David J. Hendry

Data generation for the Cox proportional hazards model with time-dependent covariates: A method for medical researchers

Article (Accepted version) (Refereed)

Original citation:
Hendry, David, J. (2013) *Data generation for the Cox proportional hazards model with time-dependent covariates: A method for medical researchers.* Statistics in Medicine, 33, (3) pp. 436-454. ISSN 0277-6715.

DOI: 10.1002/sim.5945

Data Generation for the Cox Proportional Hazards Model with Time-Dependent Covariates: A Method for Medical Researchers

David J. Hendry

Published in *Statistics in Medicine*, Vol. 33, No. 3, pp. 436–454.

Abstract

The proliferation of longitudinal studies has increased the importance of statistical methods for time-to-event data that can incorporate time-dependent covariates. The Cox proportional hazards model is one such method that is widely used. As more extensions of the Cox model with time-dependent covariates are developed, simulations studies will grow in importance as well. An essential starting point for simulation studies of time-to-event models is the ability to produce simulated survival times from a known data generating process. This paper develops a method for the generation of survival times that follow a Cox proportional hazards model with time-dependent covariates. The method presented relies on a simple transformation of random variables generated according to a truncated piecewise exponential distribution, and allows practitioners great flexibility and control over both the number of time-dependent covariates and the number of time periods in the duration of follow-up measurement. Within this framework, an additional argument is suggested that allows researchers to generate time-to-event data in which covariates change at integer-valued steps of the time scale. The purpose of this approach is to produce data for simulation experiments that mimic the types of data structures applied researchers encounter when using longitudinal biomedical data. Validity is assessed in a set of simulation experiments and results indicate that the proposed procedure performs well in producing data that conform to the assumptions of the Cox proportional hazards model.

Keywords: survival analysis; Cox proportional hazards model; time-dependent covariates; simulations; truncated piecewise exponential distribution

1 Introduction

Statistical methods for time-to-event data have long been staples of medical research. Within this class of methods, the proportional hazards model proposed by Cox [1] is certainly among the most important, with a variety of its extensions [2–11] and related diagnostic techniques [3, 7, 11–17] having become standard components of the medical researcher's toolbox. Largely due to the increasing availability of data from longitudinal studies, which collect measurements of the same units at different time points, one class of extensions that has received a great deal of attention in recent years is the set of models that augment the standard Cox proportional hazards model to include time-dependent covariates [6, 18–20]. In fact, the inclusion of time-dependent covariates within the Cox framework is now commonplace in medical research [21–28]. As such, and as pointed out by Sylvestre and Abrahamowicz [29], researchers engaged in the development of specialized extensions of survival models, and of the Cox model in particular, must be concerned with the inclusion of time-dependent covariates [30–34].

As more new extensions of the Cox model with time-dependent covariates are developed, and as medical researchers encounter as yet undiscovered complications with longitudinal data structures, validation of modelling techniques through simulations will almost certainly be crucial [29, 35]. To be sure, the value of simulations has not been lost on researchers investigating the Cox model, e.g. [34, 36–42]. However, simulation experiments for event history models in general, and for the Cox model with time-dependent covariates in particular, present a unique set of complications with respect to the generation of simulated data.

For simulations within the Cox framework, many have noted the issue of assuming a functional form for the baseline risk [29, 35, 43, 44] and satisfying the proportional hazards assumption within and across units [11]. However, the most difficult conceptual problem arises when one considers the relationship between hazard rates and survival times. Like many time-to-event models, the Cox model is parameterized in terms of the hazard rate, meaning that the relationship between a set of covariates and the hazard rate must be translated into a relationship between that set of covariates and survival times in order to generate an appropriate set of simulated data. In a series of papers, Leemis and colleagues take on this complication directly by demonstrating that

survival times for proportional hazards models can be generated from known distributions via inversion of the cumulative hazard function [45–47]. In the same vein (but working independently of Leemis and colleagues), Bender et al. offer a detailed framework for the generation of survival times that follow a Cox model without time-dependent covariates [43, 48]. However, Sylvestre and Abrahamowicz argue that Bender et al.'s algorithm cannot easily be extended to the case of the Cox model with time-dependent covariates because it involves inverting a function of the cumulative baseline hazard, which is not possible when the changes in the covariates over time are not described by a parametric function or are not defined over the entire range of the time span [29]. Sylvestre and Abrahamowicz then describe and evaluate two alternatives for the generation of survival times conditional on time-dependent covariates (one that uses a permutation algorithm based on a permutation probability law derived from the Cox model [49, 50], and another that generates events within follow-up measurement periods using a binomial model) [29]. Austin, on the other hand, builds directly upon the framework put forth by Bender et al. and Leemis and colleagues by demonstrating exactly how researchers can extend the method of inverting the cumulative hazard function to the case of proportional hazards models with time-dependent covariates. Specifically, Austin provides detailed derivations of the relationships between survival times generated according to the exponential, Weibull, and Gompertz distributions and a set of time-fixed covariates and exactly one time-dependent covariate [35].

In short, though applications of the Cox proportional hazards model with time-dependent covariates are likely to become increasingly important for medical research, the incorporation of time-dependent covariates remains as a thorny complication to overcome when generating simulated data that adhere to the assumptions of the Cox model. To date, only a small number of researchers have put forth procedures for the generation of simulated data that follow a Cox model with time-dependent covariates [29, 35, 44, 46, 47]. The objective of this paper is to advance this literature further by presenting another general means of simulating survival times conditional on time-dependent covariates that follow the assumptions of the Cox model. Like the methods presented by Sylvestre and Abrahamowicz [29], the procedure proposed here allows for an arbitrary number of time-dependent covariates of unrestricted functional form. However, unlike these methods, but similar to the presentation of Austin [35], the data generating process

that produces survival times can be presented in closed form. The expression of the relationship between time-dependent covariates and survival times in closed form is not an advantage of the procedure per se, but this ability in combination with the capacity to include an arbitrary number and form of the time-dependent covariates differentiates this method from those currently proposed in the literature.

To proceed, the paper extends upon a method advanced by Zhou [44], which thus far has received little attention in the medical literature (but see [51–54]). The method presented relies on a simple transformation of a random variable generated according to a truncated piecewise exponential distribution, where the bounds of truncation allow the user to specify the minimum and maximum number of measurements that are of interest for a particular application, and the piecewise nature of the distribution allows covariates to vary as step functions over the time scale (for a different perspective that uses the piecewise exponential distribution to directly model time-dependent effects in proportional hazards situations, see [55–57]). Within this general framework, an additional argument is suggested that allows practitioners to generate data that vary at integer-valued steps of the time scale, which would be the case in a particular empirical application if follow-up measurements are taken at days, weeks, months, etc. The goal is to suggest a means of generating simulated data that more closely match real-world empirical situations examined in longitudinal studies.

The paper proceeds as follows. The next section formally introduces the Cox proportional hazards model with time-dependent covariates in order to set up the mathematical issues that must be considered in data generation procedures intended to follow this specification. Section 3 explicates the relationship between variates generated according to a truncated piecewise exponential distribution and survival times that follow a Cox proportional hazards model. In so doing, the section provides a general result illustrating that a transformation of the truncated piecewise exponential distribution follows the Cox model, shows how this result can be used in the context of rejection sampling, presents a summary of the suggested algorithm, and offers practical guidance on the choice of a transformation. Section 4 validates the proposed method in a series of simulation experiments, and Section 5 provides guidance on how the method can be extended to examine violations of the assumptions of the Cox model. Finally, Section 6 offers concluding

remarks.

2 The Cox model with time-dependent covariates

Let $Z_{ij}(t)$ be the jth covariate of the ith unit under observation, where $i = 1, \ldots, n$, $j = 1, \ldots, p$, and t is an observed value of the time scale. The notation $Z_{ij}(t)$ indicates that the value of Z_{ij} varies as a function of the time scale. Then the Cox proportional hazards model with time-dependent covariates specifies that the hazard rate for the ith individual is given by:

$$h_i(t) = h_0(t) \exp(Z_i(t)\beta), \tag{1}$$

where h_0 is the so-called baseline hazard rate, $Z_i(t)$ is a $1 \times p$ vector of covariates for unit i that may be either time-fixed or time-dependent, and β is a $p \times 1$ vector of coefficients.

Among the advantages of the Cox model over other types of time-to-event methods is the fact that the baseline hazard can be left unspecified in practice. The only assumption about functional form that a practitioner must make is that h_0 is a nonnegative function of t. For researchers without strong substantive theory about the shape of the hazard when $Z_{ij}(t) = 0$, the Cox model offers a great deal of flexibility. However, the Cox model also imposes a rather strong constraint on the data in that it carries the assumption of proportional hazards. With time-dependent covariates, the proportional hazards assumption states that the relative hazard for any two observations i and j follows the relationship

$$\frac{h_0(t) \exp(Z_i(t)\beta)}{h_0(t) \exp(Z_j(t)\beta)} = \frac{\exp(Z_i(t)\beta)}{\exp(Z_j(t)\beta)}. \tag{2}$$

That is, researchers employing the basic Cox model must be able to reasonably assume that the relative impact of any two values of a covariate—either within or across observations—can be summarized by the single coefficient β [11].

From a data generation perspective, producing random variables that follow a Cox proportional hazards model involves translating the hazard rate given in equation (1) with the property given in equation (2) into an appropriate data generating process for survival times. Without

the presence of time-dependent covariates, this translation can be made in a straightforward way by noting that the exponential, Weibull, and Gompertz distributions, some of the most common distributions employed in parametric survival analysis, also carry the proportional hazards assumption. Therefore, if a researcher is simply interested in generating survival times as part of a data matrix with one row per unit under study, the task can be handled by simulating random variables that follow an exponential, Weibull, or Gompertz distribution conditional on a set of covariates and an assumed value of β, a relatively elementary task using standard statistical software [43]. However, if a researcher is interested in generating data that follow a Cox model with time-dependent covariates, and if the desired data structure is one in which covariates vary at integer-valued steps of the time scale (as would be the case if measurements are taken at minutes, hours, days, weeks, years, etc.), the translation from equations (1) and (2) into a data generating process is not as transparent. The following section demonstrates that such a translation can be achieved by employing a simple transformation of a truncated piecewise exponential random variable.

3 Cox via truncated piecewise exponentials

The idea of using a transformation of exponential random variables to simulate survival times that follow a Cox model was presented by Leemis et al. [46] and Zhou [44]. Expanding on this general approach, Zhou [44] developed a procedure for generating survival times that follow a Cox model with time-dependent covariates that uses a transformation of piecewise exponential random variables. Though this latter approach represents a large step forward given the state of the literature at that time, it is limited to consideration of situations with only one time-change point per unit, where the time-change point characterizes the transition of a single binary time-dependent covariate from a value of 0 to a value of 1. Such a case may be used to represent, for instance, a unit's transition from a non-exposure condition to an exposure condition in which the unit remains exposed for the remainder of the follow-up measurement period (as would be the case with a procedure such as an organ transplant [35]). This simulated circumstance may reasonably capture an important class of real-world settings of interest to biomedical researchers, but it is certainly not general enough to capture the range of situations that concern users of

longitudinal observational data, with measurements taken on a vast array of characteristics and at multiple time points. The argument presented in this paper extends on the discussion of Zhou [44], but provides a more general procedure that allows for an arbitrary number of time change points and an arbitrary form of the covariate(s).

The presentation of the truncated piecewise exponential distribution relies on various developments of piecewise exponential distributions and truncated exponential distributions [4, 44, 58–65]. Notation most closely follows that used by Hougaard [4] and Ibrahim et al. [60].

To begin, consider a prespecified partition of the time scale under investigation, $\mathcal{S} = \{s_1, s_2, \ldots, s_J\}$, $0 = s_0 < s_1 < \cdots < s_J$, forming J intervals $(0, s_1]$, $(s_1, s_2], \ldots, (s_{J-1}, s_J]$, where s_J is greater than the largest observed survival time. For the purposes of exposition, and without loss of generality, in this section only one time-dependent covariate is considered. Define a time-dependent covariate as $Z(t) = Z_j$, $s_{j-1} < t < s_j$, where $Z(t)$ is allowed to follow an arbitrary distribution. Additionally, let $\lambda_j = \exp(Z(t)\beta)$. Then we can say that the random variable T is distributed as piecewise exponential conditional on $Z(t)$ if its density is given by

$$k(t) = \prod_{h=1}^{j-1} \exp\left(-\lambda_h \left(s_h - s_{h-1}\right)\right) \left(\lambda_j\right) \exp\left(-\lambda_j \left(t - s_{j-1}\right)\right) I\{s_{j-1} < t \leq s_j\},$$

where I is the indicator function. The form of the density shows that a piecewise exponential random variable has a constant hazard in each interval, i.e., $\lambda(t) = \lambda_j$ for $t \in (s_{j-1}, s_j]$, $j = 1, \ldots, J$. Then the density of a truncated piecewise exponential random variable with support $[a, b]$ is given by

$$f(t) = \frac{k(t) I\{a \leq t \leq b\}}{K(b) - K(a)},$$

where K is the distribution function associated with k.

To see how the truncated piecewise exponential distribution allows the researcher great flexibility in generating survival times that follow a Cox model with covariates that vary at integer-valued steps of the time scale, consider an arbitrary transformation g, such that

$$g(0) = 0,\ g(t) \nearrow \text{ for } t > 0, \text{ and } g^{-1}(t) \text{ is differentiable.} \tag{3}$$

Applying these properties of g, we are able to derive the following result.

Theorem: Suppose a random variable Y is generated as piecewise exponential with density function given by

$$k_Y(t) = \prod_{h=1}^{j-1} \exp\left(-\lambda_h \left(g^{-1}(s_h) - g^{-1}(s_{h-1})\right)\right) \times (\lambda_j) \exp\left(-\lambda_j \left(t - g^{-1}(s_{j-1})\right)\right) I\{g^{-1}(s_{j-1}) < t \leq g^{-1}(s_j)\},$$

$j = 1, 2, \ldots, J$, and corresponding distribution function $K_Y(t)$, where g is defined in (3). Further, suppose that X is a truncated version of Y with density function given by

$$f_X(t) = \frac{k_Y(t) I\{g^{-1}(a) \leq t \leq g^{-1}(b)\}}{K_Y\left(g^{-1}(b)\right) - K_Y\left(g^{-1}(a)\right)},$$

where $[g^{-1}(a), g^{-1}(b)]$ is the support of X. Then $g(X)$ follows a Cox model with a time-dependent covariate and baseline hazard $h_0(t) = \frac{d}{dt}\left[g^{-1}(t)\right]$.

A proof of the theorem is presented in Appendix A.

Remark 1: Since piecewise exponential random variables can easily be generated using standard statistical software, the generation of truncated piecewise exponentials can be accomplished by using rejection sampling.

The well-known rejection sampling result states that if f and k are densities on $\mathbb{R}$ such that for some $M > 1$,

$$f(x) \leq Mk(x) \ \forall \ x,$$

and if Y is generated from k and $U \sim \mathrm{U}[0, 1]$, then X may be generated from f by calculating $Y = Mk(X)/f(X)$ and accepting each value such that $YU \leq 1$ as a random draw of X. For the context considered here, suppose that Y is a piecewise exponential random variable with density function $k_Y(t)$ and distribution function $K_Y(t)$. Further, suppose that X is a truncated piecewise

exponential random variable with support $[a, b]$, and density

$$f_X(t) = \frac{k_Y(t) I\{a \leq t \leq b\}}{K_Y(b) - K_Y(a)}.$$

Then to generate random draws of X, we need to find $M > 1$ such that $f_X(t) \leq M k_Y(t) \; \forall \; t$. Since f has density 0 for $t < a$ and $t > b$, we only need to consider t such that $a \leq t \leq b$. Note that

$$f_X(t) \leq M k_Y(t) \Longrightarrow \frac{k_Y(t)}{K_Y(b) - K_Y(a)} \leq M k_Y(t) \Longrightarrow \frac{1}{K_Y(b) - K_Y(a)} \leq M.$$

Therefore, for any given bounds of truncation $\{a, b\}$, we can let $M = \frac{1}{K_Y(b) - K_Y(a)}$, provided that $K_Y(b) - K_Y(a) < 1$. Since K is the cumulative distribution function of a piecewise exponential distribution, this relation will hold for all a and b such that $0 < a < b$.

Remark 2: To generate survival times that follow a Cox model conditional on covariates that vary at integer-valued steps of the time scale, one simply has to let the partition $\mathcal{S}$ be a subset of the natural numbers and approximate the value in the final interval of each survival time by using its ceiling value. That is, for a survival time of the ith observation, T_i, generated according to a truncated piecewise exponential distribution with rates $\lambda_1, \ldots, \lambda_J$ and partition $\mathcal{S}_i = \{s_{i1}, \ldots, s_{ij}\} \subset \mathbb{N}$ such that $s_{ij-1} < T_i \leq s_{ij}$, data for the ith observation can be constructed by associating $Z_{i1}, \ldots, Z_{ij}$ with survival times $s_{i1}, \ldots, s_{ij}$. Then the total survival time for the ith observation is given by $\sum_j s_{ij}$.

[Table 1 about here.]

Table I illustrates the structure of data generated according to the proposed method. An identification variable for each observation that is measured at multiple time points is presented along with the elapsed time at each measurement occasion and a censoring indicator. Covariate values are omitted for purposes of illustration. This is a standard format for time-to-event data with time-dependent covariates. Comparing the elapsed times for the corrected and uncorrected versions associated with the final row of each case i.d. demonstrates the effect of using the ceiling value for the final time interval. The corrected data, in which all measurements are assumed to be taken at integer valued steps of the time scale, will look familiar to applied longitudinal

researchers who employ time-to-event methods. The uncorrected data, on the other hand, vary at integer valued steps for all but the final time interval—a structure unlikely to be encountered in observational data. It will be shown later that the corrected version provides an excellent approximation.

3.1 Algorithm

All of the pieces required to generate survival times that follow a Cox proportional hazards model conditional on covariates that vary as integer-valued steps of the time scale have now been presented. The algorithm can be summarized as follows:

1. Define g such that $g(0) = 0$, $g(t) \nearrow$ for $t > 0$, and $g^{-1}(t)$ is differentiable
2. Define a maximum value for the time scale $t \in \mathbb{N}$
3. Define a finite partition of the time scale $\mathcal{S} = \{s_1, \ldots, s_J\} \subset \mathbb{N} : \max \mathcal{S} \leq t$
4. Define bounds of truncation $a, b \in \mathcal{S} : a < b < t$
5. Define number of observations n
6. Define β
7. For i in 1 to n
   (a) Generate $\{Z_{ij}\}_{j=1}^{t}$
   (b) Calculate $\{\lambda_{ij}\}_{j=1}^{t} = \{\exp(Z_{ij}\beta)\}_{j=1}^{t}$
   (c) Generate X_i as truncated piecewise exponential with rates $\lambda_{i1}, \ldots \lambda_{iJ}$, time-change points $g^{-1}(s_1), \ldots, g^{-1}(s_J)$, and bounds of truncation $g^{-1}(a)$, $g^{-1}(b)$
   (d) Calculate $T_i = g(X_i)$
8. Define a censoring indicator $\{\delta_i\}_{i=1}^{n}$, where $\delta_i \in \{0, 1\}$
9. Let data= $\varnothing$. For i in 1 to n
   (a) For j in 1 to ceiling$(T_i) - 1$, add $(0, j, j-1, Z_{ij})$ to matrix for observation i
   (b) For $j =$ ceiling(T_i), add $(\delta_i, j, j-1, Z_{ij})$ to matrix for observation i
   (c) Add matrix for observation i to data

In other words, the user begins with a time partition, along with the minimum and maximum number of measurements required per unit and a desired form for the covariate vector. Presumably, all of these choices will be made with a particular empirical analogue in mind. After T_i is generated conditional on $Z_{ij}\beta$, values 1 through ceiling(T_i) are then associated with each $Z_{i\text{ceiling}(T_i)}$, and all Z_{ij} such that $j >$ ceiling(T_i) are discarded. Additionally, the definition of the censoring indicator is left arbitrary to illustrate that this too can take a variety of forms. The algorithm is currently expressed in such a way that censoring is random, and can be uniformly distributed (and hence, uninformative), subject to a biased assignment mechanism in which, say, survival times that are relatively long or relatively short are more or less likely to be censored, or

subject to a set of cutoff points defined by a value or percentile of T_i, as in [35]. Alternatively, as in [29, 50], the user could specify a marginal distribution of censoring times, C_i, following the same set of procedures used to generate event times and take $T_i^* = \min\{T_i, C_i\}$ and $\delta_i = I\{T_i \leq C_i\}$ as the values of the survival time and censoring indicator for unit i, respectively. This alternative approach would require reexpressing steps 7 and 8 of the algorithm in a straightforward manner.

The result will be a data set having the structure of the corrected data in Table I, with each unit having a minimum and maximum number of rows a and b, respectively.

3.2 Hardware and software specifications

For the simulations that follow, the algorithm was implemented using R 2.15.2 with a *Mersenne-Twister* random number generator on a machine with an Intel Xeon 2.26 GHz processor running Windows 7 64-bit. Piecewise exponential random variables were generated using a suite of functions in the `msm` package [66], Cox parameters were estimated using the `coxph` function [67] with the Efron method for handling tied data, and diagnostic testing of the proportional hazards assumption was performed using the `cox.zph` function in the `survival` package [68]. Sample code for the procedure is presented in Appendix B.

3.3 Practical Considerations

The choices of g, β, a, b, and the data generating process for Z have all been left arbitrary in the discussion up to this point (with g only subject to the requirements in (3)). But all will have important practical consequences in terms of the computational cost of the rejection sampler (step 7(c) of the algorithm) and the form of the final distribution of survival times.

To make appropriate choices, it is important to understand the basic mechanics of the method. In short, the rejection sampler takes random draws from a piecewise exponential distribution, with rates defined by $Z_j\beta$, and only accepts draws that fall between $g^{-1}(a)$ and $g^{-1}(b)$. Intuitively, we can consider piecewise exponential random variables to represent the time until occurrence of an event. Thought of in this way, large positive expected values of $Z_j\beta$ will represent higher rates of event occurrence, and hence smaller average time values. Likewise, large negative expected values of $Z_j\beta$ will represent lower rates of event occurrence, and hence larger average time values.

Therefore, given a data generating process for Z and a value of β, the choices of g, a, and b will make the rejection sampler more or less computationally expensive, and vice versa.

3.3.1 Illustration: Choice of g given a, b, and $Z_j\beta$

Researchers performing simulation experiments in the context of event history models generally concern themselves with coefficient and covariate values that are relatively small in magnitude (often less than 1), whether positive or negative. In such cases, the greatest computational expense of the rejection sampler will often be realized in the number of draws rejected because they do not meet the specified minimum threshold of the truncated distribution. Therefore, given a choice of a minimum number of measurements desired in the resulting set of simulated data, a, it will typically make sense for the researcher to choose g such that $g^{-1}(a)$ will be close to 0. Given the requirements of g specified in (3), a natural choice is to allow g to be a power function and g^{-1} its associated positive root.

Consider a scenario in which the desired minimum and maximum number of measurements in the simulated data are 10 and 150, respectively. We let the vector of j rate parameters for unit i be defined by $\lambda_{ij} = \beta_1 Z_{1ij} + \beta_2 Z_{2ij}$, where $Z_1 \sim \text{U}[-.5, .5]$, $Z_2 \sim \text{Bin}(.5)$, $\beta_1 = 2$, and $\beta_2 = -1$, and we desire 500 simulated datasets with 1000 units each. Figure 1 illustrates the consequences of choosing g as a power function with exponents 2, 3, and 4.

[Figure 1 about here.]

Unlike the process in practical applications, for this illustration all of the piecewise exponential draws were retained in addition to the truncated draws that will be transformed into the resulting survival times. The top row of panels in Figure 1 presents densities for each of the 500 simulations for both the piecewise and truncated piecewise exponential distributions. In each case, the densities illustrate that the vast majority of piecewise exponential draws for this choice of coefficients and covariates fall well short of the desired minimum value, and the effect of assuming a power function for g and using successively larger exponents is to create successively smaller and narrower bounds of truncation. Also presented are the average number of piecewise exponential draws per simulation, $\bar{N}_Y$. For $g^{-1}(t) = \sqrt{t}$, over 18,000 piecewise exponential draws are required, on average, to produce 1000 truncated piecewise exponential draws with the desired

properties. Moving to $g^{-1}(t) = \sqrt[3]{t}$, we see that the average number of piecewise exponential draws required to produce 1000 truncated piecewise draws drops to just over 8000, a reduction of about 56%. Moving to $g^{-1}(t) = \sqrt[4]{t}$, the required number of draws decreases further still, but the rate of decrease tapers off dramatically due to the increasingly narrow bounds of truncation, and hence a larger proportion of draws being rejected because they are greater than the upper bound.

The bottom row of panels illustrates the effect of the choice of g on the resulting transformed truncated piecewise exponential draws that follow a Cox specification. The difference in baseline hazards between the three choices becomes immediately apparent and brings up an additional consideration with respect to computation. Specifically, assuming a power function for g, using successively larger exponents has the effect of vastly increasing the total time at risk in the simulated sample. In this illustration, for $g(t) = t^2$, the average survival time across all 500 simulations of 1000 units each was about 21.2. For $g(t) = t^3$, this average increases to 37.1, and for $g(t) = t^4$ to 44.7. Because this method generates survival times as a function of covariates that vary at integer-valued steps of the time scale, the practical effect of doubling the total time at risk is actually to double the number of rows in the data matrix, which can greatly increase the computational time required to estimate parameters. In another set of simulation experiments (data not shown), simulated data sets were constructed and Cox models were estimated using the same three specifications depicted in Figure 1. Assuming no censored observations, the time required to simulate and estimate parameters for the 500 data sets using $g(t) = t^2$ was approximately 1895 seconds. Using $g(t) = t^3$, the time required more than doubled to approximately 4217 seconds. For $g(t) = t^4$, time increased further to 6412 seconds.

It is important to keep in mind that although power functions and roots are natural choices for g and g^{-1}, they are by no means the only possibilities. Researchers can choose a variety of functional forms, each of which will correspond to a particular functional form for $h_0(t)$ that can be examined graphically, as in Figure 1. Further, the computational expense of each of the choices for g in relation to one another is specific to the particular choice of parameter values and data generating processes for the covariates. The type of preliminary testing undertaken in this illustration is recommended prior to performing a large-scale simulation study.

4 Simulations to assess validity

The algorithm described and tested in the previous section was employed in a series of simulation experiments to assess its performance. Across experiments, 1000 simulated datasets were generated and g was chosen such that $g(t) = t^2$. In one set of experiments, a single time-dependent covariate was generated according to a uniform distribution with support $[-.5, .5]$. In a second set of experiments, two time dependent covariates were generated, one according to a uniform distribution with support $[-.5, .5]$ and another according to a Bernoulli distribution with parameter $p = .5$ (as in the illustration in the previous section). The intention is to demonstrate the proposed procedure using both one and two covariates and, in the latter case, using both a continuous and a binary time-dependent covariate. The continuous covariates could capture, for instance, levels of exposure to some environmental toxin, while the dichotomous covariate could represent a subject's movement in and out of some treatment over the duration of follow-up. Across each of these situations, the proportion of observations that are censored was varied to take on values of 0, .1, .25, and .5 (using random and uniform censoring), and the number of observations per simulated dataset was varied to take on values of 100, 500, and 1000. Results of the Cox regression estimations across these conditions using both the corrected and uncorrected versions of simulated data are presented in Table II. Also included is the elapsed time required to simulate and estimate parameters for the 1000 datasets for each value of N, averaged across the four censoring levels.

[Table 2 about here.]

First examining the elapsed time required to simulate and analyze the data under the various scenarios, we see that moving from $N = 100$ to $N = 500$ leads to upwards of a ten-fold increase in computation time. Moving from $N = 500$ to $N = 1000$ leads to far more modest changes in elapsed time, and in the one-covariate case even a slight decrease.

A scan of the statistics for the coefficient estimates in Table II indicates that the procedure performs quite well on average. As expected, for any given level of censoring, the standard deviation of the estimates decreases as the number of units increases. Further, for any given value of N, the standard deviation decreases as the proportion of censored cases decreases. These

relationships are true for both the one- and two-covariate specifications, and for both the corrected and uncorrected data.

Importantly, the degree of similarity between model estimates from the corrected and uncorrected versions of the data suggest that using the ceiling values of the generated survival times to estimate survival times in the final interval for each observation provides a close approximation. Though the highest levels of bias are seen for the corrected data, the relative bias across all scenarios never exceeds 5 percent of the true parameter value. Across values of N, a general trend is present in which the bias is greater for the highest levels of censoring as compared to the case of no censoring. But there are clear exceptions, and the relationship is not monotonic as the level of censoring changes.

For each level of censoring and each value of N, comparing across the corrected and uncorrected situations is instructive as to the practical impact of using ceiling values to estimate the final time interval for each unit. The effect of using ceiling values rather than actual values is to slightly increase the time at risk for each unit in a given simulation. This means that for any chosen model specification, value of N, and proportion of censored cases, a particular pattern should hold such that the estimates for the corrected simulations slightly undershoot their corrected analogues. Specifically, for a positive coefficient (β in the one-covariate simulations and β_1 in the two-covariate simulations), if the uncorrected bias is positive, the corrected bias for the same situation will either be positive and smaller in magnitude or negative. If the uncorrected bias for a positive coefficient is negative, on the other hand, the corrected bias will also be negative but larger in magnitude. Likewise, for a negative coefficient (β_2 in the two-covariate simulations), if the uncorrected bias is negative, the corrected bias for the same situation will either negative and smaller in magnitude or positive. If, on the other hand, the uncorrected bias for a negative coefficient is positive, the corrected bias will be positive and larger in magnitude. This relationship holds for all simulations and shows why, in some scenarios, the corrected data actually lead to a smaller percent bias in parameter estimates than the uncorrected data.

The overall takeaway point from Table II is that the procedure performs as expected and using the ceiling values of the survival times to estimate the final measurement interval for each observation provides a useful approximation. Indeed, the differences in estimates between the

corrected and uncorrected versions become trivial as the proportion of censored cases approaches 0. And because of the relationship between the corrected and uncorrected versions of the simulated survival times, we have seen that the corrected data can, certain cases, actually lead to parameter estimates that, on average, are closer to the true values than models estimated on the uncorrected data. Given the intuitive appeal of the corrected data for applied researchers, it seems wholly appropriate therefore to use the corrected data to generate data that follow a Cox proportional hazards model with time-dependent covariates.

Producing data for which estimated values of the parameter approach the true values is obviously critical, but is not the only requirement here. As stated previously, the Cox model assumes that the relationship follows the proportional hazards assumption. One of the advantages of the Cox model over parametric proportional hazards models is that the assumption can readily be tested as part of model diagnostics. A variety of methods for assessing the validity of the proportional hazards assumption are available (see [69] for a review). This paper utilizes the method proposed by Grambsch and Therneau [13]. In brief, this test essentially involves calculation of a correlation coefficient for the relationship between scaled Schoenfeld residuals for each covariate and the time scale to determine whether individual covariates violate the proportional hazards assumption. In addition, a global test for proportional hazards uses the calculation of a weighted sum of the scaled Schoenfeld residuals across covariates to determine whether the overall fitted model is consistent with the proportional hazards assumption [13]. In either case, a statistically significant χ^2 statistic (1 d.f. for the covariate-specific tests and d.f. equal to the number of covariates in the model for the global test) indicates a violation of the proportional hazards assumption.

Table III presents summaries of the p-values of the χ^2 statistics for each model estimated on the corrected data from Table II. Results for models estimated on the uncorrected data, χ^2 statistics, and correlation coefficients for covariate-specific tests are omitted. For each value of N, the latter column indicates the number of instances in which the proportional hazards assumption was violated.

[Table 3 about here.]

As can be seen, the procedure employed does not perfectly produce data that follow the assumption of proportional hazards, but violations are rare, falling well within the range that would be expected according to the associated p-value. Specifically, using a $p < .05$ confidence level, the largest number of violations for any of the experiments was only 38, or about 3.8% of simulations. Interestingly, there is a general pattern in which more violations of proportional hazards occur as N increases. This is likely due to the fact that the scaled Schoenfeld residual tests can be sensitive to outlier survival times and simulations with a larger number of units are more likely to produce a few relatively extreme values. Graphical summaries of the residuals (not shown) support this supposition. Though the number of violations will be tolerable for most applications, in individual instances a violation of proportional hazards will lead to biased coefficient estimates and suboptimal significance tests, and it will be up to the researcher to determine how to handle these issues in the context of the particular application. The important point for the practitioner is that for any simulated data set and estimated model, the assumption of proportional hazards can and should be tested.

5 Extensions

The discussion thus far has illustrated the theoretical relationship between variates generated according to a truncated piecewise exponential distribution and survival times that adhere to the assumptions of the Cox model, as well as the adequacy of the result for use in practical software applications. However, in many cases researchers will be interested in examining the fitness of the Cox model when one or more of the assumptions of the data generating process is violated. Two common situations of interest to medical researchers are the efficacy of the Cox model when the proportional hazards assumption does not hold or when there is dependence among units in data settings with repeated events. It is shown here that the method proposed in this paper can be easily extended to handle either of these situations.

5.1 Violations of the proportional hazards assumption

The proportional hazards assumption states that the relative impact of any two values of a covariate can be summarized by a single coefficient. For survival times conditional on time-dependent covariates, this relationship must hold both within and across units. If the true value of β varies as a function of t, then proportional hazards is violated, and a variety of methods have been suggested to account for different forms of nonproportionality, e.g., [11]. In the context of the data generation method proposed here, a straightforward extension that allows the researcher to examine various forms of nonproportional hazards is immediately apparent. That is, rather than defining β as a constant, one can specify β to take on particular values as a function of the time scale.

To illustrate, suppose a researcher was interested in the adequacy of the Cox framework in a situation in which a similar change in a covariate at different points in time has a variable impact on the rate of event occurrence. Using the specifications from the previous section as a base, another set of experiments was performed by allowing β, β_1, and β_2 to increase at pre-defined steps of the time scale. This would represent situations in which a similar change in a covariate leads to a greater likelihood of event occurrence for units that have survived a longer amount of time without yet experiencing the event. Specifically, in a set of one-covariate simulations, β took a value of .1 for $t \leq 10$, 1 for $10 < t \leq 15$, 2 for $15 < t \leq 20$, and 3 for $t > 20$. In a set of two-covariate simulations, β_1 was defined similarly, and β_2 was defined such that it took a value of -5 for $t \leq 10$, -3 for $10 < t \leq 50$, and -1 for $t > 50$. All other specifications were identical to those used in the $N = 1000$ scenario in Table II. Results of the Cox estimation are not presented, and results of the scaled Schoenfeld residual tests on the corrected data are presented in Table IV.

[Table 4 about here.]

The results in Table IV are instructive on at least two points. First, results generally tend to deviate from their desired properties most often in cases of higher levels of censoring. At low levels of censoring, the scaled Schoenfeld residual tests are able to detect violations of proportional hazards in all but a few cases. Second, heavy censoring in combination with a larger number of covariates seems to be associated with greater difficulty in the ability of these diagnostic tests

to detect proportional hazards violations. Specifically, in the two-covariate experiment with 50% censoring 103 of each of the covariate-specific tests fail to detect a violation, a number that is substantially higher than expected. In cases such as these in which violations of proportional hazards are actually desired for a particular simulation study, the sensitivity of these particular diagnostic tests to outlier survival times work against the researcher's aims and the effect is most pronounced with a greater proportion of censored cases. This result reinforces the suggestion that researchers always engage in formal tests of proportional hazards and examine graphical summaries of model residuals as an additional check to ensure that simulated data have the desired properties.

5.2 Repeated events and non-independence

An increasingly common extension of time-to-event models examines situations in which units do not drop out of the risk set after an event occurs, and multiple events per unit are possible. The basic Cox model requires that interevent times among units not be correlated, and this assumption is often unreasonable in practice. Researchers examining means of accounting for non-independence within the Cox framework will wish to produce repeated events data with known forms of non-independence. But until now, virtually all simulation studies of this kind have only considered the time-fixed covariate case, e.g. [11, 38, 70]. For researchers interested in examining Cox models for repeated events with time-dependent covariates, the method presented in this paper provides a convenient starting point.

Modifying the algorithm to include multiple events per subject would begin with a model of the number of events per subject. This could be, for example, a fixed number of events across subjects, or possibly a random draw from a count model. Then if one were to assume independence among units, the process would proceed exactly as in the non-repeated events case by simply drawing the desired number of survival times per subject and stacking the rows of the data matrix. To introduce non-independence, one might consider introducing heterogeneity via a unit-specific random effect, or event dependence by specifying the rate of event occurrence for a given event as a function of the number of events that have already occurred (as in [38]). More specifically, unit heterogeneity could be introduced by specifying the piecewise exponential rate

parameter for the kth event for unit i as

$$\lambda_{ik} = \exp\left(Z_{ik}\beta + \mu_i\right),$$

where μ_i is a unit-specific random effect drawn from a known distribution, producing correlation among units across interevent times. Likewise, event dependence could be introduced by first calculating $\lambda_{i0} = \exp\left(Z_i\beta\right)$, and then specifying the piecewise exponential rate parameter for the kth event for unit i as

$$\lambda_{ik} = f(k)\lambda_{i0},$$

where f is an arbitrary function that can be constructed in such a way as to introduce a desired form of event dependence [38]. Further, unit heterogeneity and event dependence can easily be combined. Simulations using repeated events with non-independence are omitted due to space constraints, but this exposition shows that extending the algorithm proposed in this paper to the case of repeated events can be accomplished without introducing significant complexity.

6 Discussion

As statistical methods for time-to-event data have become more widely used in the biomedical literature, there has naturally been an increasing interest among medical researchers in testing properties of common time-to-event estimators. The most common tool in this endeavor is simulation. Unlike simulated data generation methods for linear and generalized linear regression models, determining the appropriate data generating process for time-to-event models generally involves a translation from a parameterization of a set of covariates' effect on the hazard rate into a relationship between those covariates and survival times. In the context of the Cox proportional hazards model with time-dependent covariates, additional complications make this translation even less transparent.

This paper presented a method of generating simulated survival times that follow the Cox proportional hazards model with time-dependent covariates. The proposed method relies on a relatively simple argument about a transformation of truncated piecewise exponential random

variables. Furthermore, for greater ease of understanding among analysts of longitudinal observational data, the paper suggested a means of ensuring that in the final simulated data structure, covariates for a given unit under follow-up observation vary at integer-valued time points. Such a data structure would mimic real-world situations in which measurements are taken at discrete points (e.g., years, months, weeks, days), which is typical of a large class of longitudinal studies. And, perhaps most importantly, the proposed procedure allows for an arbitrary number and functional form of time-dependent and time-fixed covariates. The purpose is simply to provide applied researchers with a means of producing simulated data with desired properties that achieve a greater degree of empirical realism.

Results presented here indicate that the suggested procedure performs well in practice. In the cases of both one and two time-dependent covariates that were explored in this paper, model estimates more closely approximate true parameter values as the proportion of censored cases decreases. Additionally, violations of the proportional hazards assumption are well within the expected range for given confidence levels. Overall, the evidence indicates that the proposed method provides a valid means of generating simulated data that follow a Cox proportional hazards model with covariates that vary at step functions of the time scale. Further, it was shown how the algorithm could be extended to include time-dependent coefficients that induce violations of the proportional hazards assumption, as well as data structures with repeated events and non-independence among units.

Given the widespread use of the Cox model with time-dependent covariates and the increasing availability of longitudinal biomedical data, the need to examine the model's properties through simulations will continue to grow. The flexibility of the data generation procedure described in this paper will make it a useful tool in this enterprise.

Appendix A. Proof of Theorem

A theorem and proof for the case of a two-piece exponential random variable and a covariate that transitions from a value of 0 to a value of 1 at the single change point were developed in Zhou [44]. The argument presented here follows a similar line of reasoning, but applies to situations with an arbitrary number of time change points and an arbitrary form of the covariates.

First note that the survival function of X is given by

$$S_X(t) = \frac{\prod_{h=1}^{j-1} \exp\left(-\lambda_h \left(g^{-1}\left(s_h\right) - g^{-1}\left(s_{h-1}\right)\right)\right) \exp\left(-\lambda_j \left(t - g^{-1}\left(s_{j-1}\right)\right)\right)}{K_Y\left(g^{-1}(b)\right) - K_Y\left(g^{-1}(a)\right)} \times I\{g^{-1}\left(s_{j-1}\right) < t \leq g^{-1}\left(s_j\right)\},$$

$j = 1, \ldots, J$, where the support of X is $\left[g^{-1}(a), g^{-1}(b)\right]$. Let $g(X) = T$ and note that

$$P(T > t) = P(g(X) > t) = P\left(X > g^{-1}(t)\right) = S_X\left(g^{-1}(t)\right)$$

Therefore, the survival function of T is given by

$$S_T(t) = \frac{\prod_{h=1}^{j-1} \exp\left(-\lambda_h \left(g^{-1}\left(s_h\right) - g^{-1}\left(s_{h-1}\right)\right)\right) \exp\left(-\lambda_j \left(g^{-1}(t) - g^{-1}\left(s_{j-1}\right)\right)\right)}{K_Y\left(g^{-1}(b)\right) - K_Y\left(g^{-1}(a)\right)} \times I\{g^{-1}\left(s_{j-1}\right) < g^{-1}(t) \leq g^{-1}\left(s_j\right)\},$$

$j = 1, \ldots, J$, where the support of T is $\left[g^{-1}(a), g^{-1}(b)\right]$. Noting that the distribution function of T follows the relationship $F_T(t) = 1 - S_T(t)$, it immediately follows that the density of T is given by

$$\begin{aligned} f_T(t) =& \frac{d}{dt} F_T(t) \\ =& \frac{\left[\frac{d}{dt} g^{-1}(t)\right] \prod_{h=1}^{j-1} \exp\left(-\lambda_h \left(g^{-1}\left(s_h\right) - g^{-1}\left(s_{h-1}\right)\right)\right) \left(\lambda_j\right) \exp\left(-\lambda_j \left(g^{-1}(t) - g^{-1}\left(s_{j-1}\right)\right)\right)}{K_Y\left(g^{-1}(b)\right) - K_Y\left(g^{-1}(a)\right)} \\ & \times I\{g^{-1}\left(s_{j-1}\right) < g^{-1}(t) \leq g^{-1}\left(s_j\right)\}, \end{aligned}$$

$j = 1, \ldots, J$, where, again, the support of T is $\left[g^{-1}(a), g^{-1}(b)\right]$. Further, if we define $h_0(t) = \left[\frac{d}{dt} g^{-1}(t)\right]$, then because the hazard rate follows the relationship $h(t) = \frac{f(t)}{S(t)}$, we have that

$$h_T(t) = \left[\frac{d}{dt} g^{-1}(t)\right] \lambda_j = h_0(t) \exp\left(Z(t)\beta\right)$$

Therefore, T follows a Cox proportional hazards model with baseline hazard $h_0(t) = \left[\frac{d}{dt} g^{-1}(t)\right]$, time-varying covariate $Z(t)$, and constant hazard in each interval $(s_{j-1}, s_j]$, $j = 1, \ldots, J$.

Appendix B. Example Computer Code

This appendix presents R code that implements the suggested data generation algorithm. The rejection sampling procedure used here generates 60 uniform and piecewise exponential random variables at a time for comparison. This choice was based on a series of speed optimization tests that compared the time required to generate 100 draws from a truncated piecewise distribution with the desired bounds of truncation and rate parameters. Drawing one piecewise exponential random variable at a time was computationally very expensive, taking over 150 seconds to generate 100 replicates from the desired distribution. The time required dramatically decreased between 1 and 30 draws at a time, continuing to decrease until about 60 draws at a time, peaking at a minimum of .35 seconds elapsed time. Beyond 60, the elapsed time began to steadily increase again. It should be noted, however, that this optimization result is sensitive to the expected value of the rate parameter. Therefore, it is recommended that researchers employing this method perform similar optimization tests for their desired covariate and coefficient values prior to undertaking a large-scale simulation.

This example code uses one time-dependent covariate, $Z \sim \mathrm{U}[-0.5, 0.5]$. Further, $n = 1000$, $\beta = 2$, $g(t) = t^2$; $[10, 150]$ are the bounds of truncation (corresponding to the minimum and maximum number of follow-up measurements for units) and 50% of observations are censored. Within the code, lines of comments are denoted by `#`. The procedure is as follows:

```
require(msm)
require(survival)

# CREATING g() AND g^-1()
g.inv <- sqrt
g <- function(x) {
  x^2
}

# CREATING THE TIME SCALE AND TRANSFORMED TIME SCALE
t <- 0:199
t.diff <- (t[-1] - t[1:(length(t) - 1)])[-(length(t) - 1)]
g.inv.t <- g.inv(t)
g.inv.t.diff <- (g.inv(t[-1]) - g.inv(t[1:(length(t) - 1)]))[-(length(t) - 1)]
```

```
#CREATING THE BOUNDS OF TRUNCATION
t.max <- 150
t.min <- 10
g.inv.t.max <- g.inv(t.max)
g.inv.t.min <- g.inv(t.min)

#DATA GENERATING PROCESS FOR COVARIATE
B <- function(N, m, M) {
  runif(N, m, M)
}
#BETA
b <- 2
#NUMBER OF OBSERVATIONS
n <- 1000

#CREATING DATA VECTOR
z.list <- list()
for (i in 1:n) {
  z <- B(length(t), -0.5, 0.5)
  z.list[[i]] <- cbind(z, exp(b * z))
}

#GENERATING DATA USING ACCEPT-REJECT METHOD
k <- function(x, m, M, rates, t){
  ifelse(x <= m | x >= M, 0, dpexp(x, rates, t))
}

gen.y <- function(x) {
  x1 <- x[, 2]
  d <- ppexp(g.inv.t.max, x1, g.inv.t) - ppexp(g.inv.t.min, x1, g.inv.t)
  M <- 1 / d
  r <- 60
  repeat{
    y <- rpexp(r, x1, g.inv.t)
    u <- runif(r)
    t <- M * ((k(y, g.inv.t.min, g.inv.t.max, x1, g.inv.t) / d /
      dpexp(y, x1, g.inv.t)))
    y <- y[u <= t][1]
    if (!is.na(y)) break
  }
  y
}
y <- sapply(z.list, gen.y)
g.y <- g(y)

#CREATING CENSORING INDICATOR
prop.cen <- 0.5
d <- sample(0:1, n, replace = TRUE, prob = c(prop.cen, 1 - prop.cen))
```

```
#CREATING DATASET
data <- NULL
for (i in 1:n) {
  id.temp <- rep(i, ceiling(g.y[i]))
  time.temp <- c(1:ceiling(g.y[i]))
  time0.temp <- 0:ceiling(g.y[i] - 1)
  d.temp <- c(rep(0, length(time.temp) - 1), d[i])
  z.temp <- z.list[[i]][1:(ceiling(g.y[i])), 1]
  data.temp <- cbind(id.temp, time.temp, time0.temp, d.temp, z.temp)
  data <- rbind(data, data.temp)
}
colnames(data) <- c('id', 't', 't0', 'd', 'z1')
data <- data.frame(data)
model <- coxph(Surv(t0, t, d) ~ z1, data = data)
schoenfeld <- cox.zph(model, transform = 'identity')

#RESULT
data
summary(model)
schoenfeld
```

References

1. Cox D. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**(2):187–220.

2. Andersen PK, Gill R. Cox's regression model for counting processes: A large sampe study. *The Annals of Statistics* 1982; **10**(4):1100–1120.

3. Collett D. *Modelling Survival Data in Medical Research*. 2nd edn. Chapman & Hall/CRC, Boca Raton, FL, 2003.

4. Hougaard P. *Analysis of Multivariate Survival Data*. Springer-Verlag, New York, 2000.

5. Kalbfleisch J, Prentice R. Marginal likelihoods based on Cox's regression and life model. *Biometrika* 1973; **60**(2):267–278.

6. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd edn. John Wiley & Sons, Hoboken, NJ, 2002.

7. Klein JD. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer–Verlag, New York, 1997.

8. Lin D. Cox regression analysis of multivariate failure time data. *Statistics in Medicine* 1994; **13**(21):2233–2247. DOI: 10.1002/sim.4780132105.

9. Oakes D. Frailty models for multiple event times. In *Survival Analysis: State of the Art*, Klein JP, Goel PK, eds. Kluwer, Netherlands, 1992; 371–380.

10. Schemper M. Cox analysis of survival data with non-proportional hazard functions. *Journal of the Royal Statistical Society, Series D* 1992; **41**(4):455–465.

11. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000.

12. Cox D, Snell E. A general definition of residuals (with discussion). *Journal of the Royal Statistical Society, Series B* 1968; **30**(2):248–275.

13. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**(3):515–526. DOI: 10.1093/biomet/81.3.515.

14. Harrell F. The PHGLM procedure. In *SUGI Supplemental Library User's Guide*, Hastings RP, ed., Version 5 edn. SAS Institute, Cary, NC, 1986; 437–466.

15. Hosmer DW Jr., Lemeshow S, May S. *Applied Survival Analysis*. John Wiley & Sons, Hoboken, NJ, 2008.

16. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika* 1980; **67**(1):145–153. DOI: 10.1093/biomet/67.1.145.

17. Therneau TM, Grambsch PM, Fleming TR. Martingale-based residuals for survival models. *Biometrika* 1990; **77**(1):147–160. DOI: 10.1093/biomet/77.1.147.

18. Aydemir Ülker, Aydemir S, Dirschedl P. Analysis of time-dependent covariates in failure time data. *Statistics in Medicine* 1999; **18**(16):2123–2134. DOI: 10.1002/(SICI)1097-0258(19990830)18:16<2123::AID-SIM176>3.0.CO;2-4.

19. Fisher LD, Lin D. Time-dependent covariates in the Cox proportional-hazards regression model. *Annual Review of Public Health* 1999; **20**:145–157. DOI: 10.1146/annurev.publhealth.20.1.145.

20. Lancaster T. *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge, 1990.

21. Ananthakrishnan AN, Higuchi LM, Huang ES, Khalili H, Richter JM, Fuchs CS, Chan AT. Aspirin, nonsteroidal anti-inflammatory drug use, and risk for Crohn disease and ulcerative colitis. *Annals of Internal Medicine* 2012; **156**(5):350–359.

22. Finch A, Beiner M, Lubinski J, Lynch HT, Moller P, Rosen B, Murphy J, Ghadirian P, Friedman E, Foulkes WD, Kim-Sing C, Wagner T, Tung N, Couch F, Stoppa-Lyonnet D, Ainsworth P, Daly M, Pasini B, Gershoni-Baruch R, Eng C, Olopade OI, McLennan J, Karlan B, Weitzel J, Sun P, Narod SA, for the Hereditary Ovarian Cancer Clinical Study Group. Salpingo-oophorectomy and the risk of ovarian, fallopian tube, and peritoneal cancers in women with a *BRCA1* or *BRCA2* mutation. *Journal of the American Medical Association* 2006; **296**(2):185–192. DOI: 10.1001/jama.296.2.185.

23. Freedman LS, Oberman B, Sadetzki S. Using time-dependent covariate analysis to elucidate the relation of smoking history to Warthin's tumor risk. *American Journal of Epidemiology* 2009; **170**(9):1178–1185. DOI: 10.1093/aje/kwp244.

24. Gogas H, Ioannovich J, Dafni U, Stavropoulou-Giokas C, Frangia K, Tsoutsos D, Panagiotou P, Polyzos A, Papadopoulos O, Stratigos A, Markopoulos C, Bafaloukos D, Pectasides D, Fountzilas G, Kirkwood JM. Prognostic significance of autoimmunity during treatment of melanoma with interferon. *New England Journal of Medicine* 2006; **354**(7):709–718.

25. Houston TK, Person SD, Pletcher MJ, Liu K, Iribarren C, Kiefe CI. Active and passive smoking and development of glucose intolerance among young adults in a prospective cohort: CARDIA study. *British Medical Journal* 2006; **332**(7545):1064–1069. DOI: 10.1136/bmj.38779.584028.55.

26. Okin PM, Wachtell K, Devereux RB, Harris KE, Jern S, Kjeldsen SE, Julius S, Lindholm LH, Nieminen MS, Edelman JM, Hille DA, Dahlöf B. Regression of electrocardiographic left ventricular hypertrophy and decreased incidence of new-onset atrial fibrillation in patients with hypertension. *Journal of the American Medical Association* 2006; **296**(10):1242–1248. DOI: 10.1001/jama.296.10.1242.

27. Sylvestre MP, Huszti E, Hanley JA. Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Annals of Internal Medicine* 2006; **145**(5):361–363.

28. Zhou Z, Rahme E, Abrahamowicz M, Pilote L. Survival bias associated with time-to-treatment initiation in drug effectiveness evaluation: A comparison of methods. *American Journal of Epidemiology* 2006; **162**(10):1016–1023. DOI: 10.1093/aje/kwi307.

29. Sylvestre MP, Abrahamowicz M. Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine* 2008; **27**(14):2618–2634. DOI: 10.1002/sim.

30. Abrahamowicz M, MacKenzie TA. Joint estimation of time-dependent and non-linear effects of continuous covariates on survival. *Statistics in Medicine* 2007; **26**(2):392–408. DOI: 10.1002/sim.2519.

31. Giorgi R, Gouvernet J. Analysis of time-dependent covariates in a regressive relative survival model. *Statistics in Medicine* 2005; **24**(24):3863–3870. DOI: 10.1002/sim.2400.

32. Heinzl H, Kaider A. Gaining more flexibility in Cox proportional hazards regression models with cubic spline functions. *Computer Methods and Programs in Biomedicine* 1997; **54**(3):201–208. DOI: 10.1016/S0169-2607(02)00022-6.

33. Kooperberg C, Clarkson DB. Hazard regression with interval-censored data. *Biometrics* 1997; **53**(4):1485–1494. DOI: 10.1111/1541-0420.00067.

34. Leffondré K, Abrahamowicz M, Siemiatycki J. Evaluation of Cox's model and logistic regression for matched case-control data with time-dependent covariates: A simulation study. *Statistics in Medicine* 2003; **22**(24):3781–3794. DOI: 10.1002/sim.1674.

35. Austin PC. Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine* 2012; **31**(29):3946–3958. DOI: 10.1002/sim.5452.

36. Baldi I, Ponti A, Zanetti R, Ciccone G, Merletti F, Gregori D. The impact of record-linkage bias in the Cox model. *Journal of Evaluation in Clinical Practices* 2010; **16**(1):92–96. DOI: 10.1111/j.1365-2753.2009.01119.x.

37. Benner A, Zucknick M, Hielscher T, Ittrich C, Mansmann U. High-dimensional Cox models: The choice of penalty as part of the model building process. *Biometrical Journal* 2010; **2010**(1):50–69. DOI: 10.1002/bimj.200900064.

38. Box-Steffensmeier JM, Boef SD. Repeated events survival models: The conditional frailty model. *Statistics in Medicine* 2006; **25**(20):3518–3533. DOI: 10.1002/sim.2434.

39. Huszti E, Abrahamowicz M, Alioum A, Quantin C. Comparison of selected methods for modeling of multi-state disease progression processes: A simulation study. *Communications in Statistics–Simulation and Computation* 2012; **40**(9):1402–1421. DOI: 10.1080/03610918.2011.575505.

40. Laubender R, Bender R. Estimating adjusted risk difference (RD) and number needed to treat (NNT) measures in the Cox regression model. *Statistics in Medicine* 2009; **29**(7–8):851–859. DOI: 10.1002/sim.3793.

41. Malloy EJ, Spiegelman D, Eisen EA. Comparing measures of model selection for penalized splines in Cox models. *Computational Statistics and Data Analysis* 2009; **53**(7):2605–2616. DOI: 10.1016/j.csda.2008.12.008.

42. Omurlu IK, Ozdamar K, Ture M. Comparison of Bayesian survival analysis and Cox regression analysis in simulated and breast cancer data sets. *Expert Systems with Applications* 2009; **36**(8):11341–11346. DOI: 10.1016/j.eswa.2009.03.058.

43. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* 2005; **24**(11):1713–1723. DOI: 10.1002/sim.2059.

44. Zhou M. Understanding the Cox regression models with time-change covariates. *The American Statistician* 2001; **55**(2):153–155. DOI: 10.1198/000313001750358491.

45. Leemis LM. Variate generation for accelerated life and proportional hazards models. *Operations Research* 1987; **35**(6):892–894. DOI: 10.1287/opre.35.6.892.

46. Leemis LM, Shih LH, Reynertson K. Variate generation for accelerated life and proportional hazards models with time dependent covariates. *Statistics and Probability Letters* 1990; **10**(6):335–339. DOI: 10.1016/0167-7152(90)90052-9.

47. Shih LH, Leemis LM. Variate generation for a nonhomogenous Poisson process with time-dependent covariates. *Journal of Statistical Computation and Simulation* 1993; **44**(3–4):165–186. DOI: 10.1080/00949659308811457.

48. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models (Letter to the Editor). *Statistics in Medicine* 2006; **25**(11):1778–1779. DOI: 10.1002/sim.2369.

49. Abrahamowicz M, MacKenzie T, Esdaile JM. Time-dependent hazard ratio: Modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association* 1996; **91**(436):1432–1439. DOI: 10.1080/01621459.1996.10476711.

50. MacKenzie T, Abrahamowicz M. Marginal and hazard ratio specific random data generation: Applications to semi-parametric bootstrapping. *Statistics and Computing* 2002; **12**(3):245–252. DOI: 10.1023/A:1020750810409.

51. Deo RC, Wilson JG, Xing C, Lawson K, Kao WL, Reich D, Tandon A, Akylbekova E, Patterson N, Thomas H. Mosley J, Boerwinkle E, Herman A. Taylor J. Single-nucleotide polymorphisms in *LPA* explain most of the ancestry-specific variation in Lp(a) levels in African Americans. *PLoS One* 2011; **6**(1):e14581. DOI: 10.1371/journal.pone.0014581.

52. May S, Hosmer DW. A cautionary note on the use of the Grønnesby and Borgan goodness-of-fit test for the Cox proportional hazards model. *Lifetime Data Analysis* 2004; **10**(3):283–291. DOI: 10.1023/B:LIDA.0000036393.29224.1d.

53. Moran J, Solomon P, Fox V, Salagaras M, Williams P, Quinlan K, Bersten A. Modelling thirty-day mortality in the acute respiratory distress syndrome (ARDS) in an adult ICU. *Anaesthesia and Intensive Care* 2004; **32**(3):317–329.

54. Moran JL, Bersten AD, Solomon PJ, Edibam C, Hunt T, The Australian and New Zealand Intensive Care Society Clinical Trials Group. Modelling survival in acute severe illness: Cox versus accelerated failure time models. *Journal of Evaluation in Clinical Practice* 2008; **14**(1):89–93. DOI: 10.1111/j.1365-2753.2007.00806.x.

55. Karrison T. Confidence intervals for median survival times under a piecewise exponential model with proportional hazards covariate effects. *Statistics in Medicine* 1996; **15**(2):171–182. DOI: 10.1002/(SICI)1097-0258(19960130)15:2¡171::AID-SIM146¿3.0.CO;2-U.

56. Seaman SR, Bird SM. Proportional hazards model for interval-censored failure times and time-dependent covariates: Application to hazard of HIV infection of injecting drug users in prison. *Statistics in Medicine* 2001; **20**(12):1855–1870. DOI: 10.1002/sim.809.

57. Slasor P, Laird N. Joint models for efficient estimation in proportional hazards regression models. *Statistics in Medicine* 2003; **22**(13):2137–2148. DOI: 10.1002/sim.1439.

58. Barlow RE, Proschan F. *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, Inc., New York, 1975.

59. Friedman M. Piecewise exponential models for survival data with covariates. *Annals of Statistics* 1982; **10**(1):101–113. DOI: 10.1214/aos/1176345693.

60. Ibrahim JG, Chen MH, Sinha D. *Bayesian Survival Analysis.* Springer, New York, 2001.

61. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*, vol. 1. 2nd edn. John Wiley and Sons, New York, 1994.

62. Kim JS, Proschan F. Piecewise exponential estimator of the survivor function. *IEEE Transactions on Reliability* 1991; **40**(2):134–139. DOI: 10.1109/24.87112.

63. Malla GB. Extending the piecewise exponential estimator of the survival function. *Proceedings of the 4th Annual GRASP Symposium* 2008; **4**:69–70.

64. Popov VA, Litvinov ML. Solution of queuing problems using piecewise distribution functions. *Cybernetics and Systems Analysis* 1973; **9**(3):451–455. DOI: 10.1007/BF01069200.

65. Zelterman D, Grambsch PM, Le CT, Ma JZ, Curtsinger JW. Piecewise exponential survival curves with smooth transitions. *Mathematical Biosciences* 1994; **120**(2):233–250. DOI: 10.1016/0025-5564(94)90054-X.

66. Jackson C. Multi-state modelling with R: The msm package. 2007. Available from: `http://rss.acs.unt.edu/Rdoc/library/msm/doc/msm-manual.pdf`. (Accessed on 16 May 2013).

67. R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, 2006.

68. Therneau TM. A package for survival analysis in S. 1999. Available from: `http://mayoresearch.mayo.edu/mayo/research/biostat/upload/survival.pdf`. (Accessed on 16 May 2013).

69. Ng'andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Statistics in Medicine* 1997; **16**(6):611–626. DOI: 10.1002/(SICI)1097-0258(19970330)16:6<611::AID-SIM437>3.0.CO;2-T.

70. Kelly PJ, Lim LLY. Survival analysis for recurrent event data: An application to childhood infectious diseases. *Statistics in Medicine* 2000; **19**(1):13–33. DOI: 10.1002/(SICI)1097-0258(20000115)19:1¡13::AID-SIM279¿3.0.CO;2-5.

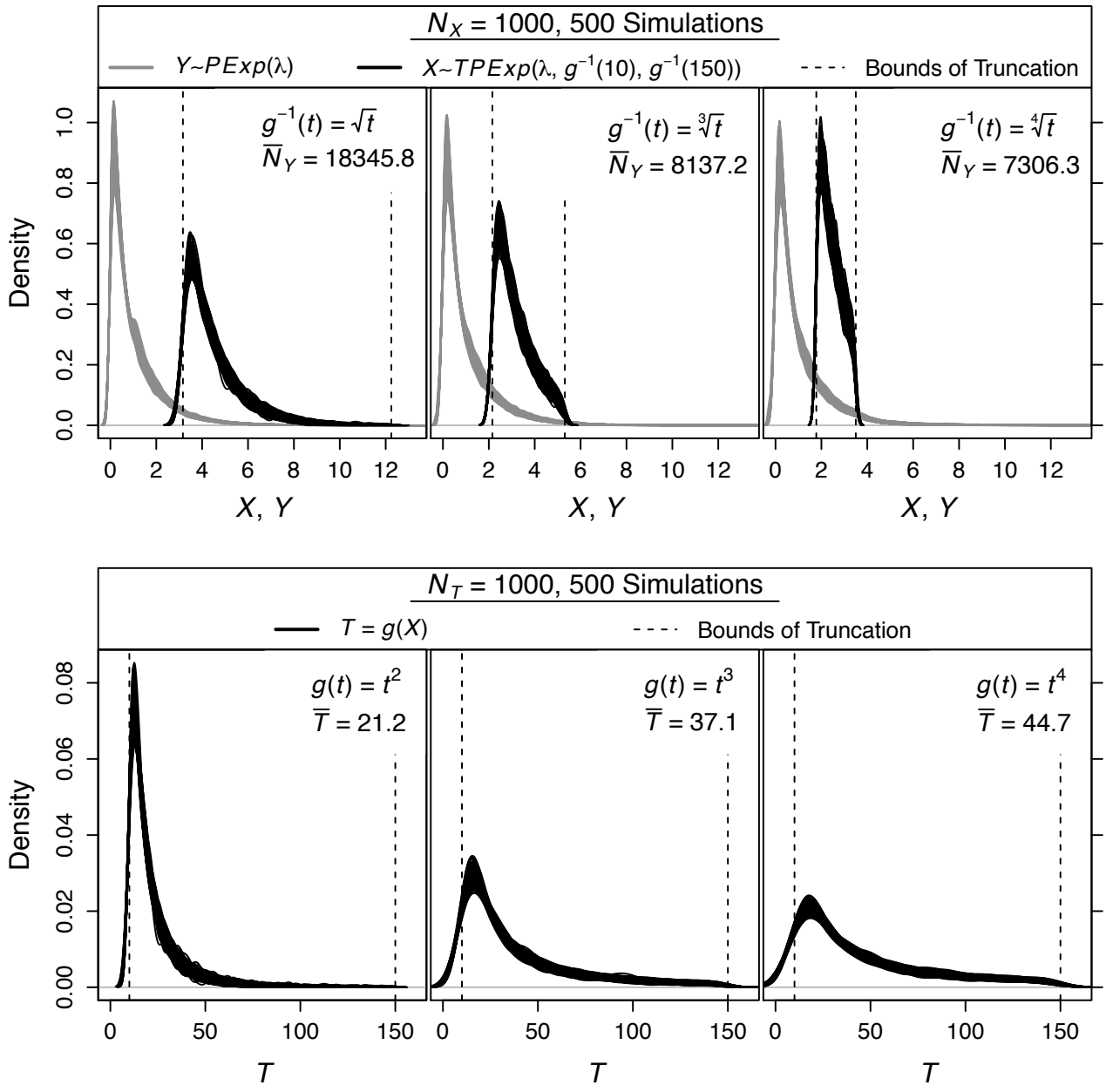

Figure 1. Effect of choice of g on rejection sampler iterations and form of survival distribution. $Y \sim$ piecewise exponential, $X \sim$ truncated piecewise exponential, $T = g(X)$, $\lambda = 2{*}Z_1 + (-1){*}Z_2$, λ, Z_1, and Z_2 are vectors of length t, $Z_{1ij} \sim \mathrm{U}[-.5, .5]$, $Z_{2ij} \sim \mathrm{Bin}(.5)$, $i = 1, \ldots, 1000$, $j = 1, \ldots, t > 150$.

Table I. Data Structures for Uncorrected and Corrected Data

Uncorrected Data			Corrected Data		
Case I.D.	Elapsed Time	Censoring Indicator	Case I.D.	Elapsed Time	Censoring Indicator
1	1.000	0	1	1	0
1	2.000	0	1	2	0
⋮	⋮	⋮	⋮	⋮	⋮
1	17.000	0	1	17	0
1	17.648	1	1	18	1
2	1.000	0	2	1	0
2	2.000	0	2	2	0
⋮	⋮	⋮	⋮	⋮	⋮
2	13.000	0	2	13	0
2	13.789	0	2	14	0
⋮	⋮	⋮	⋮	⋮	⋮
55	1.000	0	55	1	0
55	2.000	0	55	2	0
⋮	⋮	⋮	⋮	⋮	⋮
55	16.000	0	55	16	0
55	16.736	1	55	17	1
⋮	⋮	⋮	⋮	⋮	⋮

Note: Uncorrected data uses transformed draws from a piecewise exponential random variable. Corrected data uses the ceiling values of the transformed random draws to produce an estimate for survival times in the final interval for each unit.

Table II. Cox estimation with simulated survival times and time-dependent covariates

Uncorrected Data										
Prop.		$\hat{\beta}$, $N = 100$			$\hat{\beta}$, $N = 500$			$\hat{\beta}$, $N = 1000$		
Cens.	β	Mean	SD	Bias	Mean	SD	Bias	Mean	SD	Bias
0.50	β : 2	2.045	0.576	0.045	2.008	0.246	0.008	2.002	0.170	0.002
0.25	β : 2	2.011	0.453	0.011	1.993	0.198	-0.007	1.998	0.144	-0.002
0.10	β : 2	2.035	0.424	0.035	2.004	0.180	0.004	2.003	0.124	0.003
0.00	β : 2	2.006	0.387	0.006	2.000	0.173	<0.000	2.009	0.123	0.009
Elapsed Time		350.58 Seconds			3771.50 Seconds			3411.53 Seconds		
0.50	β_1 : 2	2.003	0.553	0.003	2.002	0.238	0.002	2.002	0.169	0.002
	β_2 : -1	-1.025	0.339	-0.025	-1.008	0.142	-0.008	-0.998	0.102	0.002
0.25	β_1 : 2	2.014	0.468	0.014	2.008	0.189	0.008	1.997	0.142	-0.003
	β_2 : -1	-1.001	0.267	-0.001	-0.998	0.115	0.002	-1.002	0.080	-0.002
0.10	β_1 : 2	2.021	0.432	0.021	2.009	0.183	0.009	1.998	0.126	-0.002
	β_2 : -1	-0.985	0.243	0.015	-1.004	0.108	-0.004	-0.998	0.074	0.002
0.00	β_1 : 2	2.009	0.405	0.009	2.008	0.172	0.008	1.997	0.120	-0.003
	β_2 : -1	-1.008	0.230	-0.008	-1.004	0.098	-0.004	-1.000	0.069	<0.000
Elapsed Time		340.60 Seconds			3868.95 Seconds			4476.02 Seconds		
Corrected Data										
Prop.		$\hat{\beta}$, $N = 100$			$\hat{\beta}$, $N = 500$			$\hat{\beta}$, $N = 1000$		
Cens.	β	Mean	SD	Bias	Mean	SD	Bias	Mean	SD	Bias
0.50	β : 2	1.957	0.573	-0.043	1.923	0.245	-0.077	1.917	0.169	-0.083
0.25	β : 2	1.961	0.449	-0.039	1.947	0.197	-0.053	1.952	0.144	-0.048
0.10	β : 2	2.011	0.422	0.011	1.982	0.179	-0.018	1.981	0.124	-0.019
0.00	β : 2	1.997	0.386	-0.003	1.994	0.172	-0.006	2.005	0.123	0.005
Elapsed Time		358.22 Seconds			3932.64 Seconds			3515.20 Seconds		
0.50	β_1 : 2	1.936	0.550	-0.064	1.937	0.237	-0.063	1.937	0.169	-0.063
	β_2 : -1	-0.993	0.338	0.007	-0.977	0.143	0.023	-0.967	0.102	0.033
0.25	β_1 : 2	1.975	0.465	-0.025	1.973	0.189	-0.027	1.962	0.142	-0.038
	β_2 : -1	-0.983	0.266	0.017	-0.982	0.115	0.018	-0.986	0.080	0.014
0.10	β_1 : 2	2.001	0.429	0.001	1.991	0.183	-0.009	1.981	0.125	-0.019
	β_2 : -1	-0.976	0.241	0.024	-0.996	0.108	0.004	-0.991	0.074	0.009
0.00	β_1 : 2	2.001	0.404	0.001	2.004	0.172	0.004	1.993	0.120	-0.007
	β_2 : -1	-1.005	0.229	-0.005	-1.002	0.098	-0.002	-0.999	0.069	0.001
Elapsed Time		345.10 Seconds			4050.03 Seconds			4701.83 Seconds		

Note: Uncorrected version uses raw survival times; corrected version uses ceiling values of survival times (see Table I). "Prop. Cens." refers to the proportion of cases censored. 1000 datasets were constructed and $g(t) = t^2$ for all simulations. Elapsed time is the time that was required to simulate and estimate parameters for the 1000 simulations, averaged across the four levels of censoring.

Table III. Summary of p-values from scaled Schoenfeld residual proportional hazards tests, corrected data

		$N = 100$			$N = 500$			$N = 1000$		
Prop.		p-values		Freq.	p-values		Freq.	p-values		Freq.
Cens.	Z	Mean	SD	$p < .05$	Mean	SD	$p < .05$	Mean	SD	$p < .05$
0.50	Z_1	0.590	0.267	13	0.539	0.272	23	0.532	0.286	34
0.25	Z_1	0.581	0.268	8	0.560	0.264	15	0.535	0.280	34
0.10	Z_1	0.589	0.264	11	0.553	0.270	23	0.549	0.275	26
0.00	Z_1	0.585	0.263	10	0.570	0.267	10	0.531	0.274	21
0.50	Z_1	0.586	0.269	23	0.543	0.279	25	0.529	0.279	27
	Z_2	0.590	0.265	18	0.542	0.282	29	0.517	0.276	32
	Global	0.632	0.267	20	0.562	0.279	29	0.542	0.278	27
0.25	Z_1	0.587	0.260	16	0.538	0.268	21	0.534	0.271	25
	Z_2	0.565	0.267	16	0.548	0.267	18	0.523	0.286	38
	Global	0.619	0.261	14	0.575	0.262	19	0.547	0.274	30
0.10	Z_1	0.595	0.254	7	0.544	0.275	13	0.520	0.278	28
	Z_2	0.589	0.257	13	0.549	0.274	23	0.539	0.282	34
	Global	0.642	0.243	6	0.573	0.272	12	0.548	0.282	31
0.00	Z_1	0.584	0.260	12	0.551	0.271	23	0.548	0.284	29
	Z_2	0.596	0.259	8	0.536	0.277	27	0.544	0.285	35
	Global	0.643	0.256	12	0.570	0.272	17	0.564	0.281	33

Note: Tests performed on corrected data from Table II. "Prop. Cens." refers to the proportion of cases censored. p-values associated with Z_1 and Z_2 are $\chi^2(1)$ probabilities representing a test of the null hypothesis that each covariate individually does not violate the proportional hazards assumption. p-values associated with global tests are $\chi^2(2)$ probabilities representing a test of the null hypothesis that the full model does not violate the proportional hazards assumption. Large p-values indicate evidence in favor of the proportional hazards assumption.

Table IV. Summary of p-values from scaled Schoenfeld residual tests of proportional hazards, time-dependent coefficients, corrected data

		p-values		
Prop. Cens.	Z	Mean	SD	Freq. $p < .05$
0.50	Z_1	0.008	0.030	963
0.25	Z_1	0.001	0.011	996
0.10	Z_1	0.000	0.001	1000
0.00	Z_1	0.000	0.001	1000
0.50	Z_1	0.021	0.061	897
	Z_2	0.021	0.076	897
	Global	0.004	0.035	988
0.25	Z_1	0.004	0.018	985
	Z_2	0.003	0.015	989
	Global	0.000	0.002	1000
0.10	Z_1	0.001	0.007	997
	Z_2	0.002	0.010	986
	Global	0.000	0.000	1000
0.00	Z_1	0.000	0.002	1000
	Z_2	0.002	0.022	993
	Global	0.000	0.000	1000

Note: Tests performed on corrected data. "Prop. Cens." refers to the proportion of cases censored. p-values associated with Z_1 and Z_2 are $\chi^2(1)$ probabilities representing a test of the null hypothesis that each covariate individually does not violate the proportional hazards assumption. p-values associated with global tests are $\chi^2(2)$ probabilities representing a test of the null hypothesis that the full model does not violate the proportional hazards assumption. Small p-values indicate evidence against the proportional hazards assumption.