0.90]	0.77 [0.64, 0.92]					
gender	0.58 [0.48, 0.71]	0.58 [0.48, 0.71]	0.59 [0.48, 0.71]	0.60 [0.50, 0.73]					

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

(a) Regression Coefficient Estimates					
	PG	Cox PH			
		Laplace	Analytic	Efron	Breslow
smoke	1.20 [1.00, 1.45]	1.20 [1.00, 1.45]	1.20 [1.00, 1.45]	1.20 [1.00, 1.45]	1.19 [0.99, 1.43]
CRP	1.04 [1.02, 1.07]	1.04 [1.02, 1.07]	1.04 [1.02, 1.07]	1.04 [1.02, 1.07]	1.04 [1.02, 1.07]

Table 3:

Single Covariate Simulation Studies: Average Number of Ties at Each Visit, for Each Setting of β and Sample Size n .

	$n = 200$			$n = 500$			$n = 2000$		
β	Visit 1	Visit 2	Visit 3	Visit 1	Visit 2	Visit 3	Visit 1	Visit 2	Visit 3
0	7	18	22	18	45	56	72	178	222
0.2	8	20	24	20	49	60	80	196	240
0.4	9	22	26	22	54	64	89	216	257
0.6	10	24	28	25	60	69	101	239	275
0.8	11	27	29	29	66	73	114	265	293
1	13	29	31	33	73	77	131	294	307
2	27	45	28	68	112	70	271	448	282

Bias, Standard Error (SE), Root-Mean-Square Error (RMSE), and Coverage Probability (CP) in Single Covariate Simulation Study with $n = 500$. For each value of β , the number of simulations is 10000.

Epidemiol Methods. Author manuscript; available in PMC 2021 January 13.

Table 5:Data-Based Simulation Studies: Average Number of Ties at Each Visit, for Each Sample Size n .

n	Visit 1	Visit 2
500	42	26
1000	85	52
2000	170	104

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Bias, Standard Error (SE), Root-Mean-Square Error (RMSE), and Coverage Probability (CP) in Data-Based Simulation Study with $n = 500$. The number of simulations was 10000.

	CoxPH												PG							
	Laplace				Analytic				Efron's method								Breslow's method			
	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP	Bias	SE	RMSE	CP
age	0.004	0.224	1.181	0.949	-0.004	0.221	1.174	0.952	-0.005	0.220	1.173	0.951	-0.087	0.197	1.099	0.956	0.006	0.225	1.183	0.948
education	-0.008	0.258	0.739	0.951	-0.006	0.256	0.737	0.952	-0.006	0.256	0.737	0.951	0.013	0.241	0.721	0.962	-0.008	0.258	0.739	0.951
gender	-0.012	0.281	0.910	0.949	-0.009	0.279	0.907	0.950	-0.010	0.280	0.908	0.949	0.022	0.263	0.880	0.960	-0.013	0.282	0.911	0.949
smoke	-0.004	0.264	0.626	0.949	-0.005	0.262	0.625	0.950	-0.004	0.262	0.626	0.950	-0.018	0.246	0.618	0.964	-0.003	0.264	0.626	0.948
CRP	0.006	0.043	0.568	0.936	0.006	0.043	0.568	0.938	0.005	0.042	0.568	0.936	0.003	0.039	0.568	0.960	0.006	0.043	0.568	0.936