12-1-2015

Modeling Arbitrarily Interval-Censored Survival Data with External Time-Dependent Covariates

Wei Fang
University of Northern Colorado

UNIVERSITY OF NORTHERN COLORADO

Greeley, Colorado

The Graduate School

MODELING ARBITRARILY INTERVAL-CENSORED SURVIVAL DATA WITH EXTERNAL TIME-DEPENDENT COVARIATES

A Dissertation Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

Wei Fang

College of Educational and Behavioral Sciences
Department of Applied Statistics and Research Methods

December 2015

This Dissertation by: Wei Fang

Entitled: *Modeling Arbitrarily Interval-Censored Survival Data with External Time-Dependent Covariates*

has been approved as meeting the requirement for the Degree of Doctor of Philosophy in College of Education and Behavioral Sciences, Program of Applied Statistics and Research Methods

Accepted by the Doctoral Committee

____
Trent Lalonde, Ph.D., Research Advisor

____
Jay Schaffer, Ph.D., Committee Member

____
Susan Hutchinson, Ph.D., Committee Member

____
Robert Heiny, Ph.D., Faculty Representative

Date of Dissertation Defense _______________________________

Accepted by the Graduate School

____
Linda L. Black, Ed.D.
Associate Provost and Dean
Graduate School and International Admissions

ABSTRACT

Fang, Wei. *Modeling Arbitrarily Interval-Censored Survival Data with External Time-Dependent Covariates*. Published Doctor of Philosophy dissertation, University of Northern Colorado, 2015.

Arbitrarily interval-censored survival data refer to the situation where the exact time of the occurrence of an event of interest is only known to have occurred within some two consecutive examinations. External time-dependent covariates refer to those whose values change during the periodic follow-up, and whose value at a particular time does not require individuals to be under direct observation. Regression modeling of survival data usually either handles arbitrarily interval-censored data alone (Farrington, 1996) or external time-dependent covariates alone (Cox, 1972; Therneau & Grambsch, 2000). In the current research, an adjustment has been made to the data augmentation used in Farrington's estimation method for arbitrarily interval-censored data to accommodate external time-dependent covariates. The three approaches, regression analysis of arbitrarily interval-censored survival data by Farrington (1996), the extended Cox model (Cox, 1972; Therneau & Grambsch, 2000) for handling external time-dependent covariates, and the proposed model for handling both arbitrarily interval-censored data and external time-dependent covariates, were compared in terms of hypothesis testing performance.

The simulation results revealed that the proposed model was more powerful than the other two models, and the type I error rate from the proposed model fluctuated around the nominal level .05, and was comparable to that from the extended Cox model.

Moreover, the proposed model gave the smallest absolute relative bias of parameter estimates, and always gave the correct direction of the effect from the significant external time-dependent covariate. As such, the proposed model depicted the survival experience of subjects regarding the timing of the occurrence of an event more realistically.

According to the results of the current research, the proposed model can be used in practice as an alternative to the popular extended Cox model (Cox, 1972; Therneau & Grambsch, 2000) for investigating what factors influence the survival times of subjects.

ACKNOWLEDGEMENTS

I wish to thank my committee members, Dr. Trent Lalonde, Dr. Jay Schaffer, Dr. Susan Hutchinson, and Dr. Robert Heiny, thank you for every moment of instruction and correction.

To my research advisor specifically, Dr. Trent Lalonde, I really appreciate your support for my Ph.D. journey academically and spiritually: exchanging countless emails, having a meeting with me whenever I needed it, and encouraging me not to lose confidence. The support is crucial to me.

To Dr. Jay Schaffer, thank you very much for your support since I came to the University of Northern Colorado. To Dr. Susan Hutchinson, thank you very much for your precious time you spent on my dissertation. To Dr. Robert Heiny, thank you very much for your patience.

I want to thank my dearest family for their steadfast faith and support through this long journey: my father, Zhenqing Fang, my mother, Suhua Zhang, my sister Ying Fang, my brother Kai Fang, my sister Lin Fang, and my brother-in-law Jinyan Wang. I could never have done this without each of you.

My work is a reflection of all who I have mentioned plus those unmentioned, and it is my hope that it will make a positive contribution.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I

INTRODUCTION

Background

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. An event is defined as a qualitative change that can be situated in time (Allison, 2010). Timing refers to when the change occurred. Thus, survival analysis is extremely useful for studying many different kinds of events situated in time in both the social and natural sciences, such as disease onset, equipment failures, earthquakes, stock market crashes, and retirements. Different kinds of events include both those with increasing hazards and those with decreasing hazards. Increasing hazards refer to situations where, as time goes on, the hazard of the occurrence of an event of interest increases. Equipment failure is an example. Decreasing hazards refer to the hazard of the occurrence of an event of interest decreasing as time goes on. An example of decreasing hazards is survival of burned patients. The main feature of survival analysis that renders conventional statistical methods inappropriate is that survival data are frequently censored, which refers to when the occurrence of an event of interest has not been observed for a subject during a follow-up study. In other words, survival data contain incomplete information. It is worth mentioning that throughout the dissertation, the subject of survival analysis only refer to human subjects.

In survival analysis, there are usually three basic goals. The first goal is to estimate and interpret the survival and/or hazard functions from a particular group, which may refer to a particular treatment group, as in experimental designs with manipulated independent variables, an age group, or a cohort of senior high school students. The survival function is defined to be the probability of a subject's surviving beyond some time *t*, and the hazard function is defined to be the risk of experiencing an event of interest at some time *t.* The second goal is to compare the survival and/or hazard functions between different treatment groups. Comparison can also be made between distinct values of a covariate. The third goal is to assess the effect of independent variables on the hazard of an event. Independent variables can be factors or covariates, either alone or in combination (Collett, 2003). A covariate is a variable that takes numerical values that are often on a continuous scale of measurement, such as age or blood pressure. A factor is a variable that takes a limited set of values, which are known as the levels of the factor. For example, sex is a factor with two levels, or a treatment plan might include both the standard treatment and a new treatment. In this research, only covariates were considered. Regarding the first and second goals, estimating, interpreting, and comparing the survival and/or hazard functions is in the nature of descriptive statistics, while the third goal is analogous to regression analysis. The third goal was the focus of the current research.

In survival analysis, although the most common type of survival data is right-censored data, where the event time of interest is observed either exactly or is greater than the pre-specified study end time for all subjects, a special type of survival data is often encountered. Suppose researchers are interested in the onset of an event of interest, such as AIDS. However, as the occurrence of the event of interest is occult, no one can know the exact time of its occurrence. Thus, researchers usually conduct periodic follow-ups to keep track of the status of the event. Accordingly, it is only known that the true event time is greater than the last examination time at which the change of status has not occurred and less than or equal to the first examination time at which the change of status has been observed to occur; thus giving an interval that contains the real but unobserved time of occurrence of the change of status. Data in this form are known as interval-censored data.

Different censoring mechanisms produce different types of interval-censored data, such as current status data, arbitrarily interval-censored data, doubly censored data, panel count data, and truncated interval-censored data (Sun, 2006). For the current research, only arbitrarily interval-censored data were considered. In particular, for the *i*th subject, let $\tau_{i0} = 0$ be the starting time of a periodic follow-up, i.e., study entry, τ_{im} be the *m*th examination, τ_{il} be the final examination, $m = 1,\ldots, l - 1$, and T_i be the unobservable time of the occurrence of an event of interest. Thus, when there are *l* examinations within the follow-up $(\tau_{i0}, \tau_{il}]$ per subject, where *l* might vary across all subjects, and T_i is known to have occurred within some two consecutive examinations τ_{im} and $\tau_{i(m+1)}$, with $\tau_0 < \tau_{im} < \tau_{i(m+1)} \leq \tau_{il}$, arbitrarily interval-censored data arise. The use of different types of brackets indicates that the unobservable event time is greater than τ_{i0}, but less than or equal to τ_{il}, i.e., $\tau_{i0} < T_i \leq \tau_{il}$. In other words, the event has not occurred by τ_{i0}, but has occurred by τ_{il}. For example, suppose in one study, after 200 patients were discharged healthy from a hospital, they were examined periodically to ascertain their health status. For those who get sick between two examinations, all that is known is that the time when they are still healthy is at least as long as the time of the earlier examination and is no longer than the time of the most recent examination. The exact time is not known, though. It is possible that across this cohort, each one kept the same series of examination times, thus making analysis much more straightforward, and the survival analysis methods of Prentice and Gloeckler (1978), including how to estimate regression coefficients and the survival function, would have applied. However, since event times may be censored into overlapping and non-disjoint intervals, these methods may not be directly applicable. The current research concentrated on the latter case.

When analyzing arbitrarily interval-censored data, as when analyzing any other type of survival data, estimation of the survival function or the hazard function, analogous to descriptive statistics in ordinary statistical analysis, is perhaps the first task. In doing so, information is needed, such as the status of an event of interest during the course of a periodic follow-up, the examinations during which the status has changed, and the number of subjects who are still free of the occurrence of the event after the last examination. If researchers are interested in a more detailed analysis, such as quantifying the effect of independent covariates on the survival function or the hazard function, that is, conducting regression analysis, additional information from independent covariates needs to be collected during the periodic follow-up.

Independent covariates can be either time-independent or time-dependent depending on whether they change in value over the course of a follow-up. Time-independent covariates refer to those whose values are recorded at study entry and remain constant during the periodic follow-up. Examples include randomized treatment and race. On the other hand, there may be situations where one or more of the variables are measured during the periodic follow-up and their values change over time. This type of covariate is known as time-dependent covariates. Blood pressure measured at different times is an example. Intuitively, if account can be taken of the values of covariates as they evolve, a more satisfactory model for describing the hazard of an event of interest at any given time should be obtained. For example, in connection with studies on heart disease, more recent values of blood pressure may provide a better indication of future life expectancy than the value at study entry.

Time-dependent covariates are further classified as being either internal or external. An internal time-dependent covariate is one whose value is subject-specific and requires that the subject be under periodic observation. Typical examples of internal covariates are disease complications and measurements recorded at follow-up examinations. In contrast, an external time-dependent covariate is one whose value at a particular time does not require subjects to be under direct observation. A standard example of an external covariate is the time of day or the season of the year. Certain random covariates such as measurements of air pollution can also be considered as external. The reason why it is important to distinguish between internal and external time-dependent covariates is that an internal covariate requires special treatment compared to an external one. The current research concentrates on external time-dependent covariates.

Many regression models have been proposed for quantifying the effect of independent covariates on survival times. One way to classify these models depends on whether a particular form of probability distribution for the underlying survival times is assumed. As such these models can be classified into two broad categories: semi-parametric regression models and parametric regression models. If there is no need to assume a particular form of probability distribution for the underlying survival times, semi-parametric regression models are preferred, such as those based on the Cox proportional hazards (PH) model (Cox, 1972), and those based on the odds of the survival function, like the proportional odds model (McCullagh, 1980). On the other hand, if the assumption of a particular probability distribution for the underlying survival times is valid, a class of parametric regression models is preferred, such as the exponential model, the Weibull model, the gamma model, and the Gompertz model (Lindsey, 1998). Due to their flexibility and widespread applicability, semi-parametric regression models were chosen over parametric regression models for the current research.

Among different semi-parametric models for regression analysis of survival data, which are proposed from different aspects of the association between the event time and independent covariates, those based on the Cox PH model are the most frequently used forms of the semi-parametric models due to the simplicity of implementation. Therefore, the one chosen for regression analysis of arbitrarily interval-censored data in the current research was based on the Cox PH model.

Formally, the Cox PH model assumes that the hazard function at time *t* has the form

$$h(t|\boldsymbol{X}) = h_0(t)e^{(\boldsymbol{\beta}'\boldsymbol{X})}, \tag{1}$$

given a vector of time-independent covariates $\boldsymbol{X}$, where $h_0(t)$ denotes the unspecified baseline hazard function, that is, the hazard function for subjects with $\boldsymbol{x} = \mathbf{0}$, or the infinite-dimensional nuisance parameter, and $\boldsymbol{\beta}$ denotes the vector of unknown regression parameters, or finite-dimensional regression parameters. The corresponding survival function is

$$S(t|\boldsymbol{X}) = [S_0(t)]^{e^{(\boldsymbol{\beta}'X)}}. \tag{2}$$

In terms of the type of covariates assumed in Equation 1, it is restricted to time-independent covariates alone. When external time-dependent covariates, which do not necessarily require a subject to be under direct observation, and whose values evolve along the course of a follow-up study, are incorporated into this model instead, the Cox PH in Equation 1 becomes the extended Cox model (Cox, 1972; Therneau & Grambsch, 2000), as the hazards between different time-dependent treatment groups, or distinct time-dependent covariate values, are no longer proportional as time goes on.

In terms of the form survival data could assume in Equation 1, the form is restricted to right-censored survival data alone, and thus the Cox PH model cannot be directly applied to arbitrarily interval-censored data. However, arbitrarily interval-censored data, as described above, depict the survival experience of subjects regarding the timing of the occurrence of an event more realistically.

Although external time-dependent covariates often arise in practice, most of the inference procedures developed for arbitrarily interval-censored data only apply to time-independent covariates (Sun, 2006), such as Farrington’s (1996) model, which is based on the Cox PH model. Thus, from a theoretical perspective, there is a need to propose a new modeling approach to accommodate arbitrarily interval-censored survival data and external time-dependent covariates simultaneously. More importantly, in practice, the extended Cox model, Farrington’s model, and the new modeling approach actually share the same data collection process. In particular, after each subject is recruited to a follow-up study, the time of an examination, values of covariates of interest, and the status of a subject at various examinations are recorded. However, how the collected data are used in regression analysis is different among the three approaches. The extended Cox model uses almost all the collected data, except that the mid-point imputation method (Law & Brookmeyer, 1992) is used to create an exact event time from the last two examinations, as the extended Cox model requires one event time. The new modeling approach uses all the collected data. The data used for Farrington's model is almost identical to those used for the new modeling approach, except that Farrington's approach uses covariate values recorded at study entry instead of covariate values recorded at various examinations.

The collected data for all three models contain a series of correlated binary responses, time-dependent covariates, and examinations. The extended Cox model accommodates external time-dependent covariates, a series of binary responses, and an event time created from the last two examinations. Farrington's approach accommodates covariate values recorded at study entry, two correlated binary responses, and last two examinations. The proposed approach accommodates external time-dependent covariates, a series of correlated binary responses, and every examination.

In summary, the proposed approach uses the most information from the collected data among the three approaches. In addition, the proposed approach considers correlation among serial binary responses and use external time-dependent covariates. As such, the proposed approach was expected to be more powerful than Farrington's approach which partially uses the information from the collected data and time-independent covariates alone, as compared to time-independent covariates, external time-dependent covariates are usually assumed to have closer connection to the response variable that evolves along the course of a follow-up study. Regarding the extended Cox model, it also accommodates close connection between external time-dependent covariates and the response variable, hazard. However, no comparison has been made regarding the power between the extended Cox model which describes a continuous response variable, and the proposed approach that models a binary response variable. Nonetheless, with the use of an imprecise, but more appropriate description of the time of the occurrence of an event, the proposed approach depicts the survival experience of subjects more realistically than the extended Cox model which uses a precise, but inappropriate description of the time of the occurrence of an event.

Taken all together, in the current research, an attempt was made to model arbitrarily interval-censored survival data with external time-dependent covariates. Emphasis was placed upon the method of estimating regression parameters.

There are two points worth mentioning for the current research. First, when time-independent covariates are incorporated in the Cox PH model, the coefficient of a covariate in the Cox PH model is a log-hazard ratio, and so under this model, the hazard ratio is constant over time. If this ratio depends on time, i.e., from an external time-dependent covariate, the log-hazard ratio is not constant, and as such a proportional hazards model no longer exists.

Second, in non-parametric analysis of arbitrarily interval-censored data, one basic and important assumption that is commonly used is that the censoring mechanism is independent of or non-informative about the event of interest. An easier way to understand this assumption is that all that is known is the event of interest happened between the two predetermined examination times. One possible scenario under which this assumption would not hold is, for instance, if the occurrence of the event of interest could be accompanied by symptoms, which would make one subject more likely to go for an examination. In this case, it would be reasonable to suspect that the event occurred closer to the right endpoint of the censoring interval. On the other hand, even if the occurrence of the event of interest could be accompanied by symptoms, and the subject does not change the predetermined examination times, this assumption would hold. This assumption applies to regression analysis of arbitrarily interval-censored survival data as well.

In order to investigate the effect of external time-dependent covariates on imprecise but more appropriate survival times, i.e., arbitrarily interval-censored survival data, regression analyses were conducted using the extended Cox model, Farrington's model, and the proposed approach in the current research. Besides investigating how parameters would be estimated and parameter hypothesis tests would be performed in the presence of arbitrarily interval-censored survival data with external time-dependent covariates in conducting regression analysis using the proposed approach, there were three main research questions.

Research Questions

The following research questions guided this research:

Q1 How does absolute relative bias (ARB) of parameter estimates, that is, the absolute value of the difference between parameter estimates and true values of the coefficients divided by of the coefficients, and percent of correct sign of parameter estimates (% CS) from the proposed approach compare to those from Farrington's model, and those from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates?

Q2 How does the power from the proposed approach compare to that from Farrington’s model and that from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates?

Q3 How does type I error rate from the proposed approach compare to that from Farrington’s model and that from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates?

Delimitations of the Research

There were some limitations to the current study. First, due to the unique form of Farrington’s expression for response probability, the resulting baseline hazard function decreases monotonically. Consequently, the proposed model does not apply to real world examples where the resulting baseline hazard function increases monotonically. Second, the current research concentrated on the role of external time-dependent covariates in regression analysis of survival data, while the role of commonly used internal time-dependent covariates played in modeling arbitrarily interval-censored data was not investigated. Third, the current research concentrated on arbitrarily interval-censored data alone, while in reality left-censored and right-censored survival data are collected as well. Fourth, the proposed model was based on the Cox PH model where there is a multiplicative relationship between the hazards and covariates. The additive hazards model, which accounts for an additive relationship between the hazards and covariates, was not investigated in the current study.

The Organization of the Research

The current research is organized as follows. In Chapter II, a literature review was conducted on survival analysis, arbitrarily interval-censored data, external time-dependent covariates, and the current status of research in modeling survival data regarding how to estimate parameters in models that account for either external time-dependent covariates alone or arbitrarily interval-censored data alone. In Chapter III, the data structure for conducting the corresponding regression analysis using each of the three approaches was detailed, the rationale of employing the proposed model was presented, and the inference procedures for the proposed approach were detailed. The design for conducting the simulation study was discussed as well. In Chapter IV, the simulation design was reviewed, and the simulation results comparing properties of parameter estimates obtained from the three approaches were presented in tables and figures. In Chapter V, a discussion of the simulation results was presented, and limitations of the current research and directions for future research were discussed.

CHAPTER II

REVIEW OF LITERATURE

In this chapter, the elements of the procedure for modeling arbitrarily interval-censored data with external time-dependent covariates, including concepts and features of survival analysis, arbitrarily interval-censored data and external time-dependent covariates, and basic models used in regression analysis of survival data were detailed first. Then previous modeling procedures, either handling arbitrarily interval-censored data alone or external time-dependent covariates alone, were reviewed in order to find the gap to be filled by the current research.

An Introduction to Survival Data Analysis

Basic Concepts

Survival data, or time to event data, take the form of times from a well-defined time origin until the occurrence of some particular event. Time means years, months, weeks, or days from the beginning of the follow-up of a subject until an event occurs. An event is defined as a qualitative change that can be situated in time, such as disease incidence, equipment failures, promotions, and retirements. Although survival data arise mainly in biology and medicine, they are observed in other application areas as well, such as sociology, education, epidemiology, engineering, economics, finance, and demography.

Survival data present themselves in different ways, and the main feature of different types of survival data is incomplete observation of time (Hosmer, Lemeshow, & May, 2008), which is due to two mechanisms, namely censoring and truncation. Censoring, broadly speaking, occurs when a subject's survival time is known to have occurred only in a certain period of time. There are three types of censoring mechanisms, namely right censoring, where all that is known is that the subject has not yet experienced the event of interest at a given time; left censoring, where all that is known is that the subject has experienced the event of interest prior to the first examination of a study; and interval censoring, where the only information is that the event of interest occurs within some time interval. The second mechanism, sometimes confused with censoring, is truncation. Truncation of survival data occurs when only those subjects whose event time lies within a certain observational window are observed (Klein & Moeschberger, 2005). A subject whose event time is not in this interval is not observed and no information on this subject is available. This is in contrast to censoring where there is at least partial information on each subject. An example would be a study of risk factors for time to diagnosis of colorectal cancer among subjects in a cancer registry with this diagnosis (Hosmer et al., 2008). If one subject would not enter the analysis until time 10, this type of incomplete observation of time is called truncation. If one subject entered the analysis from study entry and withdrew at time 10, this type of incomplete observation of time is called censoring. The current research only considered censoring.

Analysis of Survival Data

After survival data are collected, an initial step in the analysis is to present descriptions of the survival times for subjects receiving a particular treatment protocol.

For example, in one clinical trial, subjects are randomized to receive either a standard treatment or a new treatment. Researchers might be interested in the survival experience of subjects who receive the new treatment. Then, focus is shifted to investigating what factors influence the survival times. To do this, various models are built to explore the relationship between the survival times and independent variables.

Descriptive methods. In describing survival data, there are two functions of central interest, namely the survival function and the hazard function.

The survival function, denoted *S*(*t*), is defined to be the probability of a subject's surviving beyond some time *t*. When the random variable associated with the survival time, denoted *T*, is continuous, the survival function is the complement of the cumulative distribution function of *T*, denoted *F*(*t*), representing the probability of a subject's surviving less than or equal to *t*. That is,

$$S(t) = \mathrm{P}(T \geq t) = 1 - F(t). \tag{3}$$

The survival function is also the integral of the probability density function for *T*, denoted *f*(*t*), as

$$S(t) = \mathrm{P}(T \geq t) = \int_t^{\infty} f(s)\, ds. \tag{4}$$

Closely related to the survival function is the hazard function, denoted *h*(*t*), which, by definition, represents the risk or hazard of experiencing the event of interest at some time *t*, and is obtained from the probability that a subject experiences the event at some time *t*, conditional on that subject's having survived to that time, written $P(t \leq T \leq t + \Delta t \mid T \geq t)$, where Δt denotes a time interval. This conditional probability is then expressed as a probability per unit time by dividing by the time interval, Δ*t*, to give a rate. The hazard function is then the limiting value of this quantity, as Δ*t* tends to zero, so that

$$h(t) = \lim_{\Delta t \to 0^{+}} \left[\frac{\mathrm{P}(t \le T \le t + \Delta t |\, T \ge t)}{\Delta t}\right]. \qquad (5)$$

In Equation 5, the product of *h*(*t*) on the left hand side and Δ*t* in the denominator may be viewed as the approximate probability that a subject experiences the event of interest in the interval (*t*, *t* + Δ*t*), conditional on that subject's having survived to time *t*. According to a standard result from probability theory, the probability of an event *B*, conditional on the occurrence of another event *A*, is given by *P*(*B*|*A*) = *P*(*A* ∩ *B*)/*P*(*A*), where *P*(*A* ∩ *B*) is the probability of the joint occurrence of *A* and *B*. Using this result, the conditional probability in the hazard function in Equation 5 takes the form

$$\begin{aligned}\mathrm{P}(t \le T \le t + \Delta t |\, T \ge t) &= \frac{\mathrm{P}[(t \le T \le t + \Delta t) \cap (T \ge t)]}{\mathrm{P}(T \ge t)} \\ &= \frac{\mathrm{P}(T \ge t)\,\frac{\mathrm{P}(t \le T \le t+\Delta t)}{\mathrm{P}(T \ge t)}}{\mathrm{P}(T \ge t)} \\ &= \frac{\mathrm{P}(t \le T \le t + \Delta t)}{\mathrm{P}(T \ge t)} \\ &= \frac{F(t + \Delta t) - F(t)}{S(t)}.\end{aligned}$$

Then,

$$h(t) = \lim_{\Delta t \to 0^{+}} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right] \frac{1}{S(t)}.$$

Now, the definition of the derivative of *F*(*t*) with respect to *t* is, which is acutally *f*(*t*), takes the form

$$F'(t) = \lim_{\Delta t \to 0^+} \left[\frac{F(t + \Delta t) - F(t)}{\Delta t}\right],$$

and therefore

$$h(t) = \frac{f(t)}{S(t)}. \tag{6}$$

It then follows that

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt}[\log S(t)],$$

and therefore

$$S(t) = e^{[-H(t)]}, \tag{7}$$

where

$$H(t) = \int_0^t h(s)\, ds.$$

The function *H*(*t*) is called the cumulative hazard function.

From the above, it can be seen that knowing any one of *f*(*t*), *S*(*t*), *F*(*t*), *h*(*t*), or *H*(*t*) is enough to specify the other four expressions, which greatly facilitates the descriptions of the survival experience of subjects during a follow-up study.

Modeling survival data. After descriptive statistics for the survival times themselves are obtained, focus of analysis is shifted to investigating what factors might affect the survival times, that is, modeling survival data. Usually the first step of modeling is to collect data. In most settings that give rise to survival data, in addition to the survival times and the censoring status, supplementary information is also recorded on each subject. For example, subjects may have demographic variables recorded, such as age, behavioral variables, such as smoking history, or physiological variables, such as blood pressure. Such variables may be used as independent variables in explaining the survival times.

The next step of modeling survival data, as in ordinary linear regression analysis, is to create the specific likelihood function to be maximized. In ordinary linear regression analysis, data assume a particular form of probability distribution, while in regression analysis of survival data, usually no particular form of probability distribution is assumed. Hence, how to build the likelihood function for survival data is unique, as is discussed below. To see this, a natural place to begin is to build a likelihood function as if full knowledge of survival data were known.

In survival analysis, regarding the survival experience of subjects in a particular study, each subject either undergoes the occurrence of the event of interest, or that subject's observation is censored, which then contributes to the construction of the likelihood function accordingly. Let an indicator variable δ_i denote the censoring status of the ith subject, with $\delta_i = 1$ for an occurrence case and $\delta_i = 0$ for a censored case. Regarding the occurrence case, its role in constructing the likelihood function is represented by the density function $f(t)$ (Hosmer et al., 2008), quantifying the probability that the ith subject undergoes the event of interest at time t_i. According to the relationship shown in Equation 6, the density function is actually the product of the hazard function and the survival function, yielding

$$f(t) = h(t) \times S(t). \tag{8}$$

Regarding the censored case, its role in constructing the likelihood function is represented by the survival function *S*(*t*) (Hosmer et al., 2008), quantifying the probability the *i*th subject survives longer than some time t_i. Taken together, under the assumptions of independent observations and absolutely continuous event times, the full likelihood function for *N* subjects, each with a vector of covariates $\boldsymbol{X}_i$, is obtained by multiplying the respective contributions of the observed cases over the entire sample,

$$L(\boldsymbol{\beta}, h_0, S_0 | \boldsymbol{X}_i, t_i) = \prod_{i=1}^{N} [h(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i) \times S(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i)]^{\delta_i} [S(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i)]^{1-\delta_i}, \tag{9}$$

where t_i denotes a particular time for the *i*th subject, and $\boldsymbol{\beta}$ denotes the vector of unknown regression regression parameters. It is interesting to notice that the construction of the likelihood function for survival data in Equation 9 is analogous to that for the familiar Bernoulli distribution, which actually makes sense, in that the survival experience of a particular subject is like one Bernoulli trial, with the outcome either being the censored case or the occurrence case. It was thus anticipated that this link between survival data and the familiar Bernoulli distribution might have the potential for simplifying the inference procedures for regression analysis of survival data.

To assess the effect of independent variables on the survival experience of subjects, the method of maximum likelihood is applied to the likelihood function in Equation 9. The corresponding log-likelihood function of Equation 9 is

$$
\begin{aligned}
l(\boldsymbol{\beta}, h_0, S_0 | \boldsymbol{X}_i, t_i) = \delta_i \sum_{i=1}^{N} \{\log[h(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i) \times S(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i)]\} \\
+ (1 - \delta_i) \sum_{i=1}^{N} \{\log[S(t_i, \boldsymbol{\beta}, \boldsymbol{X}_i)]\}. \qquad (10)
\end{aligned}
$$

To simplify calculations, Equation 10 is usually maximized, as the maximum of Equation 9 and its corresponding log-likelihood function in Equation 10 occur at the same value for each component of $\boldsymbol{\beta}$ when the log function is monotone.

In summarizing the survival times, except for a particular treatment protocol, supplementary information, such as weight and smoking history, recorded on each subject is not used in the survival function and the hazard function. However, in the context of modeling survival data, a set of independent variables needs to be included into the hazard function and the survival function, as in Equation 9, in order to explore the relationship between the survival times and independent variables.

Different regression models can be built from different perspectives, i.e., different ways of describing how a set of independent variables is related to the survival times. In particular, depending on whether the underlying distribution of the survival times is specified, there are parametric models (Cox & Oakes, 1984) or semi-parametric models (Cox, 1972). Depending on whether the relationship between the baseline hazard function and the hazard function is multiplicative or not, there are multiplicative regression models or additive regression models (Aalen, 1989; Lin & Ying, 1994). If an event of interest can occur multiple times in the course of a subject’s follow-up, there are recurrent events models (Clayton, 1994). If there are factors other than the measured covariates that could significantly affect the distribution of the survival times, there are frailty models, which incorporate random effects into the models (Vaupel, Manton, &

Stallard, 1979). In the current research, survival data were modeled from the perspective of underlying distributions of the survival times.

There are two groups of models depending on whether the underlying distribution of the survival times is specified. In particular, models in which a specific probability distribution is assumed for the survival times are known as parametric models, such as the exponential model, the Weibull model, and the Gompertz model (Lindsey, 1998). As an example, in the Weibull model, which allows for dependence of the hazard on time, the hazard function takes the form $h(t) = \lambda\gamma t^{(\gamma-1)}$, where t denotes a specific point in time, λ denotes the scale parameter, and γ denotes the shape parameter. When $\gamma > 1$, the hazard increases monotonically. Therefore, for a particular study, if researchers firmly believe that the baseline hazard function increases monotonically as time goes on, the Weibull model with $\gamma > 1$ should be employed to model survival data.

In general, if the assumption of a particular probability distribution of the survival times is valid, inferences based on such an assumption will be more precise because of fewer parameters (Klein & Moeschberger, 2005). Nonetheless, justification of using a parametric model in reality will be difficult unless the sample data contain a large number of event times (Collett, 2003). If a parametric model is chosen incorrectly, it may lead to inconsistent estimators of the quantities of interest.

Models in which there is no need to specify a probability distribution of the survival times are known as semi-parametric models, among which the Cox proportional hazards (PH) model (Cox, 1972) is the most commonly applied methodology for assessing the effect of independent variables on the hazard of an event of interest. The term proportional hazards refers to the fact that, when values of all the other variables are fixed at study entry, the hazard rates of two subjects, either with distinct values of the main treatment variable or a covariate, remain constant, independent of time. A key reason for the popularity of the Cox PH model is that, even though a probability distribution for the survival times is not specified, reasonably good estimates of regression coefficients and other quantities of interest, such as hazard ratios, can be obtained for a wide variety of data situations (Kleinbaum & Klein, 2011). In other words, the Cox PH model will closely approximate the results for the correct parametric model. For example, if the correct parametric model is Weibull, then use of the Cox PH model typically will give results comparable to those obtained using the Weibull model.

In summary, researchers may not be completely certain that a given parametric model is appropriate. Thus, when in doubt, as is typically the case, the Cox PH model will give reliable enough results so that it is a “safe” choice of model, and researchers do not need to worry about whether the wrong parametric model is chosen. Therefore, the current research concentrates on the Cox PH model.

There are other reasons for choosing the Cox PH model as the basis for regression analysis in this research. The Cox PH model, which accounts for time-independent covariates, assumes that the effect of a covariate acts multiplicatively on an unknown baseline hazard function, and coefficients are unknown constants whose value does not change over time. Covariates which do not act on the baseline hazard function in this fashion are modeled either by the inclusion of a time-dependent covariate or by stratification (Klein & Moeschberger, 2005). In other words, when external time-dependent covariates are included, the hazards are not proportional across time. An alternative model that does not assume constant hazard ratios is the additive hazard model, which is based on assuming that the covariates act in an additive manner on an unknown baseline hazard function. The unknown coefficients in this model are allowed to be functions of time so that the effect of a covariate may vary over time.

Though the Cox PH model with external time-dependent covariates, i.e., the extended Cox model, and the additive hazard model share the similarity of accounting for varying hazard ratios across time, the former model was chosen over the additive hazard model in the current research for the following two reasons.

First, multiplicative models are extremely useful in practice because either the estimated coefficients themselves or simple functions of them can be used to provide estimates of hazard ratios. To illustrate, in a hypothetical Cox PH model containing sex and age, $h(t) = h_0(t)e^{(\beta_1 * sex + \beta_2 * age)}$, the estimated coefficient for β_1, can easily provide estimate of hazard ratios at a particular age value between males and females using $e^{\hat{\beta}_1}$. While in the additive hazard model, the estimated coefficients, that is, those yielding a positive hazard function, are tightly constrained by the additive form of the model. As such, the hazard ratio from the additive hazard model might take the form, $\frac{1+\beta_1+\beta_2 a}{1+\beta_2 a}$, where a denotes a particular age value. One rather obvious problem with this model is that, if inferences are based on hazard ratios, it is impossible, except in a univariate model, to isolate the effect of a single covariate. Under this model, the difference in the hazard for males and females, at a particular age value a, is $h_0(t)\hat{\beta}_1$, which depends on both the coefficient for sex and the unspecified baseline hazard function. Despite the possible clinical appeal of additive relative hazard models, they are not as practical as multiplicative models, which may be why they have not been used more frequently in applied research (Hosmer et al., 2008).

Second, most studies address multiplicative models. Moreover, statistical software is readily available and easy to use to fit the proportional hazards model, check model assumptions, and assess model fit. The widespread use of the proportional hazards model in applied settings is largely due to these factors (Hosmer et al., 2008).

Formally, assume there are *N* independent observations. Each of the observations contains information on the length of observed time, the censoring status, and a vector of time-independent variables, that is, their values are determined at study entry, and remain at those values throughout the follow-up of the subject. For the *i*th subject, denote the triplet of observed time, a vector of variables, and censoring variable as (t_i, $\boldsymbol{X}_i$, δ_i), $i = 1, 2, \ldots, N$, where $\boldsymbol{X}_i$ denotes a vector of *p* time-independent variables. Those independent variables typically include a variable indicating the main treatment group and other covariates. There are times when there is no experimentally manipulated treatment variable, that is, only covariates are used in modeling survival data (Collett, 2003). Moreover, the model also allows for non-manipulated grouping variables, such as sex, educational level, and ethnicity. Let $h_0(t)$ be the hazard function at time *t* for a subject for whom the values of all the independent variables that make up the vector $\boldsymbol{X}_i$ are zero, or the baseline hazard function. Then the corresponding hazard function at time *t* under the Cox PH model (Cox, 1972) for the *i*th subject can be written as

$$h_i(t|\boldsymbol{\beta}, \boldsymbol{X}_i) = h_0(t)e^{(\boldsymbol{\beta}'\boldsymbol{X}_i)}, \tag{11}$$

where $\boldsymbol{\beta}$ denotes the vector of unknown regression parameters. In the current study, no treatment variable was used in the modeling procedure. The Cox PH model or the extended Cox model is capable of accommodating covariates alone, although proportional hazards originally refer to hazards between different levels of a particular treatment.

As mentioned above, the survival function can be specified through the hazard function. If the relationship shown in Equation 7 is used, then the corresponding survival function is

$$S_i(t) = e^{[-H_i(t)]}, \tag{12}$$

where, under the Cox PH model,

$$\begin{aligned} H_i(t|\boldsymbol{\beta}, \boldsymbol{X}_i) &= \int_0^t h_i(s)\, ds \\ &= \int_0^t h_0(s) e^{(\boldsymbol{\beta}' \boldsymbol{X}_i)}\, ds \\ &= e^{(\boldsymbol{\beta}' \boldsymbol{X}_i)} \int_0^t h_0(s)\, ds \\ &= e^{(\boldsymbol{\beta}' \boldsymbol{X}_i)} H_0(t). \end{aligned} \tag{13}$$

Substituting Equation 13 into Equation 12, the survival function becomes

$$S_i(t|\boldsymbol{\beta}, \boldsymbol{X}_i) = e^{\left[-e^{(\boldsymbol{\beta}' \boldsymbol{X}_i)} H_0(t)\right]}. \tag{14}$$

Thus, it follows that

$$S_i(t|\boldsymbol{\beta}, \boldsymbol{X}_i) = [S_0(t)]^{e^{(\boldsymbol{\beta}' \boldsymbol{X}_i)}}. \tag{15}$$

Similarly, $S_i(t)$ denotes the survival function at time t for the ith subject, $S_0(t)$ denotes the baseline survival function for that subject for whom the values of all the independent variables that make up the vector $\boldsymbol{X}_i$ are zero, and $\boldsymbol{\beta}$ denotes the vector of unknown regression parameters. There are two reasons for deriving the survival function from the hazard function. First, in addition to facilitating the descriptions of the survival experience of subjects during a follow-up study, the survival function, like the hazard function, is a component of the likelihood function in Equation 9. Second, practitioners tend to understand the survival experience of subjects better in that in most applied settings, practitioners are typically, though not always, more interested in describing how long the study subjects live, rather than the risk of how quickly they die.

After independent variables are accommodated using the hazard function and the corresponding survival function under the Cox PH model, the log-likelihood function in Equation 10 becomes

$$l(\boldsymbol{\beta}, h_0, S_0 | \boldsymbol{X}_i, t_i) = \sum_{i=1}^{N} \{\delta_i \log[h_0(t_i)] + \delta_i(\boldsymbol{\beta}'\boldsymbol{X}_i) + e^{(\boldsymbol{\beta}'\boldsymbol{X}_i)} \log[S_0(t_i)]\}. \qquad (16)$$

Unfortunately Equation 16 cannot be maximized without specifying the form for the baseline hazard function. The reason is, as discussed by Kalbfleisch and Prentice (2002), the log-likelihood function in Equation 16 is a function of finite-dimensional regression parameters and infinite-dimensional nuisance parameters, which refer to parameters that are present in a model but are not of direct inferential interest, i.e., the baseline hazard function. Moreover, to obtain estimates of regression parameters by maximizing over the infinite-dimensional parameters is difficult.

Thus, to avoid specifying the baseline hazard function, Cox (1972) proposed using an expression based on the PH model in Equation 11, which he called a “partial likelihood function” due to the fact that the function does not actually use the full data: only the ordering of the survival times, not the actual times an event of interest occurs, is important. In particular, as indicated earlier, assume there are N independent observations, or N subjects, each consisting of the triplet (t_i, $\boldsymbol{X}_i$, δ_i), $i = 1, 2, \ldots, N$. Among those observations or subjects there are r distinct event times in total, and N - r censored survival times, which are assumed right censored. In other words, each of N subjects either experiences the event of interest or is censored. For simplicity, it is assumed that no ties exist among the uncensored event times. The r ordered event times are then denoted by $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$, so that $t_{(j)}$ is the jth ordered event time. Define the risk set, $R(t_i)$, at the event time for the ith subject t_i, as the set of all subjects, indexed by l, who have not experienced the event and thus uncensored at a time just prior to t_i. Further, it is assumed that censoring is non-informative in that, given a vector of covariates, $\boldsymbol{X}_i$, the event and censoring times for the ith subject are independent. Thus the partial likelihood (Cox, 1972) is given by

$$L(\boldsymbol{\beta}|\boldsymbol{X}_i, t_i) = \prod_{i=1}^{N} \left[\frac{e^{(\boldsymbol{\beta}'\boldsymbol{X}_i)}}{\sum_{l \in R(t_i)} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}}\right]^{\delta_i}. \qquad (17)$$

The corresponding log partial likelihood (Collett, 2003) function is given by

$$l[\boldsymbol{\beta}|\boldsymbol{X}_i, t_i] = \sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta}'\boldsymbol{X}_i - \log \sum_{l \in R(t_i)} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}\right].$$

Equation 17 is usually modified to exclude censored cases, that is, for cases with $\delta_i = 0$. Thus the modified partial likelihood function for r distinct ordered event times is

$$L[\boldsymbol{\beta}|\boldsymbol{X}_{(j)}, t_{(j)}] = \prod_{j=1}^{r} \frac{e^{[\boldsymbol{\beta}'\boldsymbol{X}_{(j)}]}}{\sum_{l \in R[t_{(j)}]} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}}, \qquad (18)$$

in which $\boldsymbol{X}_{(j)}$ is a vector of independent variables for the subject who experiences the event of interest at the jth ordered event time, $t_{(j)}$. The summation in the denominator of this likelihood function is the sum of the values of $e^{(\beta'X_l)}$ over all subjects who are at risk at time $t_{(j)}$.

Equation 18 is actually derived by multiplying conditional probabilities over all event times, which could be seen from Equation 19 to Equation 21. First consider the probability, p*, that a subject experiences the event of interest at time $t_{(j)}$, conditional on $t_{(j)}$ being one of the r ordered event times. Using the standard result from conditional probability theory described above, P($B|A$)=P($A \cap B$)/P(A), the conditional probability, given $\boldsymbol{X}_{(j)}$, is expressed as

$$p^* = \frac{\mathrm{P}\big[\text{subject } i \text{ with } \boldsymbol{X}_{(j)} \text{ has event at } t_{(j)}\big]}{\mathrm{P}\big[\text{one event at } t_{(j)}\big]}. \tag{19}$$

It can be seen that the numerator in Equation 19 is the hazard function for the ith subject. To see this, first replace the time point $t_{(j)}$ with the time interval [($t_{(j)}$, $t_{(j)} + \Delta t$)], where Δt denotes a time interval, and next divide the numerator by Δt, and then take the limiting value of the resulting expression as $\Delta t \to 0^+$. That is,

$$h_i\big[t_{(j)}\big] = \lim_{\Delta t \to 0^+}\left(\frac{\mathrm{P}\big\{\text{subject } i \text{ with } \boldsymbol{X}_{(j)} \text{ has event in } \big[t_{(j)}, t_{(j)} + \Delta t\big]\big\}}{\Delta t}\right), \tag{20}$$

which would replace the numerator in Equation 19. Regarding the denominator in Equation 19, since the event times are assumed to be independent of one another, the denominator is the sum of the probabilities of the event at time $t_{(j)}$ over all subjects who are at risk at that time. With $R[t_{(j)}]$ denoting the risk set, the denominator becomes {$\sum_{l \in}$

$R[t(j)]$ P[subject l has event at $t_{(j)}$]}. In the same vein, it can be seen that if the time point $t_{(j)}$ in the expression for probability is replaced with the time interval $[(t_{(j)}, t_{(j)} + \Delta t)]$, the denominator is divided by Δt, and then the limiting value of the resulting expression is taken as $\Delta t \to 0^+$. The result is the sum of the hazard function at $t_{(j)}$ over all subjects who are at risk at that time. Therefore, the denominator in Equation 19 becomes $\{\sum_{l\in R[t(j)]} h_l$ $[t_{(j)}]\}$. Substituting Equation 20 and the new expression for the denominator into Equation 19, the conditional probability p* becomes

$$p^* = \frac{h_i[t_{(j)}]}{\left\{\sum_{l\in R(t_{(j)})} h_l[t_{(j)}]\right\}}. \tag{21}$$

On using the Cox PH model in Equation 11, the baseline hazard function in the numerator and denominator in Equation 21 cancels out, and the part regarding conditional probabilities in Equation 18 is obtained. Finally, by taking the product of these conditional probabilities over the r distinct event times, Equation 18 above is obtained. That is,

$$\begin{aligned}\prod_{j=1}^{r} \frac{h_i[t_{(j)}]}{\sum_{l\in R(t_{(j)})} h_l[t_{(j)}]} &= \prod_{j=1}^{r} \frac{h_0[t_{(j)}]e^{[\boldsymbol{\beta}'\boldsymbol{X}_{(j)}]}}{\sum_{l\in R[t_{(j)}]} h_0[t_{(j)}]e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}} \\ &= \prod_{j=1}^{r} \frac{h_0[t_{(j)}]e^{[\boldsymbol{\beta}'\boldsymbol{X}_{(j)}]}}{h_0[t_{(j)}]\left\{\sum_{i\in R[t_{(j)}]} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}\right\}} \\ &= \prod_{j=1}^{r} \frac{e^{[\boldsymbol{\beta}'\boldsymbol{X}_{(j)}]}}{\sum_{l\in R[t_{(j)}]} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}}.\end{aligned}$$

The corresponding log partial likelihood function (Hosmer et al., 2008) takes the following form

$$l\left[\boldsymbol{\beta}|\boldsymbol{X}_l, t_{(j)}\right] = \sum_{j=1}^{r} \boldsymbol{\beta}'\boldsymbol{X}_{(j)} - \sum_{j=1}^{r} \log\left\{\left[\sum_{l\in R\left[t_{(j)}\right]} e^{(\boldsymbol{\beta}'\boldsymbol{X}_l)}\right]\right\}. \tag{22}$$

The maximum likelihood estimate of each component of $\boldsymbol{\beta}$ in the PH model can be found by maximizing Equation 22 using numerical methods, such as the Newton-Raphson algorithm.

After regression parameter estimates are obtained, the next step naturally in inferential statistics is to estimate their standard errors, which are obtained in the same manner as standard error estimators are obtained in most maximum likelihood estimation procedures. In particular, the first step is to get the variance estimator by taking the inverse of negative second derivatives of the log partial likelihood at the value of the parameter estimator (Kalbfleisch & Prentice, 2002). Formally, letting $\mathbf{I}(\boldsymbol{\beta})$ be the p by p matrix of negative second derivatives of the log partial likelihood, where p is the number of parameters in the Cox PH model, the (g, h)th element of $\mathbf{I}(\boldsymbol{\beta})$ is

$$\mathbf{I}(\boldsymbol{\beta})_{g,h} = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_g \partial\beta_h}\bigg|_{\hat{\boldsymbol{\beta}}}, \qquad g, h = 1, \dots, p.$$

The matrix $\mathbf{I}(\boldsymbol{\beta})$ is called the partial likelihood observed information matrix. Thus the variance estimator is $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$, where $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates. And the estimator of the standard error, denoted $\widehat{\mathrm{SE}}(\hat{\boldsymbol{\beta}})$, is the positive square root of each diagonal of the variance estimator. That is,

$$\widehat{\mathrm{SE}}(\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})}.$$

Letting $w = \frac{e^{\left(\hat{\boldsymbol{\beta}}^{'}\boldsymbol{X}_l\right)}}{\sum_{l \in R\left[t_{(j)}\right]} e^{\left(\hat{\boldsymbol{\beta}}^{'}\boldsymbol{X}_l\right)}}$, the above equation can be expressed in scalar notation (Klein & Moeschberger, 2005) as

$$\widehat{\mathrm{SE}}(\hat{\boldsymbol{\beta}}) = \sqrt{\left\{\sum_{j=1}^{r}\left[w \sum_{l \in R(t_{(j)})} X_{lg}X_{lh}\right] - \sum_{j=1}^{r}\left[w \sum_{l \in R(t_{(j)})} X_{lg}\right]\left[w \sum_{l \in R(t_{(j)})} X_{lh}\right]\right\}^{-1}}$$

$$= \sqrt{\left\{w \sum_{j=1}^{r} \sum_{l \in R(t_{(j)})} X_{lg}X_{lh} - w^2 \sum_{j=1}^{r}\left[\sum_{l \in R(t_{(j)})} X_{lg}\right]\left[\sum_{l \in R(t_{(j)})} X_{lh}\right]\right\}^{-1}}.$$

Moreover, in inferential statistics, after the parameter estimates and their standard errors are obtained, hypothesis testing is performed to assess the significance of the parameter estimates. Still, as parameter estimates from the Cox PH model are obtained via the maximum likelihood method, hypothesis testing is based on large-sample likelihood theory. Three such tests are the partial likelihood ratio test, the Wald test, and the score test.

The partial likelihood ratio test, denoted *G*, is calculated as twice the difference between the log partial likelihood of the model containing the independent variables and the log partial likelihood for the model not containing the independent variables. Formally,

$$G = 2\left\{L(\hat{\boldsymbol{\beta}}) - L(\mathbf{0})\right\},$$

where $\hat{\boldsymbol{\beta}}$ is a vector of maximum log partial likelihood parameter estimates and **0** is a vector of zeroes. Assuming large samples, this statistic follows an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis that H_0: $\boldsymbol{\beta} = \mathbf{0}$, where p is the difference in the number of parameters between the null model and the alternative model, and thus can be used to obtain p-values to test the significance of $\boldsymbol{\beta}$.

The Wald test, in its multiple variable version, is expressed as

$$\mathrm{Z}^2 = (\widehat{\boldsymbol{\beta}} - \mathbf{0})' \boldsymbol{I}(\widehat{\boldsymbol{\beta}})\,(\widehat{\boldsymbol{\beta}} - \mathbf{0}),$$

where $\widehat{\boldsymbol{\beta}}$ is a vector of maximum log partial likelihood parameter estimates, $\mathbf{0}$ is a vector of zeroes, and the matrix $\boldsymbol{I}(\widehat{\boldsymbol{\beta}})$ is the observed information matrix evaluated at the vector of parameter estimates. Assuming the same mathematical assumptions required for the log partial likelihood ratio test stated above, the Wald statistic asymptotically follows a chi-squared distribution with p degrees of freedom under the null hypothesis that H_0: $\boldsymbol{\beta} = \mathbf{0}$.

The score test is based on the efficient score statistics. Let $\mathrm{U}(\boldsymbol{\beta})$ be the $p \times 1$ vector of first derivatives of the log-likelihood function in Equation 22 with respect to each component of $\boldsymbol{\beta}$. This quantity is known as the vector of efficient scores. Under the null hypothesis that H_0: $\boldsymbol{\beta} = \mathbf{0}$, the vector of efficient scores $\mathrm{U}(\mathbf{0})$ has a large-sample multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix given by the information matrix evaluated at the coefficient vector equal to zero, that is, $\mathbf{I}(\mathbf{0})$. Thus the score test statistic is

$$\mathrm{S}^2 = \mathbf{U}(\mathbf{0})'\mathbf{I}^{-1}(\mathbf{0})\mathbf{U}(\mathbf{0}).$$

Again, assuming the same mathematical assumptions required for the log partial likelihood ratio test stated above, this statistic has an asymptotic chi-squared distribution with p degrees of freedom under the null hypothesis that H_0: $\boldsymbol{\beta} = \mathbf{0}$.

Summary of Features of Survival Analysis

A basic introduction to survival data analysis presented above reveals that there is one salient feature in survival data that is difficult to handle with conventional statistical methods, namely incomplete observation of time, i.e., censoring in the current research. Different mechanisms of censoring render survival analysis even more challenging. In modeling survival data, the popular Cox PH model, although it does not require a probability distribution for the survival times to be specified and still gives reliable enough results, allows for right-censored data alone. In other words, the Cox PH model cannot directly accommodate left-censored and arbitrarily interval-censored survival data. However, from the definition of arbitrarily interval-censored data, it is obvious that compared to the right-censoring mechanism, this type of censoring mechanism provides more information regarding when the event of interest occurs.

Besides censoring, time-dependent covariates pose another challenge in survival analysis. Although the extended Cox model can handle external time-dependent covariates, it assumes survival data are right-censored alone. Thus, from a theoretical perspective, it is worth studying how to model arbitrarily interval-censored data with external time-dependent covariates.

Interval-censored Data

Right-censored Data under Scrutiny

In the above introduction to regression analysis of survival data using the Cox PH model, it was assumed that survival data are right-censored, that is, the event time of interest is observed either exactly at or later than the pre-specified study end time for all subjects. Regarding the data-generating process (DGP) for right-censored survival data, there are actually three scenarios. In the first scenario, for subjects, indexed by *i*, who have already experienced the event of interest by the end of the study, their event times are known exactly. The study end time is not restricted to the pre-specified study end time. For a particular subject, it might be the end of the follow-up period, which is prior to the pre-specified study end time. Using the indicator variable in Equation 9, cases with δ_i = 1arise. In the second scenario, for subjects who have not experienced the event of interest at the pre-specified study end time, their survival times are not observed exactly, but are known to be greater than the pre-specified study end time, i.e., they are right-censored. Thus, cases with $\delta_i = 0$ arise. In the third scenario, for subjects who have not experienced the event of interest at their last follow-up, which is prior to the pre-specified study end time, all that is known is that their survival times are at least as long as the time associated with their last follow-up, and then they are either lost to follow-up or withdraw from the study for some reasons, i.e., they are deemed as right-censored. Similarly, cases also with $\delta_i = 0$ arise. As a matter of fact, survival analysis is extremely useful for studying many different kinds of events including disease onset, equipment failures, earthquakes, automobile accidents, and stock market crashes.

Justifying the Use of Interval-censored Data

In constructing the partial likelihood function in Equation 18 above, cases with $\delta_i = 0$ were excluded. The reason is that Equation 18 is constructed from Equation 17, where each component of the product with $\delta_i = 0$ is equal to 1, and thus there is no need to include cases with $\delta_i = 0$ in Equation 18. So data from uncensored observations were actually collected, that is, cases with $\delta_i = 1$, which, under the right-censoring mechanism, contain for each uncensored observation, duration in survival time, an exact event time, which is the examination when the status of the event is found to have changed, and a vector of covariates. Among the three pieces of information, duration in survival time is incidental in nature as the partial likelihood function, relying on vectors of covariates at ordered exact event times, does not make direct use of the actual length of survival times. However, in reality, for each uncensored observation more detailed information than those two pieces actually used in the Cox PH model is collected. In particular, it is common practice to collect survival data on a regular basis from each subject after entry into a follow-up study. Suppose there are η_i examinations for the ith subject, which are denoted by $t_1 < t_2 < \cdots < t_{\eta_i}$, so that t_{η_i} is the ηth examination time. For each case with $\delta_i = 1$, the change of status for that subject is known to occur between two adjacent examination times, such as between t_5 and t_6. If the last examination time alone is used, such as t_6, to specify when the event of interest occurs in order to rank event times, as in applying the Cox PH model above, the specification is not informative compared to if the event time is set between two adjacent examination times, that is, under the interval-censoring mechanism, as using t_6 will give a false impression that the event occurs at the time point t_6 instead of between t_5 and t_6.

As an example, consider one hypothetical study in which the aim is to model survival times among patients admitted to a hospital for a serious disease. Suppose patients who were discharged healthy from the hospital, with discharge deemed as study entry, were examined every three months to ascertain their health status. One patient was tested negative until the ninth month and positive at the 12th month. Apparently, the status of the disease changed between the ninth and the 12th months, yet the exact time of change of status is not clear. Under the right-censoring mechanism, the information includes the length of 12 months in survival time, the 12th month as the exact event time, and a vector of covariates whose values are recorded at study entry and remain constant during the entire follow-up. The information obtained is good enough for using the Cox PH model in Equation 18. However, taking a closer look at the exact event time, i.e., the 12th month, it is found that it is a simplified specification of the event time, as it is known that the status of the disease changed between the ninth and the 12th months, rather than at the 12th month exactly. Although this treatment does not provide a more informative picture of the survival experience of that patient in terms of the specified event time, it is still common practice in reality due to popularity of the Cox PH model which handles right-censored alone as well as reliability of the resulting parameter estimates.

If the interval-censoring mechanism is used instead to specify the event time, that is, (9, 12], this advantage of describing the event time more informatively, conditional on the fact that the status of the event has not changed from study entry to the ninth month, leads to the problem that the Cox PH model in Equation 18 is unable to account for the more informatively specified event time. In particular, from Equation 22, it can be seen that when $\delta_i = 1$, the ordered event time index j dictates specifically who is in the risk set at the jth ordered event time, which is from the original examination time, and the values of the independent variables to be used accordingly. If the alternative specification is used to denote an event time, two indices for specifying examination times must be employed, such as t_5 and t_6, which Equation 22, using distinct ordered event times, cannot handle, assuming some such time specifications may overlap and vary in length. It is because of this difficulty and loss of information due to simplified specification of an event time under the right-censoring mechanism that motivate some new methods of modeling more informatively specified survival data.

In summary, while it is common practice to monitor each subject on a regular basis to keep track of evolution of an event after entry into a follow-up study, how to specify the event time, that is, for cases with $\delta_i = 1$, is handled differently. On one hand, under the right-censoring mechanism, when an event does occur prior to the pre-specified study end time, the last examination time is usually recorded as the exact event time for that subject. On the other hand, under the interval-censoring mechanism, the approach is to bind an unknown event time between the last two examinations when the status of the event is found to have changed. It is evident that while the alternative approach accounts for more informative information regarding when the event of interest occurs, at the same time it complicates the corresponding modeling procedures, as the standard partial likelihood function under the Cox PH model is not compatible with the new way of specifying the event time.

Definition of Arbitrarily Interval-censored Data

For an occurrence case, that is, $\delta_i = 1$, let τ_i be the unobservable event time, A_i and B_i be two examination times forming the time-interval $(A_i, B_i]$ for the ith subject, $i = 1, 2,\ldots, N$, where A_i might or might not be the first examination after study entry, and B_i another examination following A_i prior to or at the pre-specified study end time. Thus for $\delta_i = 1$, if τ_i is bound in the time-interval $(A_i, B_i]$, interval-censored survival data arise. Both left-censored survival data and right-censored survival data are actually special cases of interval-censored survival data. In particular, for $\delta_i = 1$, if $A_i = 0$ and $B_i \neq \infty$, left-censored survival data arise. For $\delta_i = 0$, if $A_i \neq 0$ and $B_i = \infty$, right-censored survival data arise.

The values of A_i and B_i may or may not be the same across the cohort in a particular study. When it is further assumed that examination times may well be different for each subject in the study, arbitrarily interval-censored data arise. This type of survival data is the focus of the current research.

Regression Analysis of Arbitrarily Interval-censored Survival Data

Introduction

In conducting regression analysis of arbitrarily interval-censored data, information is collected regarding independent variables, the status of an event, and examination times either for confined data, for left-censored data, or for right-censored data.

For the time being, values of those independent variables are treated as time-independent, that is, the values taken by such variables are those recorded at study entry and remain unchanged throughout the follow-up.

As mentioned earlier, the Cox PH model is the most frequently used model in describing the relationship between the hazard of an event and independent variables. However, the primary problem in fitting the standard Cox PH model to arbitrarily interval-censored data is that the standard partial likelihood formulation in Equation 18 is not easily adapted. As information collected for a survival analysis will also include left-censored and right-censored data, some integrated approach must be used to accommodate the advantage of describing an event time more informatively through an interval rather than through an exact examination time, introduced by arbitrarily interval-censored data on one hand, and to handle left-censored and right-censored data at the same time on the other hand.

Review of Previous Approaches to Analyzing Arbitrarily Interval-censored Data

In the mid-1980s, many articles about conducting regression analysis of arbitrarily interval-censored data began to appear (Sun, 2006). I reviewed some of those approaches, regarding how likelihood functions were constructed, what the estimation methods were used, what the properties of the estimators were, and how hypothesis testing was performed.

Four types of approaches could be found in the literature. First, the seminal article by Finkelstein (1986) is the first that studied the use of the Cox PH model for interval-censored data. Her method is based on the full likelihood under the Cox PH model and required estimation of the underlying baseline survival function. Regarding estimating regression parameters, the approach uses the difference in the survival functions, specified through the Cox PH model, at two consecutive examinations, as the basis for constructing the likelihood function for each subject. In particular, for $i = 1, 2,\ldots, N$, letting $(L_i, R_i]$ denote the interval during which an event of interest occurred, the likelihood function is

$$L = \prod_{i=1}^{N} [S_i(L_i) - S_i(R_i)], \tag{23}$$

where $S_i(t)$ is the survival function in Equation 15. Let $s_0, s_1, \ldots, s_m$ correspond to the examination times of a follow-up study. From the set of L_i and R_i, the set of times, $0 = s_0$

$< s_1 < \cdots < s_m = \infty$, is determined, such that each L_i and R_i is contained in the set. Define ξ_j = log[-log $S_j(t)$]. Using Equation 15, the likelihood function in Equation 23 can be rewritten as

$$L(\boldsymbol{\beta}|\xi_j) = \sum_{i=1}^{N} \log \sum_{j=1}^{m} \alpha_{ij} \left(e^{\left\{-e^{\left[\boldsymbol{\beta}'\mathbf{X}_i+\xi_{(j-1)}\right]}\right\}} - e^{\left[-e^{\left(\boldsymbol{\beta}'\mathbf{X}_i+\xi_j\right)}\right]} \right), \tag{24}$$

where $\alpha_{ij} = 1$ if $(s_{(j-1)}, s_j]$ is a subset of $(L_i, R_i]$, and $j = 1, 2, \ldots, m - 1$. Then, the maximum likelihood method is applied. Unlike the partial likelihood function in Equation 18, however, the likelihood function involves both unknown regression parameters and the baseline survival function at consecutive examination times. In terms of the asymptotic properties of parameter estimates, they are consistent and efficient (Huang & Wellner, 1997). With regard to the baseline survival function, it could be estimated either by the maximum likelihood method, or by some non-parametric approach, such as the Breslow estimator (1972), which takes the form

$$\hat{S}_{0,B} = \prod_{j=1}^{r} \left\{ 1 - \frac{d_j}{\sum_{l \in R[t_{(j)}]} e^{(\hat{\boldsymbol{\beta}}'\mathbf{X}_l)}} \right\},$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood parameter estimates, r is the number of ordered event times, $t_{(j)}$ is the jth ordered event time, d_j is the number of events at time $t_{(j)}$, and $R[t_{(j)}]$ is the risk set at time $t_{(j)}$.

The method had several drawbacks, though. First, it relies on the grouped data assumption, that is, grouping intervals, the determination of which depends on observed data, are identical for all subjects. Second, this full likelihood approach directly estimates the finite-dimensional regression parameters and the infinite-dimensional nuisance parameter simultaneously. Moreover, since the number of parameters to be estimated may increase with the number of event times, this could be numerically unstable and computationally intensive for some data sets (Goggins, Finkelstein, Schoenfeld, & Zaslavsky, 1998).

In a second method, called "the marginal likelihood approach," each finite censoring interval is regarded as missing, and is replaced by an imputed exact event time. Then a standard method, such as the Cox PH model, is used to analyze the imputed data. In particular, the marginal likelihood approach was originally for right-censored data (Kalbfleisch & Prentice, 1973). In order to extend it into arbitrarily interval-censored data, observed censoring intervals must be converted to ranked event times, as required in the Cox PH model. Using the imputed set of rankings *R*, which are consistent with the observed censoring intervals, from the set of all possible such rankings of the ordered observations for subject *i* to *N*, denoted by φ, the marginal likelihood function *ι* takes the form,

$$\iota(R|\boldsymbol{\beta},\boldsymbol{X}_i)=\sum\nolimits_{R\in\varphi}\mathrm{P}(R|\boldsymbol{\beta},\boldsymbol{X}_i),$$

where $\mathrm{P}(R|\boldsymbol{\beta},\boldsymbol{X}_i)$ is the probability of the ranking *R* in the standard Cox PH model, given the vector of regression parameters $\boldsymbol{\beta}$ and a set of independent variables $\boldsymbol{X}_i$, for all of the observations.

Satten (1996) proposed a Gibbs sampler procedure for generating underlying rankings from the set φ. Gibbs sampler is a technique for generating random variables from a marginal distribution indirectly, without having to calculate the density (Casella &

George, 1992). Maximum marginal likelihood estimates of $\boldsymbol{\beta}$ could be obtained as usual by solving a score function.

Satten, Datta, and Williamson (1998) used a parametric model for the imputation of missing exact or right-censored failure times, and then obtained parameter estimates by solving estimating equations through Monte Carlo techniques that are the partial likelihood score equations for the full-data Cox PH model, averaged over all rankings of imputed event times consistent with the observed censoring intervals. They presented a recursive stochastic approximation scheme that converges to the zero of the estimating equations. The resulting parameter estimates were proven to be consistent and asymptotically normal (Satten et al., 1998).

Goggins, Finkelstein, Schoenfeld, and Zaslavsky (1998) proposed a Monte Carlo expectation maximization (MCEM) algorithm for fitting the Cox PH model for interval-censored data. The basic idea of an EM algorithm is to replace one difficult likelihood maximization with a sequence of easier maximizations whose limit is the answer to the original problem, and this algorithm is guaranteed to converge to the maximum likelihood estimator (Dempster, Laird, & Rubin, 1977; Wu, 1983). The algorithm generates orderings of the events from their probability distribution under the model. Goggins et al. (1998) then maximized the average of the log-likelihoods from these completed data sets to obtain updated parameter estimates. As with the standard Cox PH model, this algorithm does not require the estimation of the baseline hazard function.

Pan (2000) proposed a two-step approach where during the first step multiple imputation of missing event times based on the Breslow estimator of the survival function was conducted. Poor Man's data augmentation (PMDA; Wei & Tanner, 1991) or

Asymptotic Normal data augmentation (ANDA; Wei & Tanner, 1991) was used to impute exact event times from the interval-censored data. During the second step a standard statistical procedure for right-censored data, such as the partial likelihood approach, was applied to imputed data to update the estimates.

A disadvantage of combining multiple imputation and methods developed for right-censored data from the second approach is that they are highly computationally demanding and the fact that the procedures used to impute missing data have a relatively ad hoc nature.

A third class of methods is a trade-off approach that lies between the first approach, which directly estimates the finite dimensional regression parameters and the infinite-dimensional nuisance parameter simultaneously, and the second approach, which focuses only on the finite-dimensional regression parameters (Betensky, Lindsey, Ryan, & Wand, 2002; Cai & Betensky, 2003). This approach approximates the infinite-dimensional nuisance parameter using some smooth, finite-dimensional parameters. Betensky et al. (2002) considered approximating the baseline hazard function using some smooth, regression parameters by applying a local likelihood to fit the Cox PH model to arbitrarily interval-censored data. Interval-censored observations contribute to the baseline hazard function terms of the form

$$\ln\left\{\int_{A_i}^{B_i} h_i(t)e^{\left[-\int_0^t h_i(u)\,du\right]}\,dt\right\},$$

where $(A_i, B_i]$, $i = 1, 2, \ldots, N$, is the interval containing the event time τ_i. To obtain a smoothed estimate of the hazard function, Betensky et al. (2002) proposed a local EM algorithm. In particular, this algorithm iterates between the E-step, in which they calculate the conditional expectations of the local log likelihoods, given the observed data and the current estimate of the hazard function, and the M-step, in which these expected log likelihoods are maximized with respect to their parameters. On the other hand, this method requires manual entry of a bandwidth parameter that determines the amount of smoothing for the hazard function estimate (Betensky et al., 2002). Further, the analytic standard errors are not derived, necessitating the use of the bootstrap, which is quite computationally intensive in this setting (Cai & Betensky, 2003). Lastly, there are numerical stability problems with local likelihood in regions of sparse data, such as the right-hand tail of the hazard function. For the same problem, Cai and Betensky (2003) proposed a penalized spline-based approach. Basically, they weakly parameterized the log-hazard function with a piecewise-linear spline and provided a smoothed estimate of the hazard function by maximizing the penalized likelihood through a mixed model-based approach. One disadvantage of this approach is that the variability due to the estimation of the smoothing parameter for small samples seems out of reach in the frequentist framework from the data.

An advantage of these methods is that predictive survival and hazard curves are directly available, and moreover, they are smooth rather than stepwise as in the case of non-parametric or semi-parametric estimation (Betensky et al., 1998; Cai & Betensky, 2003; Kooperberg & Clarkson, 1997).

A fourth class of approaches takes a different path than the other three classes, in that this class considers the occurrence of an event as a response from one Bernoulli trial with only two possible outcomes; thus having the potential for placing regression analysis of survival data under the framework of the binomial distribution and logistic regression, which is conceptually simpler than the other three classes. Thus, this class of approaches was the focus of the current research.

In particular, as mentioned earlier, in constructing the likelihood function for regression analysis of survival data, Equation 9 is analogous to that for the familiar Bernoulli distribution, which might have the potential for simplifying the inference procedures for survival analysis. Thus, the fourth class of methods treats the problem of how arbitrarily interval-censored data may be fit as a binary response regression problem. Carstensen (1996) and Farrington (1996) considered this approach from different perspectives regarding how to construct the likelihood function. Farrington's method was illustrated in the current research.

In particular, under the Cox PH model, Farrington (1996) constructed the likelihood function based on the familiar Bernoulli distribution. The occurrence of an event bound in a time interval is treated as a second Bernoulli trial conditional on the fact that there has been no occurrence of the event from a first Bernoulli trial until the start of that time interval. In this way, survival analysis of arbitrarily interval-censored data is connected with a binary response regression problem. Parameter estimates are obtained from the resulting generalized non-linear model.

Suppose that the event time for the ith of N subjects is observed to occur in the interval $(A_i, B_i]$, where the values of A_i and B_i may well be different for each subject. Further, in the context of regression analysis, the values of a number of independent variables are treated as time-independent.

The survival function for the ith subject is denoted by $S_i(t)$, as in Equation 15, so that the probability of the event occurring in the interval $(A_i, B_i]$, is $S_i(A_i)$ - $S_i(B_i)$. The corresponding likelihood function for the N independent observations then takes the following form

$$L(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{N}[S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i) - S_i(B_i|\boldsymbol{\beta},\boldsymbol{X}_i)], \tag{25}$$

where $\boldsymbol{\beta}$ denotes the vector of unknown regression parameters, and $\boldsymbol{X}_i$ denotes a vector of covariates. Now suppose that the N independent observations consist of l left-censored observations, r right-censored observations, and a observations that are interval-censored, where $N = l + r + a$. For the purpose of illustration, it will be assumed that the data have been arranged in such a way that the first l observations are left-censored, i.e., $A_i = 0$, the next r are right-censored, i.e., $B_i = \infty$, and the remaining a observations are interval-censored, i.e., $0 < A_i < B_i < \infty$. Since $S_i(0) = 1$ and $S_i(\infty) = 0$, the contributions of a left-censored and right-censored observation to the likelihood function will be 1- $S_i(B_i)$ and $S_i(A_i)$, respectively. Thus the overall likelihood function, denoted L*, can be written as

$$\begin{aligned} L^*(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{l}[1 - S_i(B_i|\boldsymbol{\beta},\boldsymbol{X}_i)] \times \prod_{i=l+1}^{r}[S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i)] \\ \times \prod_{i=l+r+1}^{N}[S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i) - S_i(B_i|\boldsymbol{\beta},\boldsymbol{X}_i)], \end{aligned} \tag{26}$$

or equivalently,

$$\begin{aligned} L^*(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{l}[1 - S_i(B_i|\boldsymbol{\beta},\boldsymbol{X}_i)] \times \prod_{i=l+1}^{r}[S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i)] \\ \times \prod_{i=l+r+1}^{N} S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i)\left[1 - \frac{S_i(B_i|\boldsymbol{\beta},\boldsymbol{X}_i)}{S_i(A_i|\boldsymbol{\beta},\boldsymbol{X}_i)}\right]. \end{aligned} \tag{27}$$

It can be shown that this likelihood is equivalent to that for a corresponding set of $N + a$ independent binary observations, $y_1, y_2 \ldots, y_{(N+a)}$, where the ith is assumed to be an observation from a Bernoulli distribution with the response probability p_i, i =1, 2, ... , $N + a$. The likelihood function, denoted L^{**}, for this set of binary data is then

$$L^{**}(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{N+a} p_i^{y_i}(1 - p_i)^{1-y_i}, \tag{28}$$

where y_i takes the value 0 or 1, for i =1, 2, ... , $N + a$. The relationship can be established as follows. For left-censored data, the event of interest occurs before the first examination, and thus each of these l observations, which can be thought of as having one Bernoulli trial, contributes a binary observation with $y_i = 1$ and

$$\begin{aligned} p_i &= S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i) \\ &= S_i(0) - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i) \\ &= 1 - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i), \end{aligned}$$

as each left-censored observation is confined between study entry and the first examination and $S_i(0) = 1$, i =1, 2, ... , l. The contribution of these l observations can be expressed as

$$\prod_{i=1}^{l} p_i = \prod_{i=1}^{l} [1 - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)]. \tag{29}$$

For right-censored data, the event of interest does not occur until after the last examination, and thus each of these r observations, which can be thought of as having one Bernoulli trial as well, contributes a binary observation with $y_i = 0$ and

$$\begin{aligned} p_i &= 1 - [S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)] \\ &= 1 - [S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) - S_i(\infty)] \\ &= 1 - S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i), \end{aligned}$$

as each right-censored observation will not experience the event of interest until after the end of the follow-up study and $S_i(\infty) = 0$, $i = l + 1 , l + 2, \ldots , l + r$. The contribution of these r observations can be expressed as

$$\prod_{i=l+1}^{l+r} (1 - p_i) = \prod_{i=l+1}^{l+r} S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i). \tag{30}$$

For interval-censored data, two Bernoulli trials are needed for the occurrence within the interval $(A_i, B_i]$, where $A_i \neq 0$ and $B_i \neq \infty$. The overall probability can be expressed as

$$\mathrm{P}(\text{no event before } A_i) \times \mathrm{P}(\text{event from } A_i \text{ to } B_i | \text{no event before } A_i).$$

The first trial happens during the time interval $(0, A_i]$, where the event of interest does not occur, that is, $y_i = 0$ and the probability that no event occurs before A_i is $1 - p_i = S_i(A_i)$. The second trial happens during the time interval $(A_i, B_i]$, where the event occurs, that is, $y_{i+a} = 1$ and the corresponding probability p_{i+a} can be expressed as

$$\mathrm{P}(\text{event from } A_i \text{ to } B_i | \text{no event before } A_i),$$

which is actually a conditional probability. Since P(no event before A_i) = P(event after A_i) = $S_i(A_i)$, that is, the probability of a non-occurrence case before the time point A_i is equal to that of an occurrence case after the time point A_i,

$$\begin{aligned}
\mathrm{P}(\text{event from } A_i \text{ to } B_i | \text{event after } A_i) &= \frac{\mathrm{P}[(\text{event after } A_i) \cap (\text{event after } A_i)]}{\mathrm{P}(\text{event after } A_i)} \\
&= \frac{\mathrm{P}(\text{event after } A_i) \times \frac{\mathrm{P}(\text{event from } A_i \text{ to } B_i)}{\mathrm{P}(\text{event after } A_i)}}{\mathrm{P}(\text{event after } A_i)} \\
&= \frac{\mathrm{P}(\text{event from } A_i \text{ to } B_i)}{\mathrm{P}(\text{event after } A_i)} \\
&= \frac{[S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) - S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)]}{S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i)} \\
&= 1 - \frac{S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)}{S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i)}.
\end{aligned}$$

Combining these two terms leads to the expression of the form

$$\prod_{i=l+r+1}^{N} (1 - p_i) p_{i+a} = \prod_{i=l+r+1}^{N} S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) \left[1 - \frac{S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)}{S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i)}\right], \tag{31}$$

where p_{i+a} denotes the response probability for the second trial in each of the confined cases.

Taken together, this shows that by suitably defining a set of $N + a$ binary observations with response probabilities expressed in terms of the survival functions for the three possible forms of interval-censored observation, the likelihood function in Equation 28 is equivalent to that in Equation 27. Regarding how Equation 27 and Equation 28 are related to Equation 31, Equation 27 is the full likelihood function, which accounts for left-censored cases, right-censored cases, and interval-censored cases, while Equation 31 is one component that only accounts for interval-censored cases in Equation 27. As the full likelihood function in Equation 27 is equivalent to that for a corresponding set of independent binary observations from Bernoulli trials with the response probability p_i, that is, Equation 28, Equation 31 corresponds to a component of the likelihood function in Equation 28 of the form, $\prod_{i=l+r+1}^{N}(1-p_i)p_{i+a}$. Accordingly, maximization of the log-likelihood function for $N+a$ binary observations is equivalent to maximizing the log-likelihood for the interval-censored data.

The next step is to construct expressions for the survival functions that make up the likelihood function in Equation 27. Recall Equation 15,

$$S_i(t|\boldsymbol{\beta},\boldsymbol{X}_i) = [S_0(t)]^{e^{(\boldsymbol{\beta}'\boldsymbol{x}_i)}},$$

where $S_0(t)$ is the baseline survival function and $\boldsymbol{X}_i$ is a vector of values of p independent variables for the ith subject, $i = 1, 2, \ldots, N$, with coefficients that make up the vector of unknown regression parameters, $\boldsymbol{\beta}$.

The baseline survival function will be modeled as a step function, where the steps occur at the k ordered censoring times, $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$, where $0 < t_{(1)} < t_{(2)} < \cdots < t_{(k)}$, which are a subset of the times at which observations are interval-censored. This means that the $t_{(g)}$, $g = 1, 2, \ldots, k$, are a subset of the values of A_i and B_i, $i = 1, 2, \ldots, N$. Now define

$$\theta_g = \log\frac{S_0\left[t_{(g-1)}\right]}{S_0\left[t_{(g)}\right]}, \tag{32}$$

where $t_{(0)} = 0$, so that $\theta_g \geq 0$, and at time $t_{(g)}$, it follows that

$$S_0\left[t_{(g)}\right] = e^{(-\theta_g)}S_0\left[t_{(g-1)}\right], \tag{33}$$

for $g = 1, 2, \ldots, k$.

Since the first step in the baseline survival function occurs at $t_{(1)}$, $S_0(t) = 1$ for $0 \leq t < t_{(1)}$. From time $t_{(1)}$, the baseline survival function, using the above relationship, has the value $S_0[t_{(1)}] = e^{(-\theta_1)}S_0[t_{(0)}]$, which, since $t_{(0)} = 0$, means that $S_0(t) = e^{(-\theta_1)}$, $t_{(1)} \leq t < t_{(2)}$. Similarly, from time $t_{(2)}$, the survival function is $S_0[t_{(2)}] = e^{(-\theta_2)}S_0[t_{(1)}]$, that is, $S_0(t) = e^{[-(\theta_1 + \theta_2)]}$, $t_{(2)} \leq t < t_{(3)}$. Similar expressions for all times can be obtained, until $S_0(t) = e^{[-(\theta_1 + \theta_2 + \cdots + \theta_k)]}$, $t \geq t_{(k)}$. Consequently,

$$S_0(t) = e^{\left(-\sum_{r=1}^{g} \theta_r\right)}, \tag{34}$$

for $t_{(g)} \leq t < t_{(g+1)}$, and so the baseline survival function, at any time t_i, is given by

$$S_0(t_i) = e^{\left(-\sum_{g=1}^{k} \theta_g d_{ig}\right)}, \tag{35}$$

where $d_{ig} = 1$ if $t_{(g)} \leq t_i$ and $d_{ig} = 0$ if $t_{(g)} > t_i$, for $g = 1, 2, \ldots, k$. The quantities d_{ig} will be taken to be the values of k indicator variables, $D_{i1}, D_{i2}, \ldots, D_{ik}$, for the ith observation in the augmented data set (Collett, 2003).

How the augmented data set is formed is detailed as follows. After collected survival data are organized in a data set such that information regarding covariates, left and right censoring times for an interval, and the binary response variable for each subject are recorded using one single line, the data set is expanded by adding a further a line of data, repeating the information for subjects whose intervals are confined, so that the revised data set has $N + a$ observations. The values, for example, y_i, of the binary response variable, Y, are then added. These are such that $Y = 1$ for a left-censored observation, and $Y = 0$ for a right-censored observation. For confined observations, where the data are duplicated, one of the pairs of observations has $Y = 0$ and the other observation $Y = 1$. The values of the D_{ig}, $g = 1, 2, \ldots, k$, will differ at each observation time, t_i.

Now combining the results together, the survival function for the *i*th subject, at times A_i and B_i, can now be obtained. In particular,

$$\begin{aligned} S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i) &= S_0(A_i)^{e^{(\boldsymbol{\beta}' X_i)}} \\ &= \left[e^{\left(-\sum_{g=1}^{k} \theta_g d_{ig}\right)}\right]^{e^{(\boldsymbol{\beta}' X_i)}} \\ &= e^{\left\{\left[-e^{(\boldsymbol{\beta}' X_i)}\right] \sum_{g=1}^{k} \theta_g d_{ig}\right\}}, \end{aligned} \tag{36}$$

where $d_{ig} = 1$ if $t_{(g)} \leq A_i$ and $d_{ig} = 0$, otherwise. Similarly,

$$S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i) = e^{\left\{\left[-e^{(\boldsymbol{\beta}' X_i)}\right] \sum_{g=1}^{k} \theta_g d_{ig}\right\}}, \tag{37}$$

where $d_{ig} = 1$ if $t_{(g)} \leq B_i$ and $d_{ig} = 0$, otherwise.

From Equations 36 and 37 for $S_i(A_i)$ and $S_i(B_i)$, respectively, the response probabilities, p_i, used in Expression 28, can be expressed in terms of the unknown parameters $\theta_1, \theta_2, \ldots, \theta_K$ and the unknown coefficients of the p independent variables in the model, $\beta_1, \beta_2, \ldots, \beta_p$. In particular, for a left-censored observation, $p_i = 1 - S_i(B_i)$, and for a right-censored observation, $p_i = 1 - S_i(A_i)$. In the case of an interval-censored observation, $p_i = 1 - S_i(A_i)$ for one of the two binary observations. For the other,

$$\begin{aligned} p_{i+c} &= 1 - \frac{S_i(B_i|\boldsymbol{\beta}, \boldsymbol{X}_i)}{S_i(A_i|\boldsymbol{\beta}, \boldsymbol{X}_i)} \\ &= 1 - \frac{e^{\left\{\left[-e^{(\boldsymbol{\beta}' X_i)}\right] \sum_{g=1}^{k} \theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\boldsymbol{\beta}' X_i)}\right] \sum_{g=1}^{k} \theta_g d_{2g}\right\}}} \\ &= 1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}' X_i)}\right] \sum_{g=1}^{k} \theta_g d_{ig}\right\}}, \end{aligned} \tag{38}$$

where $d_{1ig} = 1$ if $t_{(g)} \leq B_i$ and $d_{1ig} = 0$ otherwise, and $d_{2ig} = 1$ if $t_{(g)} \leq A_i$, and $d_{2ig} = 0$ otherwise. Consequently, the θ-terms in the numerator for which $t_{(g)} \leq A_i$ cancel with those θ-terms in the denominator, and this gives the probability expression for the other binary observation,

$$p_{i+c} = 1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}, \tag{39}$$

where $d_{ig} = 1$ if $A_i < t_{(g)} \leq B_i$ and $d_{ig} = 0$, otherwise. It then follows that in each case, the response probability can be expressed in the form

$$p_i = 1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}, \tag{40}$$

where $d_{ig} = 1$ if $t_{(g)}$ is in each corresponding interval. In particular, for left-censored data, $t_{(g)}$ is in $(0, B_i]$, for right-censored data, $t_{(g)}$ is in $(0, A_i]$, and for confined data, $t_{(g)}$ is in $(A_{i-c}, B_{i-c}]$, for $g = 1, 2, \ldots, k$, and $d_{ig} = 0$ otherwise.

Thus, the likelihood function in Equation 28 becomes

$$L(\boldsymbol{\beta}|\boldsymbol{X}_i, Y_i) = \prod_{i=1}^{N+a}\left(1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}\right)^{Y_i}\left[1 - \left(1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{x}_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}\right)\right]^{1-Y_i}.$$

This leads to a non-linear model for a set of binary response variables with values y_i, and corresponding response probabilities p_i, found from Equation 40, for $i = 1, 2, \ldots, N + a$. The model contains $k + p$ unknown parameters, namely $\theta_1, \theta_2, \ldots, \theta_k$ and $\beta_1, \beta_2, \ldots, \beta_p$. This model is actually known as a generalized non-linear model, since it is not possible to express a simple function of p_i as a linear combination of the unknown parameters, except in the case where there are no explanatory variables in this generalized non-linear model (Collett, 2003).

For non-linear models, parameter estimates typically do not have closed form. An estimate can be obtained via the nonlinear least squares approach (Bates & Watts, 2007). The consistency and asymptotic normality of parameter estimates can be established using uniform laws of large numbers and the mean value theorem, respectively (Shi, 2012). Alternatively, maximum likelihood estimation, implemented by either the Newton-Raphson procedure or the method of Fisher scoring, can be used, and the resulting parameter estimates are consistent, efficient, and asymptotically normal (Tang, He, & Tu, 2012).

There are several advantages to Farrington's approach. First, it is conceptually simpler to understand than the other three classes of methods, as construction of the likelihood function is based on the familiar Bernoulli distribution. Second, it uses an existing data set and does not need to impute data. Third, it does not introduce smoothing techniques or the MCEM. Therefore, for the current research, Farrington's approach was adopted.

Time-dependent Covariates

Introduction

In the Cox PH model introduced in Equation 11, it is assumed that the hazard depends only on time-independent covariates whose values are those recorded at study entry and remain constant throughout the course of the study, such as weight at baseline, gender, and randomized treatment. As is typical in many studies that generate survival data, subjects are monitored for the duration of the study. During this period, values of certain covariates may be recorded on a regular basis. If only time-independent covariates are used, for example, weight at baseline, recorded at the time origin of a two-year study, this constant value may not provide a better indication of health condition than more recent values of weight, such as weight measured in the fifteenth month. In other words, if in a regression analysis time-dependent covariates whose values evolve along the course of the study are used, a more satisfactory model for the hazard of an event of interest at any given time would be obtained. Two previous studies showed through real data analysis that inclusion of external time-dependent covariates into the Cox model enabled a better understanding of predictors' role in describing dynamically the survival experience of subjects in a follow-up study (Andersen, 1992; Christensen et al., 1986).

Types of Time-dependent Covariates

Time-dependent covariates are usually classified as being either external or internal (Kalbfleisch & Prentice, 2002). The reason why this classification is important is that an internal covariate requires special treatment compared to an external one.

External time-dependent covariates. External time-dependent covariates do not necessarily require a subject to be under direct observation. A standard example is the time of the day or the season of the year, which does not require a subject to be under direct observation. A covariate process is external with respect to the outcome process if the covariate at time t is conditionally independent of all preceding response measurements (Luo, 2011). Let T* denote the random variable of event times, $\boldsymbol{x}_{it}$ denote the covariate vector at time t for the ith subject, and $\boldsymbol{X}_{it} = \{x_{iu}; 0 \leq u < t\}$ denote the covariate history up to t. The formal definition of external time-dependent covariates requires such covariates to satisfy the condition (Kalbfleisch & Prentice, 2002)

$$\mathrm{P}\{u \leq T^* < u + \Delta u | \boldsymbol{X}_u, T^* \geq u\} = \mathrm{P}\{u \leq T^* < u + \Delta u | \boldsymbol{X}_t, T^* \geq u\} \tag{41}$$

for all u and t such that $0 < u \leq t$, as $\Delta u \rightarrow 0$. Hence, the hazard function at time u is influenced by the observed covariate history up to time u by the regression model, but its future path up to any time $t > u$ is not affected by the occurrence of an event at time u (Kalbfleisch & Prentice, 2002).

There are two types of external time-dependent covariates (Aalen, Borgan, & Gjessing, 2008). For a defined time-dependent covariate, the complete path of the covariate is given at the outset of the study, so that the covariate changes in such a way that its value will be known in advance at any future time. Examples include the age of a subject and a planned schedule of treatments. An ancillary time-dependent covariate is the observed path of a stochastic process whose development over time is not influenced by the occurrences of the event being studied. An example of such a covariate would be one that measures airborne pollution as a predictor for the frequency of asthma attacks. In all of these examples it is clear that the value of these external time-independent covariates at any time point is not affected by the true event time.

Internal time-dependent covariates. In contrast, for an internal time-dependent, the condition implied in Equation 41 does not hold (Kalbfleisch & Prentice, 2002). Internal time-dependent covariates relate to a particular subject in a study, and can only be measured while that subject is still under direct observation. Such data usually arise when repeated measurements of certain characteristics are made on a subject over time. Examples include biomarkers and clinical parameters, such as white blood cell count, systolic blood pressure, and serum cholesterol level. There are three important features that complicate statistical analysis with such covariates (Rizopoulos, 2012). The first important characteristic is that internal time-dependent covariates typically require the survival of the subject for their existence, so that the path of these covariates carries direct information about the event process. The second important characteristic is that internal time-dependent covariates are typically measured with error. This measurement error primarily refers to the biological variation induced by the subject rather than to the error induced by the procedure or machinery that determines the value of a covariate. The final important characteristic is that their complete path up to any time is not fully observed. That is, the levels of a time-dependent covariate for a subject are only known at some specific examination, and not in between these examinations.

The nature of time-dependent covariates. How time-dependent covariates are handled in regression analysis of survival data using the extended Cox model, as is described shortly below, depends on the nature of the time-dependence. An internal time-dependent covariate is one where the change of the covariate over time is related to the behavior of the subject. For example, the internal time-dependent covariate white blood cell count increases as one subject begins to eat more tomatoes. An external covariate is one whose path is generated externally (Zhang, 2005). A covariate of this sort, like an ancillary time-dependent covariate, can be the output of a stochastic process that is external to the subject under study and whose probability laws do not involve the parameters in the event time model under study (Kalbfleisch & Prentice, 2002). Ancillary covariates play the role of ancillary statistics for the event time model.

However, the extended Cox model is not appropriate when the time-dependent covariates are of internal nature. To see this, external time-dependent covariates are discussed first. In particular, for external time-dependent covariates, using the same set of notation as that in Equation 41, the conditional survival function for a given covariate history is defined in general by

$$\begin{aligned} S_i(t|\boldsymbol{X}_{it}) &= \mathrm{P}(T^* \geq t|\boldsymbol{X}_{it}) \\ &= e^{\left[-\int_0^t h_i(s|\boldsymbol{X}_{it})\,ds\right]} \\ &= e^{\left[-\int_0^t h_0(s)e^{(\boldsymbol{X}_{it})}\,ds\right]}. \end{aligned}$$

By contrast, the conditional hazard function bears no relationship to the conditional survival function for internal time-dependent covariates, which in fact requires the survival of the subject for its existence. For an internal covariate X_i such as white blood cell count, $S_i[t|\boldsymbol{X}_{it}] = 1$ provided that $X_i(t)$ does not indicate that the subject has died. A measurable value of white blood cell count indicates that the subject is still alive (Fisher & Lin, 1999).

To account for the special features of internal time-dependent covariates, the joint modeling framework for longitudinal and survival data (Faucett & Thomas, 1996; Wulfsohn & Tsiatis, 1997) is needed, which, however, is beyond the scope of the current research. The current study only examined external time-dependent covariates.

Regression Analysis with External Time-dependent Covariates

A defined covariate can vary in a predetermined way, that is, its total path up to any time t, $X_i(t)$, is determined in advance for each subject under study (Kalbfleisch & Prentice, 2002). Therefore, inference can be based on the partial likelihood conditioning on the covariates, as usually done in the case of time-independent covariates. Age of a subject is an example. An ancillary covariate, carrying more randomness covariates, can also be considered as external, since its stochastic process has a distribution that does not involve the parameters of the regression model for survival times (Cortese & Andersen, 2009). An example of such a covariate would be when studying how long someone remains employed, the inflation rate is essentially external to the subject's employment duration.

The classical estimation method. Recall that under an independent right-censoring mechanism, the standard Cox PH model with a vector of time-independent covariates has the form

$$h_i(t|\boldsymbol{\beta}, \boldsymbol{X}_i) = h_0(t)e^{(\boldsymbol{\beta}'\boldsymbol{X}_i)},$$

which can be extended to incorporate external time-dependent covariates. Letting $\boldsymbol{X}_{it}$ be a p-dimensional vector of values of independent variables at time t for the ith subject, $\boldsymbol{\beta}$ the p-dimensional vector of unknown parameters, and $h_0(t)$ the baseline hazard function, the corresponding extended Cox model is written as

$$h_i(t|\boldsymbol{\beta}, \boldsymbol{X}_{it}) = h_0(t)e^{(\boldsymbol{\beta}'\boldsymbol{X}_{it})}, \tag{42}$$

and the partial log-likelihood function of Equation 17 can be generalized to

$$l[\boldsymbol{\beta}|\boldsymbol{X}_{it}] = \sum_{i=1}^{N} \delta_i \left[\boldsymbol{\beta}'\boldsymbol{X}_{it} - \log \sum_{l \in R(t_i)} e^{(\boldsymbol{\beta}'\boldsymbol{X}_{lt})}\right], \tag{43}$$

in which $R(t_i)$ is the risk set at time t, the event time of the ith subject in the study, $i = 1, 2, \ldots, N$, and $\delta_i = 0$ if the survival time of the ith subject is censored and $\delta_i = 1$ otherwise. This expression can then be maximized to obtain parameter estimates.

The estimates of the associated standard errors are obtained in a manner identical to the one described for the Cox PH model, using Equation 43 in place of Equation 22.

And the partial likelihood ratio test, the Wald test, or the score test can be conducted to assess the significance of the coefficient.

However, while the extended Cox model accounts for external time-dependent covariates, it assumes right-censored survival data alone. Hence, in the case of regression analysis of other types of survival data, such as arbitrarily interval-censored or left-censored survival data, with external time-dependent covariates, the extended Cox model is not appropriate. No previous studies have been conducted on modeling arbitrarily interval-censored survival data alone with external time-dependent covariates. For example, both Van Der Laan and Robins' (1998) and Martinussen and Scheike's studies (2002) investigated current status data. Chen, May, Ibrahim, Chu, and Cole (2014) developed a procedure that models left-censored survival data and internal time-dependent covariates. Modeling arbitrarily interval-censored with external time-dependent covariates was the focus of the current research.

Other estimation methods. Besides the traditional maximum partial likelihood approach for estimating parameters for external time-dependent covariates in the extended Cox model, there are other estimation methods.

Murphy and Sen (1991) used a sieve estimation procedure (Grenander, 1981) to estimate a time-dependent coefficient in a Cox-type parameterization of the stochastic intensity of a point process. Weak consistency and asymptotic normality for the sieve estimator were demonstrated by Murphy and Sen (1991). To show weak consistency, the idea is to expand the log-partial likelihood about a point which is close to the true parameter, instead of expanding about the true parameter. To show asymptotic normality, the idea is to use the Skorohod topology on $D[0,1]$ (Billingsley, 1999).

Heinze and Dunkler (2008) used the bias reduction approach proposed by Firth (1993) to obtain the penalized maximum partial likelihood estimates for external time-dependent covariates, such as CYCB1 gene expression, under the extended Cox model. Their approach works best whenever monotone likelihood is encountered, the number of events is unusually small or the number of covariates unusually large. Monotone likelihood occurs in the fitting process of the extended Cox model if at least one parameter estimate diverges to infinity. With very small data sets, their approach tends to underestimate strong effects as opposed to standard maximum likelihood estimation method.

From a theoretical perspective, the above literature review showed that there was a need to model arbitrarily interval-censored data with external time-dependent covariates. From an applied perspective, practitioners also need such a modeling procedure, but one did not yet exist prior to the current study.

As an example, in a study conducted by Hartmann et al. (2012), serial measurement of the cardiovascular biomarker midregion proadrenomedullin (MR-proADM) was collected at study entry, days three, five, and seven, and then the extended Cox model was applied at day 30 to assess risk of lower respiratory tract infection. At the end of each subject’s follow-up, an overall status of the event for the subject, i.e., the event happened or did not happen, was recorded. Apparently, the study used day 30 as the event time for an occurrence case, which most probably was not true. Practitioners thus can only evaluate roughly the actually risk of lower respiratory tract infection at a particular day. As another example, Collett (2003) applied Farrington’s approach to investigate the effect of the combination of chemotherapy and radiotherapy on one type of tumor. The exact time of occurrence of the event of interest was unknown, and the only information available concerned whether or not retraction was identified when a patient visited the clinic. Since the visit times, measured in months, were not the same for each patient, and a number of patients failed to keep appointments, the data are regarded as arbitrarily interval-censored. Moreover, at study entry, each patient was treated with either radiotherapy or the combination of chemotherapy and radiotherapy, and the treatment remained unchanged during the entire follow-up. The study lasted for 61 months, but the status of one particular patient, even at 60th month, was modelled using the covariate value collected at study entry. As such, the connection between covariates and the responsible variable is in doubt more or less. Practitioners need a modeling procedures that can establish closer connection between covariates and the responsible variable.

Regression Analysis of Arbitrarily Interval-censored Data with External Time-dependent Covariates

From the literature review above, it is easy to see the advantages of collecting arbitrarily interval-censored survival data and using external time-dependent covariates instead of time-independent covariates from both theoretical and applied perspectives. In particular, compared to arbitrarily interval-censored survival data, right-censored survival data cannot provide a more informatively specified event time, as the status of an event might have changed well before the last examination. Further, external time-dependent covariates allow updating the hazard of an event for a subject according to the evolution of such covariates along the follow-up, thus providing a more informative description of the hazard of occurrence of an event of interest. Therefore, it is natural to deem regression analysis of arbitrarily interval-censored survival data with external time-dependent covariates as a powerful analytical tool for describing the survival experience of subjects.

However, in practice, modeling with arbitrarily interval-censored data is often mimicked by methods developed for right-censored data for the sake of simplicity. For this, the interval needs to be replaced by an imputed time. For example, in one such method, mid-point imputation, the analysis is performed as though the mid-point of each subject's interval were the exact event time (Law & Brookmeyer, 1992). For example, a cohort of subjects was initially uninfected and at risk of infection in the interval month one to month nine. Screening tests for evidence of infection occurred periodically over the interval, and subjects were followed for onset of AIDS. Mid-point imputation refers to imputing the date of infection by the mid-point of the interval which is the average of month one and month nine. Then the resulting imputed time is used as the infection time. Applying methods for right-censored data on the artificial fixed points can lead to biased and misleading results, such as biased estimation and underestimation of the true error variance (Odell, Anderson, & D'Agostino, 1992; Rücker & Messerer, 1988), and biased hazards and hazard ratios (Dorey, Little, & Schenker, 1993; Law & Brookmeyer, 1992). On the other hand, as for external time-dependent covariates, which are often essential predictors for the hazard, they are either disregarded or substituted for by the baseline values of time-independent covariates for the purpose of simplifying the corresponding analysis. Further, most of the inferential procedures developed for interval-censored data only apply to time-independent covariates (Sun, 2006). Although some exceptions exist, they are either for a model not based on the Cox PH model, such as the additive hazards regression model (Lin, Oakes, & Ying, 1998; Martinussen & Scheike, 2002), or for data other than arbitrarily interval-censored data, such as current status data (Martinussen & Scheike, 2002; Van Der Laan & Robins, 1998).

Review of Literature on Properties of Parameter Estimates from the Extended Cox Model

To answer the three hypotheses in the current research, previous literature on properties of parameter estimates, such as absolute relative bias (ARB) of parameter estimates, that is, the absolute value of the difference between parameter estimates and true values of the coefficients divided by of the coefficients, percent of correct sign of parameter estimates (% CS), power, and type I error rate, from the extended Cox model and Farrington's model was reviewed. However, regarding properties of parameter estimates from Farrington's model, no one has yet conducted such research. In most research on regression analysis of interval-censored data (Ma & Kosorok, 2005; Muggeo, Attanasio, & Porcu, 2010), Farrington's model was only introduced as one way of modeling interval-censored data. Even in Farrington's article (1996) where Farrington proposed the model, he did not conduct a simulation study on properties of parameter estimates, either.

Regarding properties of parameter estimates from the extended Cox model, bias can be as low as .001(Hendry, 2014; Xiao, Abrahamowicz, & Moodie, 2010). Power can be as high as .906 (Chen, Ibrahim, & Chu, 2011). Type I error rate, however, is inflated (Abrahamowicz, Mackenzie, & Esdaile, 1996). No one has yet conducted research on percent of correct sign of parameter estimates from the extended Cox model.

Chapter Summary

In summary, while methods of modeling right-censored data either with time-independent or time-dependent covariates, and methods of modeling arbitrarily interval-censored data with time-independent covariates, are available in the literature, there is no estimation method of modeling arbitrarily interval-censored data and external time-dependent covariates simultaneously yet. Moreover, in reality, practitioners need the results from the new modeling procedure that could help them diagnose the status of a particular event of a subject more realistically. Regarding properties of parameter estimates, only the extended Cox model was investigated in previous literature. Thus, a new method, which is based on the Cox PH model and which extends Farrington’s approach, was proposed and evaluated in the current study. The corresponding parameter estimation and inferential procedures were explored as well.

CHAPTER III

METHODOLOGY

The literature review presented in Chapter II has revealed that it is reasonable to incorporate arbitrarily interval-censored data and external time-dependent covariates into regression analysis of survival data. However, two challenges arising from the corresponding modeling procedure ensue due to such an inclusion.

The first challenge is how external time-dependent covariates are handled under the framework of Farrington's (1996) modeling procedure, as it employs time-independent covariates alone. In particular, in the case of confined data, although one subject might have undergone more than two examinations after study entry, which form more than two intervals, only two intervals are used under Farrington's approach, with one from study entry to *A*, and the other from *A* to *B*, where *A* denotes the left endpoint and *B* denotes the right endpoint of the censoring interval. As such, although values of covariates can be collected at each examination, Farrington's approach does not have the mechanism to handle varying covariate values. Thus, when external time-dependent covariates are employed instead, Farrington's modeling procedure has to be extended to such covariates.

The second challenge is the resulting inferential procedure from such an extension. As Farrington's approach has not been extended to external time-dependent covariates, how to infer from regression analysis of such a data situation, including estimation methods and hypothesis testing, has not been identified in the literature. The purpose of estimation is to investigate how the survival experience of a group of subjects depends on the values of one or more independent variables. The purpose of hypothesis testing is to test whether the null hypothesis that one or more coefficients is equal to zero is rejected or not. As such, an attempt was made to fill in this gap, which is the main purpose of the current research. In particular, I first proposed for the current study the non-likelihood-based estimation method, generalized estimating equations (GEE; Liang & Zeger, 1986; Zeger & Liang, 1986), and then conducted hypothesis testing and one simulation study to explore the properties of parameter estimates.

In addition, there are three main research questions for the proposed study.

First, how does ARB and percent of correct sign of parameter estimates from the proposed approach compare to those from Farrington's approach, and those from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates? Second, how does the power from the proposed approach compare to that from Farrington's approach and that from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates? Third, how does type I error rate from the proposed approach compare to that from Farrington's approach and that from the extended Cox model, as applied to arbitrarily interval-censored survival data with external time-dependent covariates?

For the first research question, compared to Farrington's model and the extended Cox model, lower ARB from parameter estimates for the external time-dependent covariates was expected from the proposed approach. For the second research question, compared to Farrington's model and the extended Cox model, greater power related to the external time-dependent covariate was expected from the proposed model. For the third research question, type I error rate related to external time-dependent covariates was expected to be close to the nominal level of .05 for the proposed model, but higher for Farrington's model and the extended Cox model.

Statistical Inference for the Extended Generalized Non-linear Model

As described in Chapter II, Farrington's approach is capable of converting the problem of regression analysis of arbitrarily interval-censored data to a binary response regression analysis. The resulting model is a logistic model with correlated binary responses and time-independent covariates. Since it is not possible to express a simple function of the probability of success as a linear combination of the unknown parameters, except in the case where there are no independent variables in the model, the resulting model is actually a generalized non-linear model (Collett, 2003) with a binary response. Various generalized linear models are in fact special cases of generalized non-linear models.

When external time-dependent covariates, such as repeated measurements on the outdoor levels of air pollutants, are further incorporated in this binary response generalized non-linear model, the model becomes an extended generalized non-linear model (EGNM). Two difficulties regarding the corresponding inferential procedure arise. The first difficulty is how to formulate an expression which serves as the basis for parameter estimation and the corresponding hypothesis testing. The key point is the formulated expression must reflect that probabilities depend on time via external time-dependent covariates, which has not been explored under Farrington's modeling framework. The second difficulty is how to estimate parameters for external time-dependent covariates in the EGNM, which has not been explored, either. The first difficulty is discussed in the section, "Estimation Using GEE," below. The second difficulty is discussed first.

Estimation Methods for the Binary Response Generalized Non-linear Model

In terms of estimation methods for the binary response generalized non-linear model that handles time-independent covariates, usually two classes of methods can be used. Depending on whether the response variable assumes a particular probability distribution, those methods can be classified into either a likelihood-based or a non-likelihood-based method.

When a likelihood-based method is applied to the estimation procedure, the joint probability distribution of the response variable is constructed first. The resulting joint likelihood function is then evaluated using numerical methods.

Non-likelihood-based methods, such as generalized estimating equations (GEE; Liang & Zeger, 1986; Zeger & Liang, 1986), can also be used in estimating parameters for the binary response generalized non-linear model. These methods avoid constructing a likelihood function as the basis for estimation. In particular, in setting up GEE, assuming the distribution of the response variable is from an exponential family (McCullagh & Nelder, 1989), all that is needed is specification of a mean model and the mean-variance relationship in the response variable, and a working correlation structure, that is, the pairwise within-subjects association among the responses. Estimation may be accomplished either via generalized weighted least-squares or through an iterative process (Zorn, 2001).

The Data Structure

Before discussing estimation methods, the data structure for arbitrarily interval-censored data with external time-dependent covariates is described. Regarding the response, consider a survival study that gives rise to arbitrarily interval-censored data,

$$\{(A_{i(t-1)}, B_{it}], \boldsymbol{X}_i; i = 1, \dots, N; t = 1, \dots, T_i\}, \tag{44}$$

for the event times of interest. In Equation 44, $(A_{i(t\text{-}1)}, B_{it}]$ denotes an interval formed by the $(t$ - $1)$th and the tth examinations for the ith subject, $t = 1$,…, T_i denotes the number of examinations after study entry, $(t$ - $1) = 0$ denotes study entry, $A_{i(t\text{-}1)}$ and B_{it} denote the left endpoint and the right endpoint for the interval, respectively, and N denotes the number of subjects. Within each interval of this sequence, the event of interest for that subject is observed either to occur or not to occur. Let $\boldsymbol{y}_i = [y_{i1},\dots, y_{iTi}]'$ be a T_i by one vector of binary responses corresponding to the formed intervals for the ith subject, where $y_{it} = 1$ denotes the occurrence of the event and $y_{it} = 0$ otherwise.

Regarding external time-dependent covariates, the associated design matrix for the ith subject, $\boldsymbol{X}_i$, in Equation 44 takes the form

$$\boldsymbol{X}_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1P} \\ \vdots & \ddots & \vdots \\ x_{iT_i1} & \cdots & x_{iT_iP} \end{bmatrix},$$

where $p = 1,\dots, P$ denotes different external time-dependent covariates. For the ith subject at the tth examination, the row vector $\boldsymbol{x}_{it} = [x_{it1},\dots, x_{itP}]$ gives the P covariate values, and for the pth covariate for the ith subject the column vector $\boldsymbol{x}_{i.p} = [x_{i1p},\ldots, x_{Tip}]$ gives values for that covariate across all T_i examinations. Note that values collected at study entry are not included in the matrix, as they are not used in the modeling procedure, which is discussed below. For simplicity, the current research only used one external time-dependent covariate.

Taken together, the full response vector for all N subjects is given by the column vector $\boldsymbol{y} = [\boldsymbol{y}_1,\ldots, \boldsymbol{y}_N]'$, and the full design matrix is similarly given by $\boldsymbol{X} = [\boldsymbol{X}_1',\ldots, \boldsymbol{X}_N']'$.

There are a few more assumptions made for the data situation in the current research. In particular, examination times differ across N subjects, who might have different numbers of examinations and hence different numbers of responses. The number of examinations at which the ith subject is observed is smaller relative to N, that is, $T_i < N$. Further, the between-subjects responses are assumed independent. Moreover, all intervals formed by consecutive examinations for a particular subject can be described using a sequence of $(A_{i(t\text{-}1)}, B_{it}]$.

As described in Chapter II, arbitrarily interval-censored data in fact entail three types: left-censored data with $A_i = 0$ and $B_i \neq \infty$, right-censored data with $A_i \neq 0$ and $B_i = \infty$, and confined data with $A_i \neq 0$ and $B_i \neq \infty$. For left-censored data, there is only one examination after study entry, and thus there is only one response with $y_i = 1$. For right-censored data, there is at least one examination after study entry, and thus there is at least one response with $y_i = 0$. For confined data, there are at least two examinations after study entry, and thus there is at least one response with $y_i = 0$ and only one response with $y_i = 1$, with the responses correlated. For simplicity, the current research only considers confined data.

Choosing an Estimation Method for the EGNM

When it comes to estimation methods for the EGNM that handles external time-dependent covariates, neither a likelihood-based method nor a non-likelihood-based method has been identified in the literature. While both classes of methods have the potential for incorporating this type of covariates, a non-likelihood-based estimation method, GEE, was chosen in the proposed study for the following reasons.

First, a likelihood-based method could be computationally very burdensome. Two types of problems tend to occur. In some cases, the constructed likelihood function is extremely difficult to evaluate numerically with available computer technology. In other cases, the likelihood function must be maximized subject to a set of nonlinear constraints implied by the model, which further adds to the computational burden. Moreover, the successful implementation of a likelihood-based method depends greatly on good starting values (Vonesh & Chinchilli, 1996), which involves specifying initial estimates of the coefficients of the independent variables. If starting values are far from their optimal estimated values, then the corresponding optimization method may fail to converge.

Second, the response probability under Farrington's approach, p_i, in Equation 40 is actually identified through a generalized linear model using a complementary log-log link. Now that the distribution of the response variable is from an exponential family, i.e., the binomial distribution, and the mean model and the mean-variance relationship are readily specified, GEE is a natural candidate for the estimation method.

Third, besides its computational simplicity compared with likelihood-based estimation methods, the GEE approach produces consistent parameter estimates even with misspecification of the working correlation structure (Zeger & Liang, 1986).

Although the estimates are not optimal compared to those obtained from likelihood-based estimation methods, a trade-off is attained between computation and statistical properties. In the current research, priority was given to computation. Therefore, the non-likelihood-based estimation method GEE was applied in the current research.

Estimation Using GEE

Formally, the response probability of the EGNM, with one external time-dependent covariate *X*, takes the form

$$p_{it} = 1 - e^{\left\{\left[-e^{(\beta_0+\beta_1 X_{it})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}, \tag{45}$$

where β_0 denotes the parameter for the intercept constant to be estimated, β_1 denotes the parameter for the covariate to be estimated, p_{it} denotes the response probability and X_{it} denotes the external time-dependent covariate value at the *t*th examination for the *i*th subject, and $\theta_g d_{ig}$ is from Equation 40. Although this model in Equation 45 is only partly linearized using the complementary log-log link, that is, a weakly parametric generalized linear modeling framework (Farrington, 1996), GEE, designed to model correlated data under generalized linear models, can still be used.

Decomposition of GEE. In using GEE, a mean model and the mean-variance relationship of the response variable, which is from the exponential family of distributions, and a working correlation structure representing the correlation believed to be present among responses within subjects must be specified.

In particular, according to multivariate statistics theory, a variance-covariance matrix of data is expressed as

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i \mathbf{A}_i^{1/2}, \tag{46}$$

where $\mathbf{V}_i$ is a matrix representing the marginal response variance-covariance for the *i*th subject, $\mathbf{A}_i$ is a diagonal matrix representing the response variance under the assumption of independence, $\mathbf{R}_i$ is the working correlation for the response, and ϕ is the overdispersion factor. Hence, the generalized estimating equations for the mean parameters $\boldsymbol{\beta}$ for N independent subjects take the form

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' [\mathbf{V}_i(\boldsymbol{\alpha})]^{-1}[\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0}, \qquad (47)$$

where $\boldsymbol{\mu}_i = [\mu_{i1},\ldots, \mu_{iTi}]'$ and $\mathbf{V}_i$ denotes the mean and the variance-covariance, respectively, of the response $\boldsymbol{y}_i = [y_{i1},\ldots, y_{iTi}]'$ for the *i*th subject, and $\boldsymbol{\alpha}$ denotes a *s* by one vector of correlation parameters that fully describes the working correlation structure. Solving these estimating equations provides parameter estimates $\hat{\boldsymbol{\beta}}$. Each coefficient in $\boldsymbol{\beta}$ can be interpreted similarly to that of the standard regression model, with the added condition that the autocorrelation has been accounted for (Zeger & Liang, 1992).

Application of GEE to the current research. When it comes to the current research, the estimation method for the EGNM fits with the GEE scenario. When external time-dependent covariates are included, the response probability under Farrington's approach in Equation 40 will be modified to reflect the resulting dynamic relationship between such type of covariates and the response variable. As such, the mean vector for the *i*th subject takes the form

$$\boldsymbol{\mu}_i = \boldsymbol{p}_i = \begin{bmatrix} 1 - p_{i1} \\ 1 - p_{i2} \\ \vdots \\ 1 - p_{i(T_i-1)} \\ p_{iT_i} \end{bmatrix},$$

where p_{it} depends on time via the external time-dependent covariate, and the corresponding response variance is represented by the diagonal elements of $\mathbf{A}_i$, that is,

$$
\begin{aligned}
\operatorname{diag}(\mathbf{A}_i) &= \operatorname{diag}[\boldsymbol{\mu}_i(\boldsymbol{I}_i - \boldsymbol{\mu}_i)] \\
&= \operatorname{diag}[\boldsymbol{p}_i(\boldsymbol{I}_i - \boldsymbol{p}_i)] \\
&= \operatorname{diag}\begin{bmatrix} (1 - p_{i1})p_{i1} \\ (1 - p_{i2})p_{i2} \\ \vdots \\ [1 - p_{i(T_i-1)}][p_{i(T_i-1)}] \\ p_{iT_i}(1 - p_{iT_i}) \end{bmatrix},
\end{aligned}
$$

where $\boldsymbol{I}_i$ is an identity matrix.

Constructing GEE for the current research. As mentioned previously, one of the difficulties regarding the corresponding inferential procedure for the current data situation is how to formulate an expression which serves as the basis for inference and reflects that probabilities depend on time. When GEE is used as the estimation method, the response probability vector, i.e., the mean vector, which is based on Farrington's response probability in Equation 40, serves as the basis for inference and is constructed to reflect that each component of the probability vector depends on time. Moreover, an appropriate working correlation structure for the response variable is chosen to account for the inclusion of external time-dependent covariates.

Constructing the expression for probabilities. Recall from Farrington's approach that the basis of the likelihood function is the response probability

$$
p_i = 1 - e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}},
$$

which does not depend on time due to time-independent covariates. When external time-dependent covariates replace time-independent covariates, the number of intervals to be used is greater than or equal to two to allow the collection of varying values of covariates. The resulting response probabilities based on Farrington's approach become more complicated. The reason is except for the response probability during the first interval, all other responses are conditional on their precedents via varying values of covariates. As an example, the response probability for the fifth interval is conditional on all four response probabilities prior to it. It is evident that as time goes on each component of a mean vector has to account for more terms.

To see how to construct a particular response probability, suppose for the *i*th subject, values of one single external time-dependent covariate collected at each examination are denoted as x_{i0}, x_{i1}, x_{i2}, x_{i3},…, x_{iT_i}, respectively, where x_{i0} is the value collected at study entry. The symbol W_{it}^{1} is used to denote an occurrence case, that is, y_i = 1, within the *t*th interval for the *i*th subject, and W_{it}^{0} to denote a non-occurrence case, that is, y_i = 0, within the *t*th interval for the *i*th subject. The event of interest is observed to occur between (T_i -1)th and T_ith examination, where T_i is the number of examinations. The first response probability corresponding to the interval (τ_{i0}, τ_{i1}], where τ_{it} denotes an end time at the *t*th examination for an interval and τ_{i0} = 0 denotes study entry, takes the form

$$\begin{aligned}\mathrm{P}(W_{i1}^{0}) &= 1-[S_i(\tau_{i0})-S_i(\tau_{i1})]\\ &= 1-[1-S_i(\tau_{i1})]\\ &= S_i(\tau_{i1})\\ &= e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}},\end{aligned} \tag{48}$$

where d_{ig} = 1 if the ordered censoring time $t_{(g)}$ $\leq \tau_{i1}$, the right endpoint of the first interval, and d_{ig} = 0 otherwise. Equation 48 is also the mean of the response variable corresponding to the first interval. In Equation 48, the covariate value collected at τ_{i0} is ignored, which is discussed in the section, “Simulation Design,” below. Note that this mean is an unconditional mean.

From the second interval, the response probability corresponding to the interval $(\tau_{i1}, \tau_{i2}]$ becomes conditional on its precedent, $\mathrm{P}(W_{i1}^{0})$, and takes the form

$$\begin{aligned}
\mathrm{P}(W_{i2}^{0}|W_{i1}^{0}) &= \frac{\mathrm{P}(W_{i1}^{0} \cap W_{i2}^{0})}{\mathrm{P}\left(W_{i1}^{0}\right)} \\
&= \frac{\mathrm{P}(W_{i1}^{0})\left[1 - \frac{S_i(\tau_{i1}) - S_i(\tau_{i2})}{\mathrm{P}\left(W_{i1}^{0}\right)}\right]}{\mathrm{P}\left(W_{i1}^{0}\right)} \\
&= 1 - \frac{S_i(\tau_{i1}) - S_i(\tau_{i2})}{S_i(\tau_{i1})} \\
&= \frac{S_i(\tau_{i2})}{S_i(\tau_{i1})} \\
&= \frac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}},
\end{aligned} \tag{49}$$

where $d_{1ig} = 1$ if the ordered censoring time $t_{(g)} \leq \tau_{i2}$, the right endpoint of the second interval, and $d_{1ig} = 0$ otherwise, and $d_{2ig} = 1$ if $t_{(g)} \leq \tau_{i1}$, the left endpoint of the second interval and $d_{2ig} = 0$ otherwise. Equation 49 is also the mean of the response variable corresponding to the second interval.

In the same vein, the third response probability corresponding to the interval $(\tau_{i2}, \tau_{i3}]$ is conditional on its precedents, $\mathrm{P}(W_{i1}^{0})$ and $\mathrm{P}(W_{i2}^{0})$, and takes the form

$$\begin{aligned}
\mathrm{P}(W_{i3}^{0}|W_{i1}^{0}\cap W_{i2}^{0}) &= \frac{\mathrm{P}(W_{i1}^{0}\cap W_{i2}^{0}\cap W_{i3}^{0})}{\mathrm{P}\left(W_{i1}^{0}\right)\mathrm{P}\left(W_{i2}^{0}|W_{i1}^{0}\right)}\\
&= \frac{\mathrm{P}(W_{i1}^{0})\mathrm{P}(W_{i2}^{0}|W_{i1}^{0})\left[1-\frac{S_i(\tau_{i2})-S_i(\tau_{i3})}{\mathrm{P}(W_{i1}^{0})\mathrm{P}(W_{i2}^{0}|W_{i1}^{0})}\right]}{\mathrm{P}\left(W_{i1}^{0}\right)\mathrm{P}\left(W_{i2}^{0}|W_{i1}^{0}\right)}\\
&= 1-\frac{S_i(\tau_{i2})-S_i(\tau_{i3})}{S_i(\tau_{i1})\frac{S_i(\tau_{i2})}{S_i(\tau_{i1})}}\\
&= \frac{S_i(\tau_{i3})}{S_i(\tau_{i2})}\\
&= \frac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}},
\end{aligned}\tag{50}$$

where $d_{1ig} = 1$ if the ordered censoring time $t_{(g)} \leq \tau_{i3}$, the right endpoint of the third interval, and $d_{1ig} = 0$ otherwise, and the value $d_{2ig} = 1$ if $t_{(g)} \leq \tau_{i2}$, the left endpoint of the third interval and $d_{2ig} = 0$ otherwise. Equation 50 is also the mean of the response variable corresponding to the third interval.

Suppose that the event of interest occurred in the fourth interval $(\tau_{i3}, \tau_{i4}]$. The fourth response probability corresponding to this interval is conditional on its precedents, $\mathrm{P}(W_{i1}^{0})$, $\mathrm{P}(W_{i2}^{0})$ and $\mathrm{P}(W_{i3}^{0})$, that is, three consecutive non-occurrence cases, and takes the form

$$\begin{aligned}
\mathrm{P}(W_{i4}^{1}|W_{i1}^{0}\cap W_{i2}^{0}\cap W_{i3}^{0}) &= \frac{\mathrm{P}(W_{i1}^{0}\cap W_{i2}^{0}\cap W_{i3}^{0}\cap W_{i4}^{1})}{P}\\
&= \frac{P\left[\frac{S_i(\tau_{i3})-S_i(\tau_{i4})}{P}\right]}{P}\\
&= \frac{S_i(\tau_{i3})-S_i(\tau_{i4})}{S_i(\tau_{i1})\frac{S_i(\tau_{i2})}{S_i(\tau_{i1})}\frac{S_i(\tau_{i3})}{S_i(\tau_{i2})}}\\
&= 1-\frac{S_i(\tau_{i4})}{S_i(\tau_{i3})}\\
&= 1-\frac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i4})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}},
\end{aligned}\tag{51}$$

where $P = \mathrm{P}(W_{i1}^{0})\mathrm{P}(W_{i2}^{0}|W_{i1}^{0})\mathrm{P}(W_{i3}^{0}|W_{i1}^{0} \cap W_{i2}^{0})$, $d_{1ig} = 1$ if the ordered censoring time $t_{(g)} \leq \tau_{i4}$, the right endpoint of the fourth interval, and $d_{1ig} = 0$ otherwise, and $d_{2ig} = 1$ if $t_{(g)} \leq \tau_{i3}$, the left endpoint of the fourth interval and $d_{2ig} = 0$ otherwise. Equation 51 is also the mean of the response variable corresponding to the fourth interval.

Thus, each response probability corresponding to that interval, or each component of the mean vector is constructed for the *i*th subject. Similarly, mean vectors for all other subjects can be established. These mean vectors are substituted into Equation 47 to obtain parameter estimates.

More generally, the mean vector for the *i*th subject in the case of confined data takes the form

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \begin{bmatrix} e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}} \\ \dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ \dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ \vdots \\ \dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-2)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ 1 - \dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{iT_i}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \end{bmatrix}, \tag{52}$$

where x_{iTi} is the covariate value collected at the last examination. The corresponding response variance under the assumption of independence is represented by the diagonal elements of $\mathbf{A}_i$, that is, $\mathrm{diag}(\mathbf{A}_i) =$

$$\begin{bmatrix} e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}\left(1-e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}}\right) \\ \dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\left(1-\dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\right) \\ \dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\left(1-\dfrac{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\beta_0+\beta_1 x_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\right) \\ \vdots \\ \dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-2)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\left(1-\dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-2)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\right) \\ \left(1-\dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{iT_i}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}}\right)\dfrac{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{iT_i}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\beta_0+\beta_1 x_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \end{bmatrix}. \qquad (53)$$

Choosing an appropriate working correlation structure for the response variable. To reflect the correlation present within the cluster of responses for the *i*th subject, a working correlation matrix, $\mathbf{R}_i$, permitting dependence, such as the compound symmetry structure, is normally selected. Therefore, an identity matrix ***I*** which treats clustered responses as independent may be inappropriate to represent the true relationship among the responses. However, with external time-dependent covariates replacing time-independent covariates, Hu, Goldberg, Hedeker, Flay, and Pentz (1998) and Pepe and Anderson (1994) have pointed out that the consistency of parameter estimates using GEE is not assured with arbitrary working correlation structures unless a subject's repeated measurements are independent, i.e., the independent working correlation is satisfied. Pepe and Anderson (1994) thus recommended the use of the independent working correlation as a safe choice of analysis. Hence, in the current research, an identity matrix,

$\mathbf{R}_i$ ($\boldsymbol{\alpha}$) = $\boldsymbol{I}$, with the number of its elements equal to the number of responses from the *i*th subject, was used to construct the score functions $U(\boldsymbol{\beta})$ in Equation 47.

Thus, GEE in Equation 47 is fully specified, that is,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' [\mathbf{V}_i(\boldsymbol{\alpha})]^{-1}[\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0},$$

where $\boldsymbol{y}_i$ denotes the response vector, $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ was expressed in Equation 52, and

$$\mathbf{V}_i(\boldsymbol{\alpha}) = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2},$$

where the diagonal elements of $\mathbf{A}_i$ were expressed in Equation 53. Solving these estimating equations provides the parameter estimates.

Investigating Properties of the Parameter Estimates

Hypothesis Testing

Typically, the first step following the fit of a regression model is the assessment of the significance of the estimated parameters, that is, hypothesis testing. Because the estimation method GEE does not have a likelihood function, likelihood-ratio methods are not available for conducting inference tests about the estimated parameters. Instead inference uses either the Wald test or generalized score tests (Boos, 1992). Both tests are based on the asymptotic normality of the estimators together with the empirically based standard errors. As the Wald test is reliable mainly for very large samples, generalized score tests are preferable to the Wald test (Agresti, 2007). Consequently, generalized score tests were employed. This test statistic, for a vector of responses, takes the following form

$$Q(\boldsymbol{\mu}; \boldsymbol{y}) = \sum_{i=1}^{N} Q_i(\boldsymbol{\mu}_i; \boldsymbol{y}_i),$$

where

$$Q_i(\boldsymbol{\mu}_i; \boldsymbol{y}_i) = \int_{\boldsymbol{y}_i}^{\boldsymbol{\mu}_i} (\boldsymbol{y}_i - \boldsymbol{t})'[\phi \boldsymbol{V}_i(\boldsymbol{t})]^{-1} d\boldsymbol{t},$$

where $\boldsymbol{\mu}_i$ is the mean vector, and $\boldsymbol{y}_i$ is the response vector for the *i*th subject.

Simulation Design

Other properties, such as power of hypothesis testing, type I error rate, ARB, and standard errors of the parameter estimates, were further investigated via a series of simulation studies. The significance level used was .05. The power used was .90, which is considered adequate power (Lachin, 2013; Loewy, 2015). However, the focus of this study was on simple descriptive comparisons of performance among the three methods. In conducting the simulation study for the current research, the data needed for the simulation study were generated first. Then results from the simulation study were saved and reported.

Typical simulation design conditions for GEE. As in the current research the GEE approach is employed to obtain parameter estimates, typical simulation design conditions for GEE are discussed first. The purpose was to obtain the most commonly seen simulation design conditions for GEE, such as sample sizes, numbers of replicates, examination times and responses, true parameter values, and types of distributions of the independent variables and working correlation structures in order to help design the simulation design conditions for the current research. Depending on the purpose of a particular simulation study where the GEE approach is employed, some simulation studies contain all the simulation design conditions, while other simulation studies do not. For example, Westgate (2014) conducted a simulation study to compare the quadratic inference function (QIF) approach to GEE for the marginal analysis of correlated data, where examination times were not employed in the design conditions. From prior empirical evidence, sample sizes have ranged from 20 to 4,077, with a majority ranging from 50 to 400 (Chen & Zhou, 2013; Touloumis, Agresti, & Kateri, 2013; Westgate, 2014). The numbers of replicates have ranged from 100 to 10,000, with a majority ranging from 500 to 1,000 (Wang, Lee, Zhu, Redline, & Lin, 2013; Westgate & Braun, 2013). The numbers of examination times have ranged from 3 to 10, with a majority ranging from 3 to 5 (Chen & Zhou, 2012; Shen & Chen, 2012). True parameter values for risk factors have ranged from - 4.5 to 5, with a majority ranging from -1 to 1(Mehrotra, Li, Liu, & Lu, 2012; Zhang & Paul, 2013). Distributions of the independent variables have included the binominal (Shen & Chen, 2012), uniform (Zhang & Paul, 2013), and normal (Westgate & Braun, 2013) distributions. Working correlation structures have included exchangeable (Chen & Zhou, 2013), autoregressive (Westgate & Braun, 2013), unstructured (Zhang & Paul, 2013), and independent (Paul & Zhang, 2014) structures.

Typical simulation design conditions for survival analysis. As in the current research arbitrarily interval-censored survival data are modeled, typical simulation design conditions for survival analysis are discussed as well. Typical design conditions for regression analysis of survival data include sample sizes, the number of replicates and examination times, the censoring rate, true parameter values, distributions of the independent variables and survival times, and the length of the follow-up. Depending the purpose of a particular simulation study, some simulation studies contain all the design conditions, while other simulation studies do not. For example, He and Schaubel (2014) conducted a simulation study to evaluate the finite-sample properties of the proposed estimators, where the censoring rate was not employed in the design conditions. From prior empirical evidence, sample sizes have ranged from 15 to 10,000, with a majority ranging from 50 to 600 (Combescure, Foucher, & Jackson, 2014; He & Schaubel, 2014; Wynant & Abrahamowicz, 2014). The numbers of replications have ranged from 50 to 100,000, with a majority ranging from 100 to 1,000 (Bhatt & Tiwari, 2014; Pan, Bao, Dai, & Fang, 2014; Salim, Ma, Fall, Andrén, & Reilly, 2014). Censoring rates have ranged from 5% to 55% (Carlin & Solid, 2014; Wallace, 2014). The numbers of examinations have been less than or equal to 7, and have ranged between 2 and 7 (He & Schaubel, 2014; Shen, Anderson, Sinha, & Li, 2014). True parameter values for risk factors have ranged from - 4 to 9, with a majority ranging from -1 to 1 (Crowther, Look, & Riley, 2014; Schaubel, Zhang, Kalbfleisch, & Shu, 2014; Whitehead, 2014). Distributions of the independent variables have included the Bernoulli (He & Schaubel, 2014), logistic (He & Schaubel, 2014), and normal (Carlin & Solid, 2014) distributions. Distributions of survival times have included the exponential (Whitehead, 2014), gamma and log normal (Bhatt & Tiwari, 2014), and Weibull (Crowther, Look, & Riley, 2014) distributions. The lengths of the follow-up have ranged from 60-240 days (Lyman, Reiner, Morrow, & Crawford, 2015) to 18 years (Molyneux, Birks, Clarke, Sneade, & Kerr, 2015).

There are three studies that examined interval-censored data and time-dependent covariates. In Van Der Laan and Robins' study (1998), which investigated current status data, a special type of interval-censored data, the sample size was 500, the number of replicates was 625, and $\beta = 2$. The independent variables assumed the binomial and normal distributions. In Martinussen and Scheike's study (2002), which also investigated current status data, the sample sizes were 100 and 200, the number of replicates was 10,000, the numbers of examinations were 4 and 6, and $\beta = .5$. The independent variable assumed the uniform distribution. In Lin, Oakes, and Ying's article (1998), which investigated the additive hazards regression model, the sample sizes were 100 and 200, the number of replicates was 10,000, and $\beta = .5$. The independent variable assumed the uniform distribution.

The numbers of subjects, sample sizes, and the number of replications for the simulation study. As the current research involves both the GEE approach and arbitrarily interval-censored survival data, the corresponding simulation design conditions drew upon prior empirical evidence from the research using the GEE approach, survival analysis, and the research related to the regression analysis of interval-censored survival data.

In the current research, the term "the number of subjects" rather than "the sample size" was used to refer to the number of a cohort of subjects enrolled in a follow-up study for the following reason. Usually, the sample size of a data set refers to the number of subjects enrolled in a study, and information collected from one subject comprises one single row in the data set. However, in the current research, information collected from one subject for the three models was augmented to multiple rows. Thus, the number of rows in the data set did not match the number of subjects enrolled. In order to avoid the confusion, the term "the number of subjects" was used to refer to the number of a cohort of subjects, and “the sample size” was used to refer to the number of rows in the resulting augmented data set.

Recall from the above literature review, the optimal simulation design for this simulation study was found to be 50, 250, and 600 representing small, medium, and large numbers of subjects, respectively, and 1,000 representing the number of replications. However, the computers used for the simulation study do not have enough RAM (Random Access Memory) installed, which was found through trial and error. In particular, each time 600 subjects and 1,000 replications were used, the software package R (Version 3.2.2) stopped working and the computer gave the following warning message:

R for Windows GUI front-end has stopped working.

Through trial and error, it was found that when the number of replication was 150, the maximum possible number of subjects was 1,000; when the number of subjects was 350, the maximum possible number of replication was 500. As such, in order to show how the simulation results behaved as the number of subjects increased in the simulation study, two sets of simulation results were presented: in the first set, the number of replication was 150, and the numbers of subjects used were 50, 250, 500, and 1,000; in the second set, the number of replication was 500, and the numbers of subjects used were 50, 150, 250, and 350.

As the mean number of examinations was around 2.1, and the resulting mean number of rows in the augmented data set for each subject was 2.1 + 2 = 4.1, when the numbers of subjects used were 50, 250, 500, and 1,000, with one to six examinations, a total sample size of between 200 and 4,000 or so was obtained; when the numbers of subjects used were 50, 150, 250, and 350, with one to six examinations, a total sample size of between 200 and 700 or so was obtained.

The Data Generation Process

Steps in the Data Generating Process

The procedure for generating data for the simulation study was as follows. In summary, the first step was to simulate data for fitting the EGNM, which for each subject included 100 values each for the two external time-dependent covariates X_{1t} and X_{2t}, the status of an event of interest (whether an event of interest has occurred), an event time, a corresponding censoring interval where an event of interest was assumed to have occurred, and the number of follow-ups between study entry and the left endpoint of the censoring interval. The reason the number 100 was chosen is two-fold. The first is as no previous simulation studies have been conducted on the number of external time-dependent covariate values used to simulate event times, the number 100 was chosen arbitrarily. The second is 100 values, without replacement, are enough to be assigned to each of the simulated examinations, the number of which for all subjects is less than 100. Then, the simulated data were modified for fitting the extended Cox model and Farrington's model.

Simulating data for fitting the EGNM. The procedure was as follows. Values of the significant external time-dependent covariate were generated first, as the corresponding process x_{it}, $x_{i(t+1)}$,…, x_{iT_i}, is not affected by the response $y_{i(t-1)}$ at the (t - 1)th examination, conditional on $x_{i(t-1)}$, that is, it rules out feedback from the response process to the covariate process (Lai & Small, 2007), or the covariate process does not depend on the response process. T_i denotes the number of examinations, (t - 1) = 0 denotes study entry, and x_{it} denotes the covariate value collected at the tth examination. Then, based on the simulated covariate process, the response process and the corresponding arbitrarily interval-censored data were generated.

Simulating values for the external time-dependent covariates. The first step of the data generation process was to simulate external time-dependent covariate values. There are two statements made about generating the covariate process. The first statement is how many covariate values should be simulated for each subject, as numbers of examinations vary across the cohort. Since covariate values are collected at examinations, the number of simulated covariate values was based upon the number of simulated examinations. The second statement is values of external time-dependent covariates were assumed piecewise constant for the model, as was suggested by Farrington (1996), that is, they remain constant between two consecutive examinations.

The significant external time-dependent covariate for the model, which was used to simulate event times, was denoted as X_{1t}. Regarding the distribution X_{1t} could assume, prior empirical evidence showed that independent variables assumed various distributions. As continuous time-dependent covariates were investigated in the current research, the normal distribution was chosen for X_{1t}. By definition, external time-dependent covariates do not depend on a subject's survival. Thus, values of X_{1t} to be simulated were based on national nitrogen dioxide concentrations (United States Environmental Protection Agency, 2013), assuming a normal distribution, $X_{1t} \sim N\,(79, 484)$, where 79 is the mean, and 484 is the variance of the national nitrogen dioxide concentrations from 1980 to 2012. For the ith subject, 100 values of X_{1t} were generated, and then the average of the 100 values was used to generate the corresponding event time. The other external time-dependent covariate X_{2t}, also assuming the normal distribution, though deemed as one potential factor influencing the hazard of experiencing the respiratory disease, bears no relation to simulating event times. The purpose of including X_{2t} in the three models was to conduct the analysis of type I error rate. Still, this external time-dependent covariate X_{2t} does not depend on a subject's survival. Thus, values of X_{2t} to be simulated were based on total precipitation in centimeters by state in the United States (National Climatic Data Center, 2001), assuming the normal distribution, $X_{2t} \sim N\,(94, 204)$, where 94 is the mean, and 204 is the variance of total precipitation in centimeters by state from 1971 to 2000. As most of the nitrogen dioxide comes from motor vehicle exhaust, X_{2t} is independent of X_{1t}. However, in order to help convergence in the algorithm used in GEE, the original distributions of both X_{1t} and X_{2t} had to be scaled, through trial and error, to $N\,(0.3, 0.06)$ and $N\,(0.3, 0.36)$, respectively, which would be discussed shortly.

Simulating event times. The second step of the data generation process was to generate an event time for each subject using the values of X_{1t} simulated for that subject. The event time variable was denoted by Υ, $0 \leq \Upsilon < \infty$, and Υ was measured in days.

Three assumptions were made regarding simulating event times. First, the simulated event times are non-informative in the sense that given external time-dependent covariates, an interval $(A_i, B_i]$ is not influenced by the specific value of the event time confined in $(A_i, B_i]$, that is, the occurrence of some particular event and the censoring time for the ith subject are independent. Second, the event times were assumed to follow the gamma distribution, $\Upsilon \sim GAM\,(\lambda, \rho)$, $\lambda > 0$, $\rho > 0$, where λ is the scale parameter, and ρ is the shape parameter. The reason for choosing the gamma distribution was two-fold. The first reason was the gamma distribution was used in prior empirical studies (Bhatt & Tiwari, 2014; Sastry, 1997). The second reason was the gamma distribution can accommodate a decreasing, monotonically baseline hazard function by letting $\rho < 1$, which is required in the current research. Third, it was assumed that the associated hazard $h(t)$ in the current research decreases monotonically during the follow-up, which was attained by letting $\rho < 1$. Here the scale parameter λ took the value of 50, and the shape parameter ρ took the partial form of Farrington's response probability in Equation 40,

$$\rho_{it} = 1 - e^{\left[-e^{(\beta_0 + \beta_1 \bar{X}_{1\cdot})}\right]}, \tag{54}$$

where the true parameter values $\beta_0 = 1.5$ and $\beta_1 = -3.6$, which were found through trial and error, and $\overline{X}_{1\cdot}$ refers to the mean of 100 simulated values of X_{1t}. The simulated expected event time for the ith subject at the tth examination was the product of the scale parameter and the shape parameter, that is,

$$\mathrm{E}(\Upsilon) = 50 * \rho_{it}. \tag{55}$$

In selecting λ and the true values of β_0 and β_1 for the shape parameter ρ in Equation 54 through trial and error, originally $\lambda = 75$, $\beta_0 = .04$, and $\beta_1 = -.011$. However, when the original distribution of X_{1t}, that is, $X_{1t} \sim N\,(79, 484)$, together with $\lambda = 75$, $\beta_0 = .04$, and $\beta_1 = -.011$, was used to generate event times via Equation 55, one unexpected situation occurred: the estimation of β_0, β_1, and β_2 from the EGNM failed to converge. The reason was found to be that the values, which were calculated from Equation 54 and were required in GEE for obtaining the parameter estimates from the EGNM, were very close to 0, which in turn produced noninvertible matrices. As such, in order to help convergence, the original distributions of both X_{1t} and X_{2t} were then scaled, through trial and error, to N (0.3, 0.06) and N (0.3, 0.36), respectively, and accordingly $\lambda = 50$, $\beta_0 = 1.5$, and $\beta_1 = -3.6$.

The expression for the shape parameter ρ, compared to Equation 40, actually dropped the term that summarized log ratios of the baseline survival functions at consecutive examinations. The reason was two-fold as well. On one hand, the parameters θ_gs for the indicator variables in the summation were nuisance parameters per se, which were not of direct inferential interest. On the other hand, no previous simulation studies were conducted on how these indicator variables were generated. As such, the response probability for the proposed model in Equation 40 that took the summation of log ratios of the baseline survival functions into account was different than the model used solely to generate event times in Equation 54.

The reason $\lambda = 50$ was chosen was two-fold. First, it was assumed that the mean of all simulated expected event times in this simulation study was 40 days or so. Second, it was further assumed that a majority of the simulated events happened at a later time in the follow-up study. In other words, if all the simulated expected event times were represented by a histogram, the histogram would be left-skewed.

There are three theoretical considerations and also empirical evidence for the choice of true parameter values $\beta_0 = 1.5$ and $\beta_1 = -3.6$. The first theoretical consideration is that true parameter values are chosen such that $\rho < 1$ is guaranteed, for when $\rho < 1$, the associated hazard function decreases monotonically. As the mean of event times is linked to true parameter values via the exponential function, which is invertible, there are one-to-one relationships between true parameter values and event times. Second, the follow-up study was assumed to last for around 60 days. The length of 60 days was chosen for three reasons. The first reason is 60 days is a reasonable time length for a healthy infant to be infected with some chronic respiratory disease under polluted air due to environmental factors (Cherian, Simoes, John, Steinhoff, & John, 1988), such as nitrogen dioxide. The second reason is the length of 60 days is based on prior empirical evidence. The third reason is the 60th day, which denotes the end of the follow-up, is later than a simulated expected event time, which is around the 40th day, as the simulated expected event time is confined in the last two examinations. Third, the majority of events were assumed to occur later in the follow-up study, for it takes time for some chronic respiratory disease to develop, and the hazard of experiencing some chronic respiratory disease was assumed to decrease monotonically during the follow-up study. The empirical evidence for the choice of true parameter values is that, as was seen from prior empirical evidence, unstandardized true parameter values for risk factors have ranged from -4 to 9, with a majority ranging from -1 to 1. Thus for the current study, β_1 = -3.6 was chosen to have similar and also large enough magnitude. Regarding the coefficient for the intercept constant term, which was also unstandardized, β_0 = 1.5 was chosen so that the mean of all simulated expected event times, via Equation 55, was 40 days or so, that is,

$$\mathrm{E}(\Upsilon) = 50 * \left\{1 - e^{\left[-e^{(1.5-3.6*\bar{X}_{1\cdot})}\right]}\right\}.$$

These simulated values were then rounded off to the nearest integer. Through trial and error, it was confirmed that the product of λ = 50 and the scale parameter ρ gave the mean of all simulated expected event times around 40 days, and most of the simulated events were clustered near the end of the follow-up study.

Simulating arbitrarily interval-censored survival data. The third step of the data generation process was to generate arbitrary intervals, where the simulated event times are confined, from the simulated event times. There are many ways to generate arbitrarily interval-censored survival data. For example, in Calle and Gómez's (2005) method, the censoring mechanism of the event time mimics a longitudinal study in which there is a periodic follow-up with scheduled examinations, taking into account that subjects might miss some of their examinations. For the current research, Zhang's (2009) naive way of simulating intervals was modified to generate arbitrarily interval-censored data. In particular, for the *i*th subject with a generated event time τ_i, which was rounded off, two random quantities, denoted by $U^{(1)}$ and $U^{(2)}$, respectively, were taken from a uniform distribution in the interval (0, c). These two quantities were then subtracted from or added to the simulated event time τ_i to form an interval $(\tau_i - U_i^{(1)}, \tau_i + U_i^{(2)}]$, that is, $A_i = \tau_i - U_i^{(1)}$ and $B_i = \tau_i + U_i^{(2)}$. However, this naive censoring interval is not non-informative, as the above uniform distribution is known to have bounded support. One way to go around this problem is by constructing $A_i^* = \max\left(\tau_i - U_i^{(1)}, \tau_i + U_i^{(2)} - c\right)$ and $B_i^* = \min\left(\tau_i + U_i^{(2)}, \tau_i - U_i^{(1)} + c\right)$, where A_i^* denotes the left censoring point, and where B_i^* denotes the right censoring point for a censoring interval. Roughly speaking, the purpose of this modification is to ensure that the width of the censoring interval does not exceed c, which is the upper bound of the above uniform distribution used to generate $U^{(1)}$ and $U^{(2)}$. It can be shown that this modified censoring interval satisfies the non-informative condition. Thus, censoring intervals were generated.

The upper bound of the uniform distribution c dictates the width of a censoring interval in Zhang's (2009) method. A comparatively wide censoring interval carries more uncertainty about when the event occurs than a comparatively narrow censoring interval. As such, how the width of a censoring interval affects the estimation of parameters becomes an interesting topic. Unfortunately, Zhang (2009) did not investigate this topic. Thus in the current research, both $c = 2$, which was thought to produce comparatively narrower censoring intervals, and $c = 5$, which was thought to produce comparatively wider censoring intervals, were investigated.

Simulating the number of examinations for each subject. For the fourth step of the data generation process, after an event time and the corresponding censoring interval were generated, the number of examinations for each subject was generated.

There were two questions associated with this step. The first question was: how many examinations should be simulated for each subject? In this simulation study, the numbers of examinations of all subjects were assumed to follow a binomial distribution, ranging from one to six randomly between study entry and the left endpoint of a simulated censoring interval. Four justifications were made for this choice of range. First, the range is similar to that in prior empirical evidence (He & Schaubel, 2014; Shen, Anderson, Sinha, & Li, 2014). Second, the range is reasonable as it considers both those who often have examinations and those who do not. Third, the range satisfies the non-informative condition as the specific number for one subject is random. Fourth, the range guarantees that 100 simulated external time-dependent covariate values are enough to be assigned to each simulated examination. The generated examinations for each subject were bound in an interval $(0, A_i)$, where A_i is the left endpoint of the censoring interval.

The second question was: how was the association between the external time-dependent covariate X_{1t} and the number of examinations established? As the response process, including that with all non-occurrence cases before the left censoring point A_i, and hence the number of examinations, is supposedly strongly associated with X_{1t} in the current study, the mean of the simulated X_{1t} values for each subject was used to generate the corresponding number of examinations.

It is worth mentioning that as it usually took some time for an event of interest, such as a certain type of respiratory disease, to display syndrome, i.e., the event of interest has occurred, the left endpoint of the censoring interval for each subject was assumed to be at least seven days from study entry. The reason for using seven was two-fold. The first reason was that six was the maximum possible simulated number of examinations for a subject. If six was simulated, the simulated sixth examination, where the event has not occurred, would be on the sixth day. When at least a one day gap was assumed between two consecutive examinations, the left endpoint of the censoring interval must be greater than six. The smallest possible left endpoint greater than six was seven. The second reason was when a simulated number of examinations was less than six, the left endpoint of a censoring interval could be any integer between two and six inclusive, which nevertheless caused the process of simulating censoring intervals across all subjects to be very complicated. In particular, if the left endpoint of a simulated censoring interval was six, the binomial distribution for the number of examinations per subject could not be used anymore as it might produce the sixth examination, which overlapped the left endpoint of the censoring interval. To avoid this situation, a new binomial distribution had to be employed for that subject. Thus, to facilitate the data generation process, no matter how many examinations from the binomial distribution were simulated for each subject, seven was used as the smallest left endpoint.

Arranging the sampled external time-dependent covariate values in descending order. After a number of examinations for each subject were simulated, the same number of X_{1t} and X_{2t} values were sampled without replacement from the 100 values of X_{1t} and X_{2t} that had been generated. The question was: how were the sampled X_{1t} values assigned to the simulated examinations? Figure 1 showed the relationship between a series of hypothetical X_{1t} values, represented by the x-axis, and the corresponding response probability, represented by the y-axis.

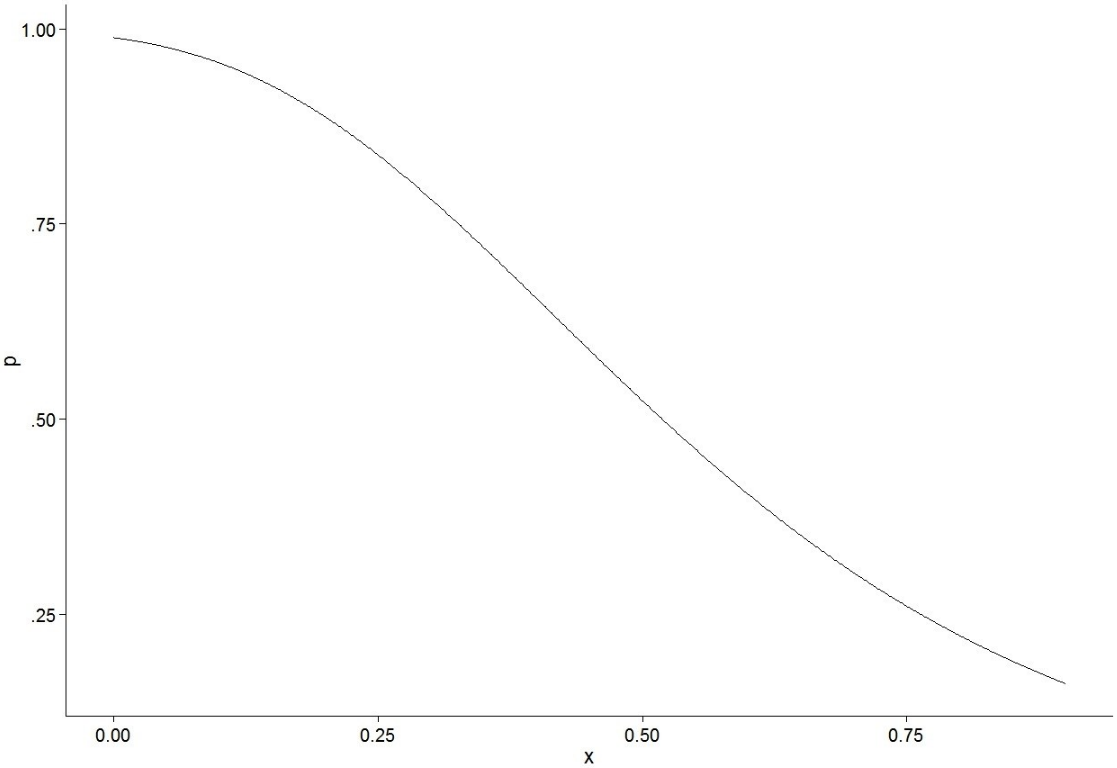

Figure 1. Association between the response probability and a series of hypothetical X_{1t} values.

The response probability was calculated via Equation 45, and $\beta_0 = 1.5$, and $\beta_1 = -3.6$, that is,

$$p_{it} = 1 - e^{\left[-e^{(1.5-3.6*X_{it})}\right]}.$$

Figure 1 revealed an apparent monotonically decreasing pattern between X_{1t} and the response probability. As the values of X_{1t} increase, the corresponding response probabilities decrease. As such, sampled X_{1t} values, which were in random order, had to be arranged in descending order in order to establish strong association between X_{1t} and the response probability. However, the pattern shown in Figure 1 was too perfect to be true in reality. Besides the descending association, for example, the ascending association exists as well. Then how the probability of the descending association between X_{1t} and the response probability affects the estimation of parameters becomes another interesting topic. Still no one has yet investigated this topic. In the current research, $\varrho = .3$ and $\varrho = .7$ were investigated, where ϱ represents the probability of the descending association between X_{1t} and the response probability. The reason for choosing $\varrho = .3$ and $\varrho = .7$ was both were equal distance from $\varrho = .5$. In particular, after X_{1t} values were sampled for all subjects, X_{1t} values of 30% or 70%, that is, $\varrho = .3$ or $\varrho = .7$, of all subjects were arranged in descending order within each subject. In this way, 30% and 70% of subjects, respectively, had the event at the smallest value of X_{1t}.

Organizing the Simulated Data for Fitting the Three Models

Before the simulated data were used for fitting each of the three models, they had to be organized, respectively, which is required for analysis of such data.

For the EGNM, the method of augmenting the collected data in Farrington's approach was modified accordingly, as Farrington's approach further incorporates external time-dependent covariates. In particular, after simulated survival data were organized in a data set such that information regarding covariates, left and right censoring times for an interval, and the binary response variable indicating the status of an event for each subject was recorded using one single line, the data set was expanded by adding a further $\left(\sum_{i=1}^{N} a_i - a\right)$ line of data, repeating the information for subjects whose intervals were confined, so that the revised data set had $\left(N + \sum_{i=1}^{N} a_i - a\right)$ observations, where a_i denoted the number of line of data for a confined case, and a denoted the number of confined cases. Note that the use of a_i referred to the fact that subjects had different numbers of examinations, and hence different numbers of responses. The values, for example, y_{it}, of the binary response variable, Y_{it} denoting the response at the tth examination for the ith subject were then added. For confined cases, where the data were duplicated, all examinations prior to the last one has $Y_{it} = 0$ and the last examination where $t = T_i$ has $Y_{iT_i} = 1$. The values of the D_{ig}, $g = 1, 2, \ldots, k$, differed at each examination time, t_i. Regarding values of the external dependent covariate X_{1t}, they were simply incorporated to each of $\left(N + \sum_{i=1}^{N} a_i - a\right)$ lines of data accordingly. For the purpose of analyzing type I error rate, values of another external dependent covariate X_{2t} were incorporated to the data similarly.

The data modeled using the extended Cox model were from the simulated data, and were almost identical to those used for fitting the EGNM, including the use of $\varrho = .3$ and $\varrho = .7$, which represented the probability of the descending association between X_{1t} and the hazards associated with an occurrence case. The only difference lay in that the event time for the ith subject was imputed from the left and right endpoints of the simulated censoring interval using the mid-point imputation method (Law & Brookmeyer, 1992), as the extended Cox model required an exact event time.

The way the data for fitting Farrington's model were organized from the simulated data is similar to that for fitting the EGNM. The major difference lay in that the data for each subject was augmented in two lines. One line was for study entry to the left endpoint of the censoring interval, and the other line was for the censoring interval. Moreover, in both lines, covariates assumed values simulated for study entry alone.

Software Used for the Current Research

The platform on which the simulated data were generated is the software package R (Version 3.2.2). The packages *survival*, *bbmle*, *foreach*, *iterators*, *optimx*, *plyr*, *dplyr*, and *ggplot2* were used for the analyses.

Data Analyses

Steps in the Data Analyses

After the data for fitting the three models in the simulation study were simulated and augmented, and the numbers of subjects, sample sizes and the number of replications were determined, the simulation study in the current research was conducted. In summary, the first step was to fit each of the three models to the simulated and organized data to obtain the parameter estimates. The second step was to evaluate properties of the obtained parameter estimates across the three models, including precision of the parameter estimates, power, and type I error rate.

Fitting models. After the data needed for the simulation study were simulated and organized, they were fitted into three models, respectively, namely, the EGNM that accounts for both arbitrarily interval-censored data and external time-dependent covariates, the extended Cox model that accounts for external time-dependent covariates but ignores arbitrarily interval-censored data, and Farrington's model that accounts for arbitrarily interval-censored data but ignores external time-dependent covariates.

Fitting the EGNM. The non-likelihood-based estimation method GEE was applied to estimate the parameters of the EGNM. GEE for the proposed model was,

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left[\frac{\partial \boldsymbol{\mu}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]' \left(\phi \mathbf{A}_i^{1/2} \boldsymbol{I} \mathbf{A}_i^{1/2}\right)^{-1} [\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = \mathbf{0},$$

where $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ represented the mean vector for the ith subject in the case of confined data, an identity matrix, $\boldsymbol{I}$, represented the correlation present within the cluster of responses for one particular subject, the diagonal elements of $\mathbf{A}_i$ shown in Equation 53 represented the response variance under the assumption of independence, and $\boldsymbol{y}_i$ represented the response vector, which referred to all responses during the follow-up for one subject, and took the form $\boldsymbol{y}_i = [y_{i1},\ldots, y_{iTi}]'$. The mean vector for the ith subject $\boldsymbol{\mu}_i(\boldsymbol{\beta})$ took the form,

$$\boldsymbol{\mu}_i(\boldsymbol{\beta}) = \begin{bmatrix} e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}} \\ \dfrac{e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_{i1})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ \dfrac{e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_{i3})}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{(\boldsymbol{\beta}'\boldsymbol{X}_{i2})}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ \vdots \\ \dfrac{e^{\left\{\left[-e^{\left(\boldsymbol{\beta}'\boldsymbol{X}_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\boldsymbol{\beta}'\boldsymbol{X}_{i(T_i-2)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \\ 1 - \dfrac{e^{\left\{\left[-e^{\left(\boldsymbol{\beta}'\boldsymbol{X}_{iT_i}\right)}\right]\sum_{g=1}^{k}\theta_g d_{1ig}\right\}}}{e^{\left\{\left[-e^{\left(\boldsymbol{\beta}'\boldsymbol{X}_{i(T_i-1)}\right)}\right]\sum_{g=1}^{k}\theta_g d_{2ig}\right\}}} \end{bmatrix},$$

where $\boldsymbol{X}_{iT_i}$ was the three-dimensional vector of the intercept constant and two time-dependent covariates X_{1t} and X_{2t} collected at the T_ith examination for the ith subject, $\boldsymbol{\beta}$ is the three-dimensional vector of unknown parameters β_0, β_1, and β_2. Solving these estimating equations provided the parameter estimates.

Fitting the extended Cox model. The extended Cox model in this simulation study took the form,

$$h_i(\tilde{t}|\boldsymbol{\beta}, \boldsymbol{X}_{it}) = h_0(\tilde{t})e^{(\boldsymbol{\beta}'\boldsymbol{X}_{it})},$$

where $\tilde{t}$ was the imputed event time, $\boldsymbol{X}_{it}$ was the three-dimensional vector of the intercept constant and two external time-dependent covariates X_{1t} and X_{2t} collected at time t for the ith subject, $\boldsymbol{\beta}$ was the three-dimensional vector of unknown parameters β_0, β_1, and β_2, and $h_0(\tilde{t})$ was the baseline hazard function. The corresponding partial log-likelihood function took the form

$$l(\boldsymbol{\beta}|X_{it}) = \sum_{i=1}^{N} \delta_i \left[(\boldsymbol{\beta}'\boldsymbol{X}_{it}) - \log \sum_{l \in R(t_i)} e^{(\boldsymbol{\beta}'\boldsymbol{X}_{lt})} \right],$$

where $\delta_i = 0$ if the survival time of the ith subject is censored and $\delta_i = 1$ otherwise. This equation was then maximized using numerical methods to obtain parameter estimates.

It is worth mentioning that the extended Cox model does not estimate an intercept term. This is because the parameter is unidentifiable, as the exponentiated intercept term is subsumed by the unknown baseline hazard function, thus any intercept term would simply change the baseline hazard function. As such, the inclusion of the intercept term in the EGNM would help estimation of parameters in general.

Fitting Farrington's model. The likelihood function for Farrington's model, denoted L^{**} ($\boldsymbol{\beta}|\boldsymbol{X}$), took the form,

$$L^{**}(\boldsymbol{\beta}|\boldsymbol{X}) = \prod_{i=1}^{N+a} p_i^{y_i}(1-p_i)^{1-y_i},$$

where y_i was the binary response variable, for i =1, 2, ... , $N + a$, indicating the number of rows in the augmented data set, and the response probability took the form,

$$p_i = 1 - e^{\left\{\left[-e^{(\beta' X_i)}\right]\sum_{g=1}^{k}\theta_g d_{ig}\right\}},$$

where $\boldsymbol{X}_i$ was the three-dimensional vector of the intercept constant and two time-independent covariates X_1 and X_2, $\boldsymbol{\beta}$ was the three-dimensional vector of unknown parameters β_0, β_1, and β_2, θ_g was the log ratio of the baseline survival functions at the (g – 1)th and the gth ordered examinations, and d_{ig} was the indicator variable for the gth ordered examination. The maximum likelihood estimation method via numerical methods was used to obtain parameter estimates.

Evaluating properties of the parameter estimates. After the parameter estimates were obtained from each model, their properties were evaluated from four perspectives: ARB and percent of correct sign of the parameter estimates, power, and type I error rate.

First, regarding precision of the parameter estimate, ARB of the parameter estimates from each model, that is, the absolute value of the difference between the parameter estimates and the true values of the coefficients divided by of the coefficients, was calculated. Smaller ARB means more precise parameter estimates. Although the accepted bias in previous simulation studies on survival analysis ranged from -0.001 (He & Schaubel, 2014) to 0.014 (Schaubel, Zhang, Kalbfleisch, & Shu, 2014), the criteria used to evaluate ARB in the current study was the cutoff point 0.01, which was chosen due to that the computers used to conduct the simulation study were capable of accommodating 1,000 subjects and 150 replications, and 350 subjects and 500 replications at most.

Second, regarding the percent of correct sign of the parameter estimates, which represented the feasibility of the parameter estimates, eighty percent (McCombie & Thirlwall, 2004) was used as the criterion. Thus, in the current study, a model with 80% percent of correct sign or higher of the parameter estimates was acceptable, indicating the model fit the simulated data well.

Third, regarding power of a model, which represented that model's capability of detecting the significance of covariates when covariates are significant indeed, although .85 (Brendel, Janssen, Mayer, & Pauly, 2014) was acceptable, .90 (Whitehead, 2014) was used as the criterion, which was the percent of the *p*-values of X_{1t} obtained from the hypothesis testing in all replications less than or equal to .05. If the power from a model was greater than or equal to .90, the model fit the simulated data well.

Fourth, regarding analysis of type I error rate, which meant the parameter estimates with the *p*-values less than or equal to .05 in hypothesis testing are not significant indeed, the nominal level of .05 (Pocock, Geller, & Tsiatis, 1987), also the typical choice, was used as the criterion, which was the percent of the *p*-values of X_{2t} obtained from hypothesis testing in all replications less than or equal to .05. The model which gave type I error rate closer to the nominal level of .05 was preferable.

Chapter Summary

In this chapter, I proposed the non-likelihood-based estimation method GEE (Liang & Zeger, 1986; Zeger & Liang, 1986) for the EGNM, which accommodates both arbitrarily interval-censored survival data and external time-dependent covariates simultaneously. However, it was found through trial and error that only when the distribution of the significant covariate X_{1t} was scaled did the parameter estimation converge.

In the simulation design conditions, for each subject, a censoring interval, a number of examinations and the corresponding number of X_{1t} and X_{2t} values, were simulated. Moreover, due to the unique form of the proposed expression for the response probability, ϱ, denoting probability the smallest X_{1t} value is associated with the response probability, was introduced to establish strong association between X_{1t} and the response probability, and c, dictating the width of simulated intervals, was also introduced.

In order to show how the simulation results behaved as the number of subjects increased in the simulation study, two sets of simulation results were presented.

Properties of the parameter estimates were evaluated from four perspectives: ARB and percent of correct sign of the parameter estimates, power, and type I error rate. The criterion used to evaluate ARB was the cutoff point .01. Eighty percent was used as the criterion to evaluate percent of correct sign of the parameter estimates. For power and type I error rate, the criteria used were .90 and the nominal level of .05.

CHAPTER IV

RESULTS

The simulation results are reported, and presented in tables and figures in this chapter, including selected descriptive statistics from the simulated data, precision of the parameter estimates of β_0, β_1, and β_2, percent of correct sign of the parameter estimate of β_1, confidence intervals of the parameter estimates of β_1 and β_2, power, and type I error rate.

As in the simulation study, in addition to two different sets of numbers of subjects and replications, $c = 2$ and $c = 5$, and $\varrho = .3$ and $\varrho = .7$ were used to investigate the impact of the upper bound of the uniform distribution, which dictates the width of a simulated censoring interval, denoted by c, and the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, denoted by ϱ, respectively, on the estimation of the parameters. The simulation results are first represented and then summarized under each combination of the conditions.

Moreover, at the end of this chapter, four comprehensive tables were created to show under each combination of c and ϱ, how ARB of the mean parameter estimate of β_1, denoted by ARB($\bar{\bar{\beta}}_1$), power, and type I error rate behaved as the number of subjects increased.

Simulation Results

Descriptive Statistics

To demonstrate correct data generation, selected descriptive statistics are reported first, including mean, maximum, and minimum of the left and right censoring points, the mean of the numbers of examinations from the simulated data for fitting the EGNM and Farrington's model, and the mean of the event times from the simulated data for fitting the extended Cox model. These statistics are displayed in Table 1-Table 8. The selected statistics from 150 replications are shown first, followed by the selected statistics from 500 replications.

Table 1

Selected Descriptive Statistics from the Simulated Data (R150, c = 2, ϱ =.3)

S	*Min(l)*[a]	*Max(l)*[b]	*Mn(l)*[c]	*Min(r)*[d]	*Max(r)*[e]	*Mn(r)*[f]	*Mn(e)*[g]
50	21	68	38	22	69	40	39
250	18	68	38	19	69	40	39
500	18	68	38	19	69	40	39
1000	18	71	38	19	72	40	39

Note. R = the number of replications. *S* = the number of subjects.
[a]*Min(l)* refers to the minimum simulated left censoring point. [b]*Max(l)* refers to the maximum simulated left censoring point. [c]*Mn(l)* refers to the mean of the simulated left censoring points. [d]*Min(r)* refers to the minimum simulated right censoring point. [e]*Max(r)* refers to the maximum simulated right censoring point. [f]*Mn(r)* refers to the mean of the simulated right censoring points. [g]*Mn(e)* refers to the mean imputed event time.

Table 2

Selected Descriptive Statistics from the Simulated Data (R150, c = 5, ϱ =.3)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	19	67	37	23	69	41	39
250	16	67	37	20	69	41	39
500	16	67	37	20	71	41	39
1000	16	69	37	20	73	41	39

Table 3

Selected Descriptive Statistics from the Simulated Data (R150, $c = 2$, $\varrho = .7$)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	21	68	38	22	69	40	39
250	18	68	38	19	69	40	39
500	18	68	38	19	69	40	39
1000	18	71	38	19	72	40	39

Table 4

Selected Descriptive Statistics from the Simulated Data (R150, $c = 5$, $\varrho = .7$)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	19	67	37	23	69	41	39
250	16	67	37	20	69	41	39
500	16	67	37	20	71	41	39
1000	16	69	37	20	73	41	39

Table 5

Selected Descriptive Statistics from the Simulated Data (R500, $c = 2$, $\varrho = .3$)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	19	66	38	20	67	40	39
150	18	65	38	20	66	40	39
250	18	68	38	19	69	40	39
350	15	71	38	16	73	40	39

Table 6

Selected Descriptive Statistics from the Simulated Data (R500, $c = 5$, $\varrho = .3$)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	17	68	37	22	72	41	39
150	17	68	37	21	72	41	39
250	16	67	37	20	71	41	39
350	16	70	37	20	74	41	39

Table 7

Selected Descriptive Statistics from the Simulated Data (R500, c = 2, ϱ =.7)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	18	68	38	19	69	40	39
150	18	68	38	19	69	40	39
250	18	68	38	19	69	40	39
350	18	71	38	19	72	40	39

Table 8

Selected Descriptive Statistics from the Simulated Data (R500, c = 5, ϱ =.7)

S	*Min(l)*	*Max(l)*	*Mn(l)*	*Min(r)*	*Max(r)*	*Mn(r)*	*Mn(e)*
50	16	67	37	20	69	41	39
150	16	67	37	20	71	41	39
250	16	67	37	20	71	41	39
350	16	69	37	20	73	41	39

Across all eight tables with various combination of conditions, the mean of the simulated expected event time was around 40 days, which lay between the mean left censoring point and the mean right censoring point, and thus satisfied the intended design. The maximum right endpoint was around 70 days, which roughly satisfied the intended design that this simulation study lasted for around 60 days. The mean imputed event time used for the extended Cox model was around 39 days, which was roughly in the middle of a censoring interval formed by the mean left censoring point and the mean right censoring point, and thus satisfied the intended design.

Precision of the Parameter Estimates

Regarding precision of the parameter estimates of β_1 and β_2, the mean parameter estimates of β_1 and β_2, denoted by $\bar{\bar{\beta}}_1$ and $\bar{\bar{\beta}}_2$, respectively, the corresponding mean standard errors of $\bar{\bar{\beta}}_1$, denoted by se($\bar{\bar{\beta}}_1$), ARB($\bar{\bar{\beta}}_1$), and percent of correct sign of the parameter estimate of $\hat{\beta}_1$, denoted by % CS($\hat{\beta}_1$), for the three models are presented in Table 9-Table 16, and the confidence intervals for $\bar{\bar{\beta}}_1$ and $\bar{\bar{\beta}}_2$ are presented in Table 17-Table 32. The simulation results regarding precision of $\bar{\bar{\beta}}_0$, i.e., the mean parameter estimate of β_0, are represented in Table 33-Table 40.

To visually check the results regarding precision of the parameter estimate of β_1, two figures were created for each table with each combination of the conditions to show ARB($\bar{\bar{\beta}}_1$) and % CS($\hat{\beta}_1$), respectively.

The results regarding precision from 150 replications are shown first, followed by the results from 500 replications.

Table 9

Precision of the Parameter Estimates of β_1 and β_2 (R150, c = 2, ρ =.3)

	M	$S = 50$	$S = 250$	$S = 500$	$S = 1000$
$\bar{\bar{\beta}}_1$(se)	C[i]	1.1286(0.6341)	0.9954(0.2649)	1.0145(0.1845)	0.9984(0.1303)
	F[j]	-0.0020(0.3474)	-0.0108(0.1349)	-0.0139(0.0944)	-0.0184(0.0688)
	E[k]	-1.4187(0.6233)	-1.4370(0.2757)	-1.4481(0.1951)	-1.4648(0.1378)
$\bar{\bar{\beta}}_2$(se)	C	0.0023(0.2633)	-0.0019(0.1100)	-0.0022(0.0773)	-0.0049(0.0546)
	F	-0.0020(0.2408)	-0.0089(0.0917)	-0.0112(0.0647)	-0.0150(0.0471)
	E	0.0125(0.2369)	0.0042(0.1059)	0.0007(0.0752)	-0.0029(0.0532)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3135	1.2765	1.2818	1.2773
	F	0.9995	0.9970	0.9961	0.9949
	E	0.6059	0.6008	0.5977	0.5931
% CS ($\hat{\beta}_1$)	C	8.0	0	0	0
	F	78.7	100	100	100
	E	99.3	100	100	100

Note. M = Model. se = Standard errors. The true value of β_1 is -3.6. The true value of β_2 is 0.
[i]C refers to the extended Cox model. [j]F refers to Farrington's model. [k]E refers to the EGNM.

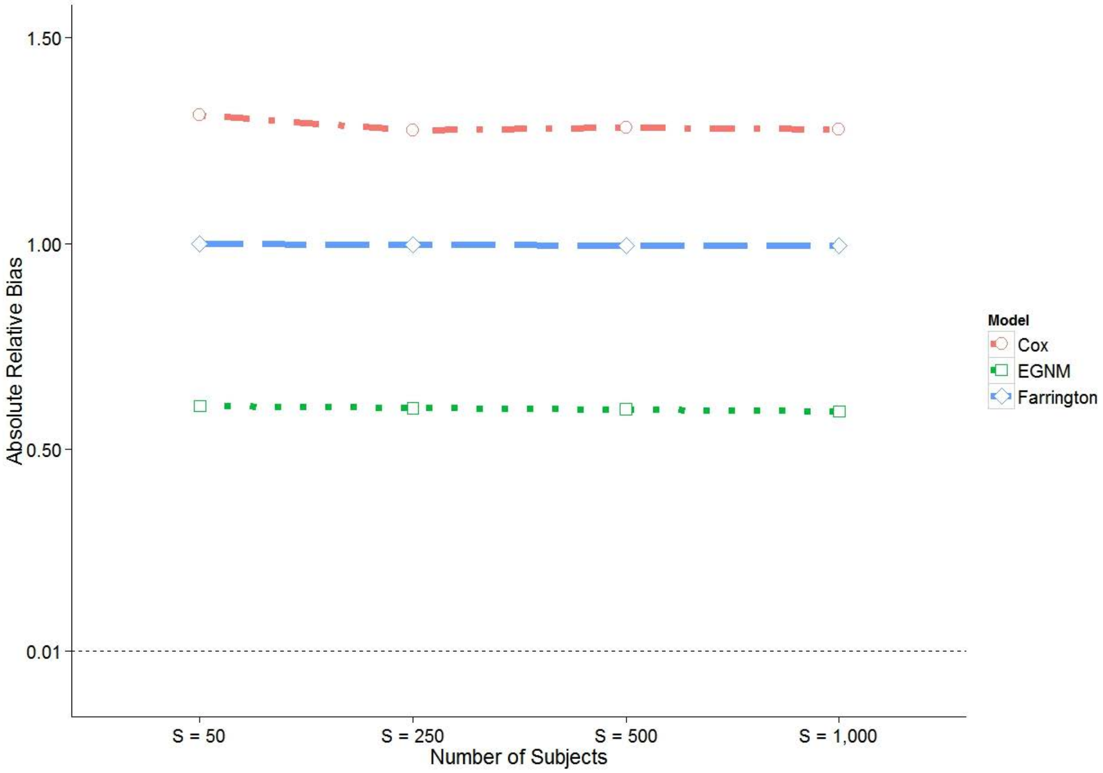

Figure 2. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R150, c = 2, ϱ = .3).

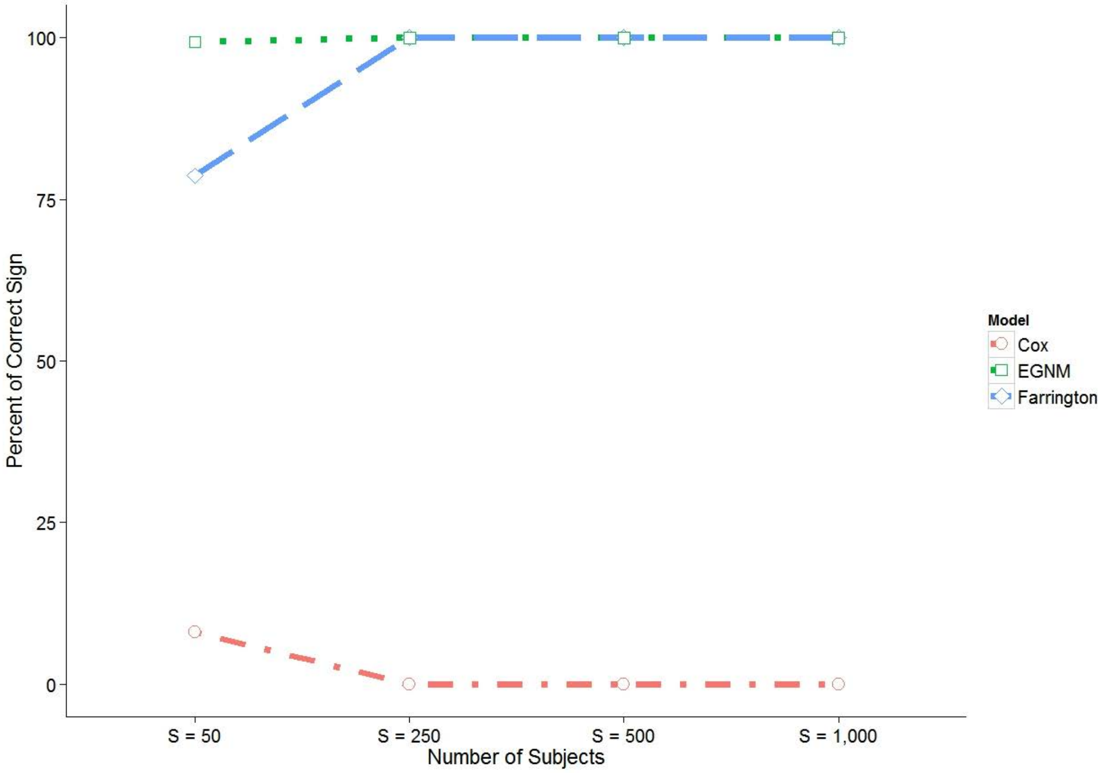

Figure 3. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R150, *c* = 2, ϱ = .3).

When the number of replications is 150, and *c* = 2 and ϱ = .3, that is, the widths of simulated censoring intervals are comparatively narrow, and the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is comparatively low, $\bar{\bar{\beta}}_1$ from any model was far from the true value -3.6 and was substantially underestimated, but $\bar{\bar{\beta}}_2$ from any model was close to 0, the true value of β_2; ARB($\bar{\bar{\beta}}_1$) from any model was not acceptable at the .01 level; % CS($\hat{\beta}_1$) from the EGNM was acceptable at the 80% level, % CS($\hat{\beta}_1$) from Farrington's model was acceptable only when the number of subjects was greater than 50, but % CS($\hat{\beta}_1$) from the extended Cox model was not acceptable in any case.

Table 10

Precision of the Parameter Estimates of β_1 and β_2 (R150, c = 5, ϱ =.3)

	M	*S* = 50	*S* = 250	*S* = 500	*S* = 1000
$\bar{\bar{\beta}}_1$(se)	C	1.0862(0.6195)	0.9637(0.2587)	0.9798(0.1800)	0.9635(0.1271)
	F	0.0069(0.3641)	-0.0028(0.1544)	-0.0078(0.1013)	-0.0105(0.0674)
	E	-1.4187(0.6233)	-1.4392(0.2758)	-1.4481(0.1951)	-1.4648(0.1378)
$\bar{\bar{\beta}}_2$(se)	C	0.0025(0.2617)	-0.0164(0.1099)	-0.0013(.0769)	-0.0030(0.0543)
	F	0.0069(0.2522)	-0.0025(0.1062)	-0.0065(.0693)	-0.0086(0.0462)
	E	0.0125(0.2369)	-0.0001(0.1063)	0.0007(.0752)	-0.0029(0.0532)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3017	1.2677	1.2722	1.2676
	F	1.0019	0.9992	0.9978	0.9971
	E	0.6059	0.6002	0.5977	0.5931
% CS ($\hat{\beta}_1$)	C	7.3	0	0	0
	F	8.0	86.7	100	100
	E	99.3	100	100	100

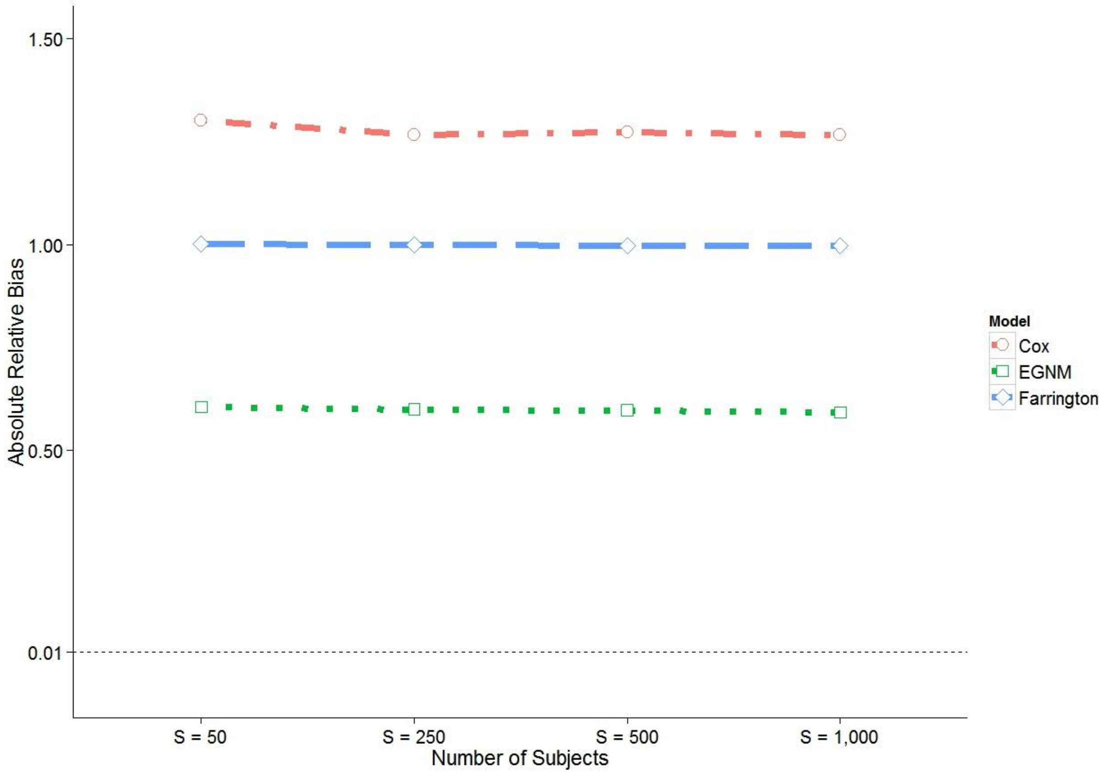

Figure 4. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R150, *c* = 5, ϱ = .3).

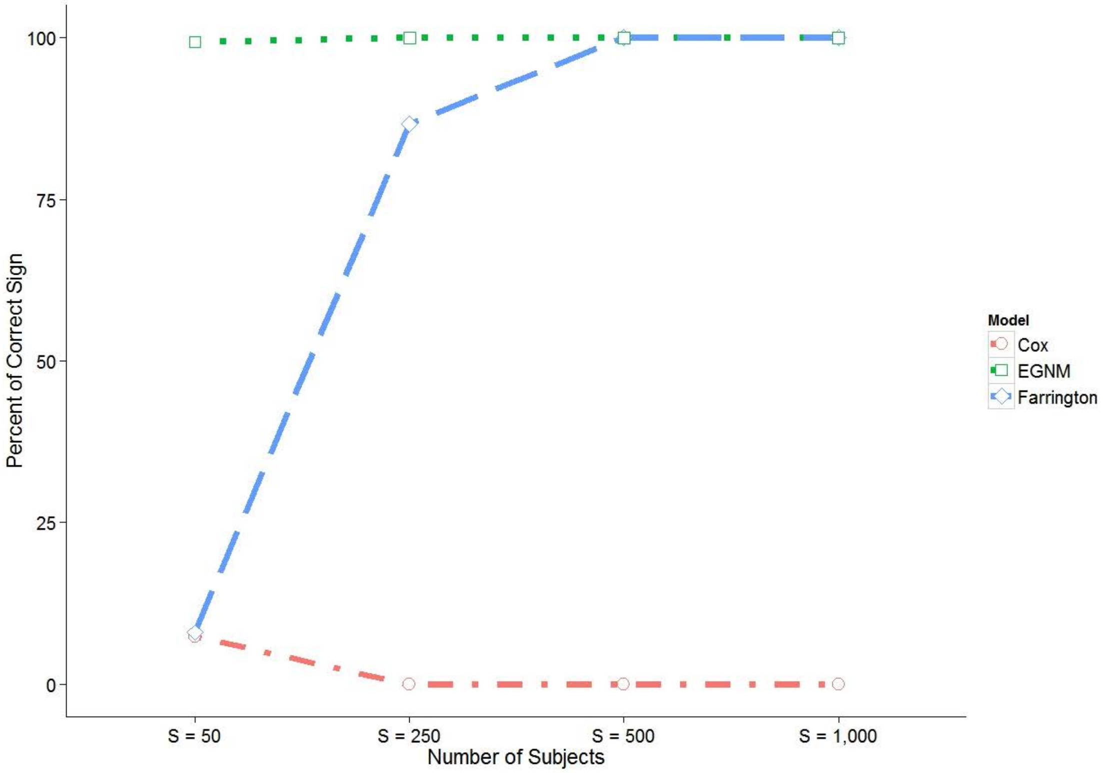

Figure 5. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R150, *c* = 5, ϱ = .3).

When the number of replications is 150, and *c* = 5 and ϱ = .3, that is, compared to *c* = 2 and ϱ = .3, the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is still comparatively low, but simulated censoring intervals are lengthened, the results were similar to those with *c* = 2 and ϱ = .3. However, % CS($\hat{\beta}_1$) from Farrington's model was acceptable at the 80% level only when the number of subjects was at least 250.

Table 11

Precision of the Parameter Estimates of β_1 and β_2 (R150, c = 2, ϱ =.7)

	M	$S = 50$	$S = 250$	$S = 500$	$S = 1000$
$\bar{\bar{\beta}}_1$(se)	C	1.4385(0.7924)	1.3819(0.3344)	1.4020(0.2327)	1.4274(0.1641)
	F	-0.0072(0.2963)	-0.0140(0.1155)	-0.0178(0.0808)	-0.0240(0.0593)
	E	-3.7621(0.7911)	-3.6785(0.3623)	-3.6681(0.2620)	-3.6503(0.1856)
$\bar{\bar{\beta}}_2$(se)	C	0.0105(0.2778)	-0.0085(0.1145)	-0.0041(0.0802)	-0.0034(0.0567)
	F	-0.0029(0.2407)	-0.0088(0.0931)	-0.0111(0.0648)	-0.0151(0.0475)
	E	0.0099(0.2435)	0.0032(0.1100)	-0.0013(0.0777)	-0.0046(0.0551)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3996	1.3839	1.3894	1.3965
	F	0.9980	0.9961	0.9950	0.9933
	E	0.0450	0.0218	0.0189	0.0140
% CS ($\hat{\beta}_1$)	C	8.0	0	0	0
	F	80.0	100	100	100
	E	100	100	100	100

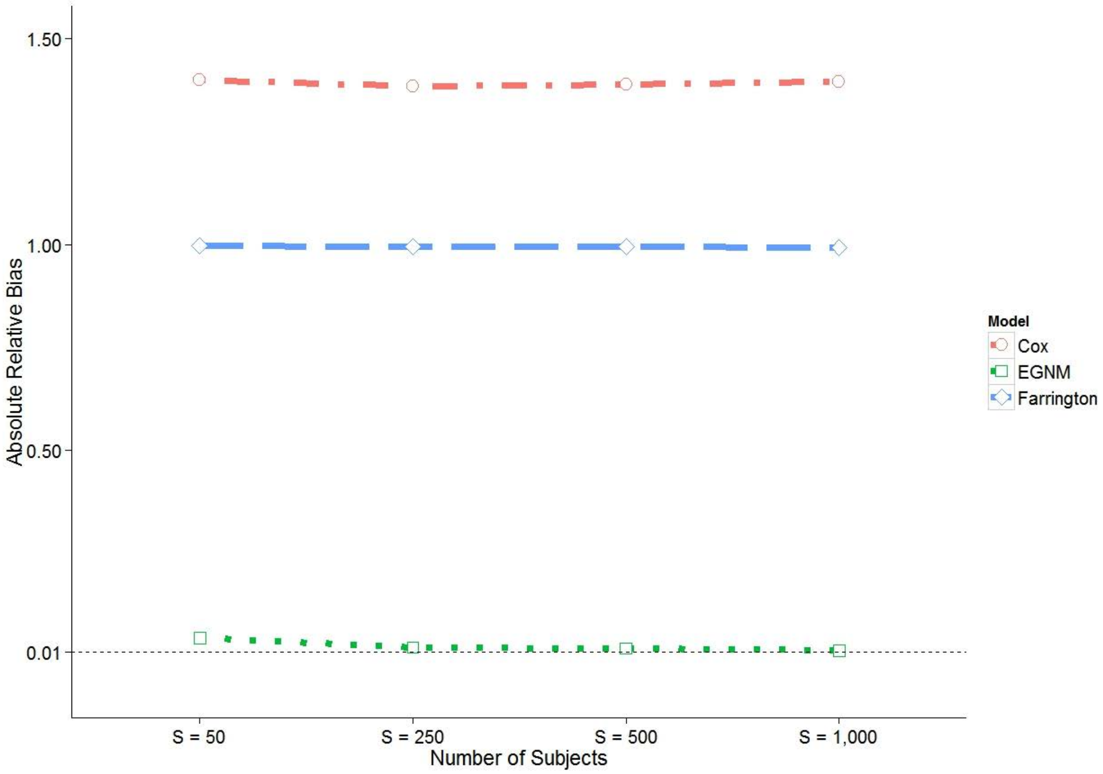

Figure 6. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R150, *c* = 2, ϱ = .7).

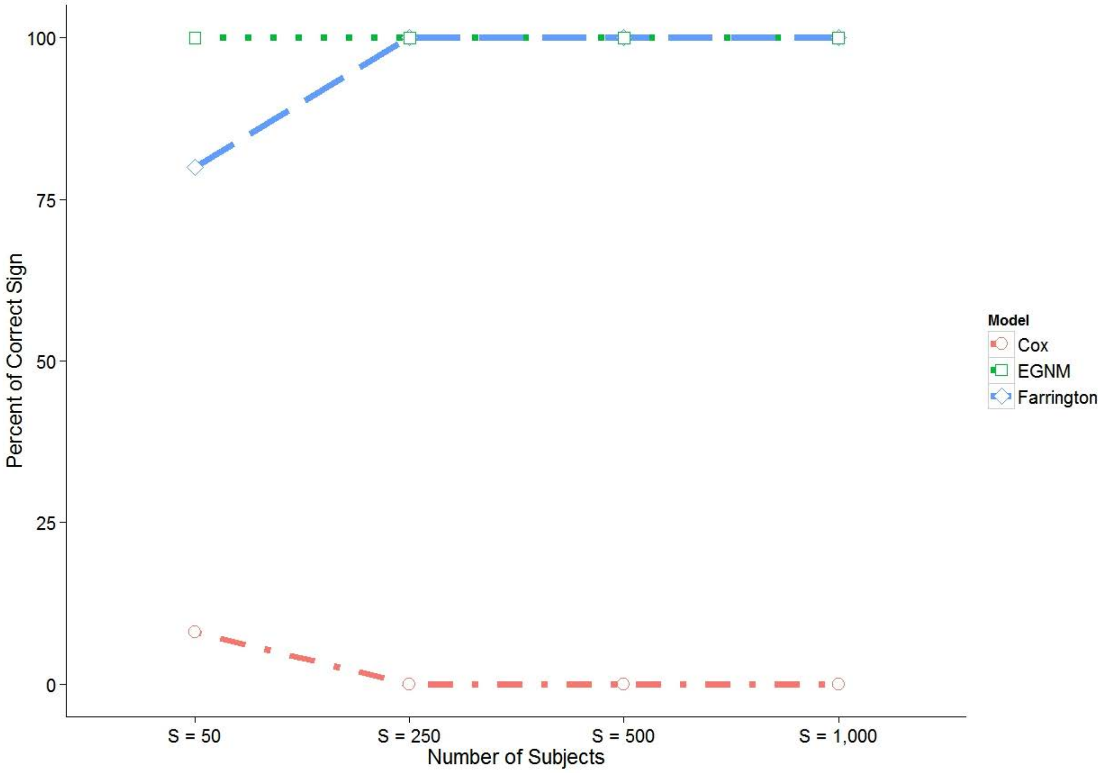

Figure 7. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R150, *c* = 2, ϱ = .7).

When the number of replications is 150, and *c* = 2 and ϱ = .7, that is, compared to *c* = 2 and ϱ = .3, the widths of simulated censoring intervals are still comparatively narrow, but the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, becomes high, only $\bar{\bar{\beta}}_1$ from the EGNM was close to the true value of -3.6, and $\bar{\bar{\beta}}_2$ from any model was close to 0; ARB($\bar{\bar{\beta}}_1$) from the EGNM was acceptable at the .01 level with at least 500 subjects, but ARB($\bar{\bar{\beta}}_1$) from the other two models was not acceptable; % CS($\hat{\beta}_1$) from both the EGNM and Farrington's model was acceptable at the 80% level, but % CS($\hat{\beta}_1$) from the extended Cox model was not acceptable in any case.

Table 12

Precision of the Parameter Estimates of β_1 and β_2 (R150, c = 5, ϱ =.7)

	M	S = 50	S = 250	S = 500	S = 1000
$\bar{\bar{\beta}}_1$(se)	C	1.4580(0.7652)	1.3598(0.3201)	1.3957(0.2230)	1.4180(0.1571)
	F	0.0397(0.3097)	-0.0034(0.1319)	-0.0100(0.0866)	-0.0134(0.0578)
	E	-3.7621(0.7911)	-3.6785(0.3623)	-3.6681(0.2620)	-3.6503(0.1856)
$\bar{\bar{\beta}}_2$(se)	C	0.0089(0.2747)	-0.0035(0.1131)	-0.0022(0.0794)	-0.0023(0.0562)
	F	0.0093(0.2514)	-0.0025(0.1062)	-0.0065(0.0694)	-0.0086(0.0462)
	E	0.0099(0.2435)	0.0032(0.1100)	-0.0013(0.0777)	-0.0048(0.0551)
ARB ($\bar{\bar{\beta}}_1$)	C	1.4050	1.3777	1.3877	1.3939
	F	1.0110	0.9990	0.9972	0.9963
	E	0.0450	0.0218	0.0189	0.0140
% CS ($\hat{\beta}_1$)	C	6.0	0	0	0
	F	6.0	84.7	100	100
	E	100	100	100	100

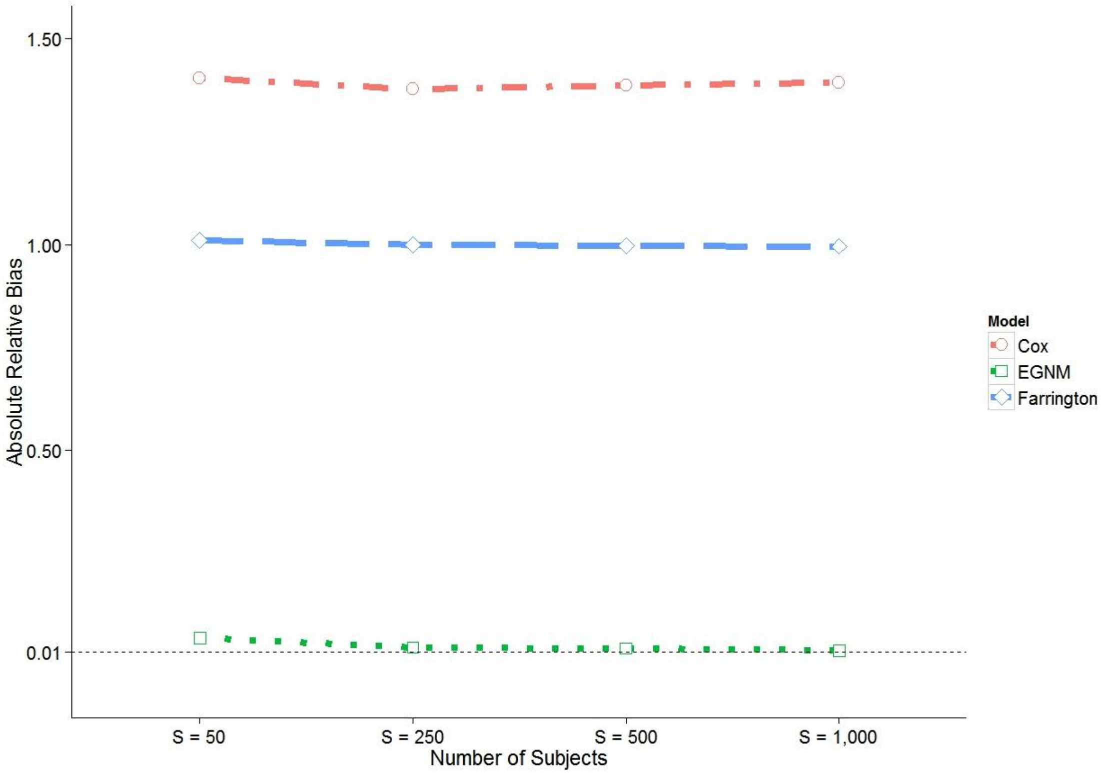

Figure 8. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R150, c = 5, ϱ = .7).

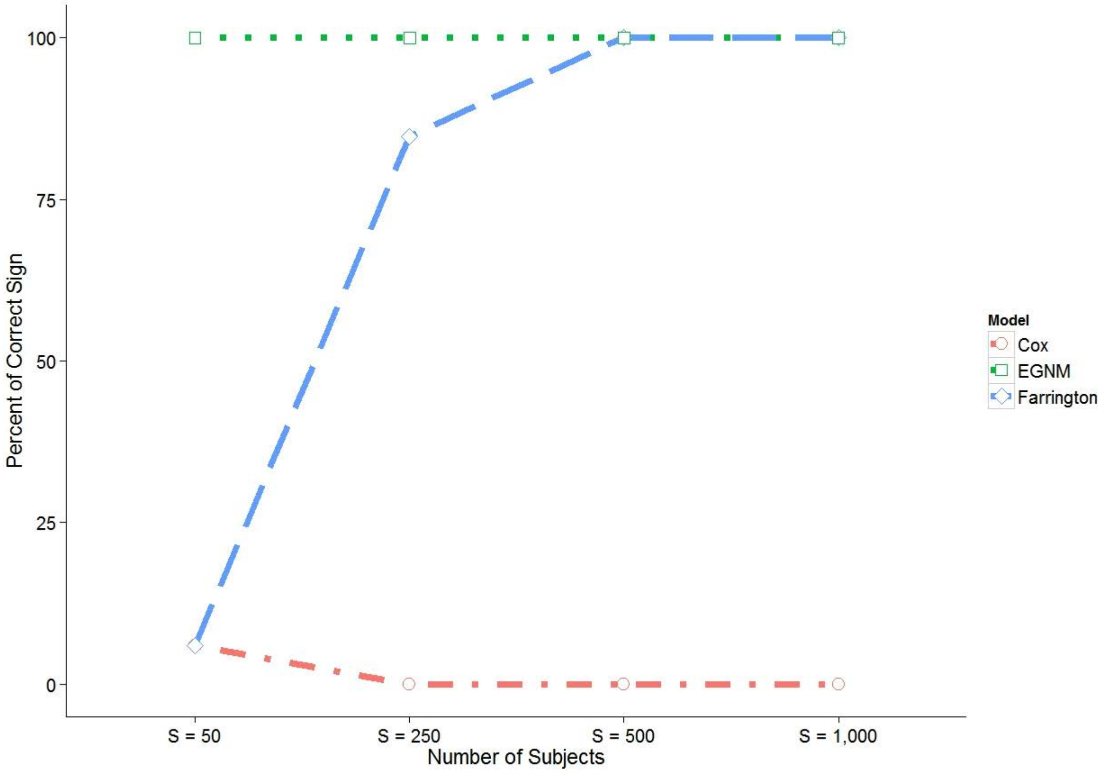

Figure 9. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R150, c = 5, ϱ = .7).

When the number of replications is 150, and c = 5 and ϱ = .7, that is, compared to c = 2 and ϱ = .7, the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is still comparatively high, but simulated censoring intervals are lengthened, the results were similar to those with c = 2 and ϱ = .7. However, % CS($\hat{\beta}_1$) from Farrington's model was acceptable at the 80% level only when the number of subjects was at least 250.

Compared to c = 5 and ϱ = .3, that is, simulated censoring intervals are still comparatively wide, but the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, was low, the results with c = 5 and ϱ = .7 were similar to those with c = 5 and ϱ = .3. The only difference lay in ARB($\bar{\bar{\beta}}_1$) from the EGNM with c = 5 and ϱ = .7 was acceptable at the .01 level with at least 500 subjects, but ARB($\bar{\bar{\beta}}_1$) with c = 5 and ϱ = .3 was not acceptable in any case.

Table 13

Precision of the Parameter Estimates of β_1 and β_2 (R500, c = 2, ϱ =.3)

	M	$S = 50$	$S = 150$	$S = 250$	$S = 350$
$\bar{\bar{\beta}}_1$(se)	C	1.0855(0.6316)	1.0356(0.3444)	0.9913(0.2643)	1.0086(0.2220)
	F	0.0010(0.3492)	-0.0090(0.1798)	-0.0108(0.1350)	-0.0117(0.1118)
	E	-1.4550(0.6200)	-1.4634(0.3557)	-1.4682(0.2766)	-1.4694(0.2330)
$\bar{\bar{\beta}}_2$(se)	C	-0.0084(0.2648)	0.0024(0.1452)	-0.0096(0.1104)	0.0007(0.0926)
	F	-0.0028(0.2401)	-0.0075(0.1235)	-0.0088(0.0925)	-0.0096(0.0764)
	E	-0.0048(0.2355)	0.0026(0.1365)	-0.0030(0.1063)	0.0036(0.0894)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3015	1.2877	1.2754	1.2802
	F	1.0003	0.9975	0.9970	0.9968
	E	0.5958	0.5935	0.5922	0.5918
% CS ($\hat{\beta}_1$)	C	7.4	0	0	0
	F	72.4	100	100	100
	E	99.6	100	100	100

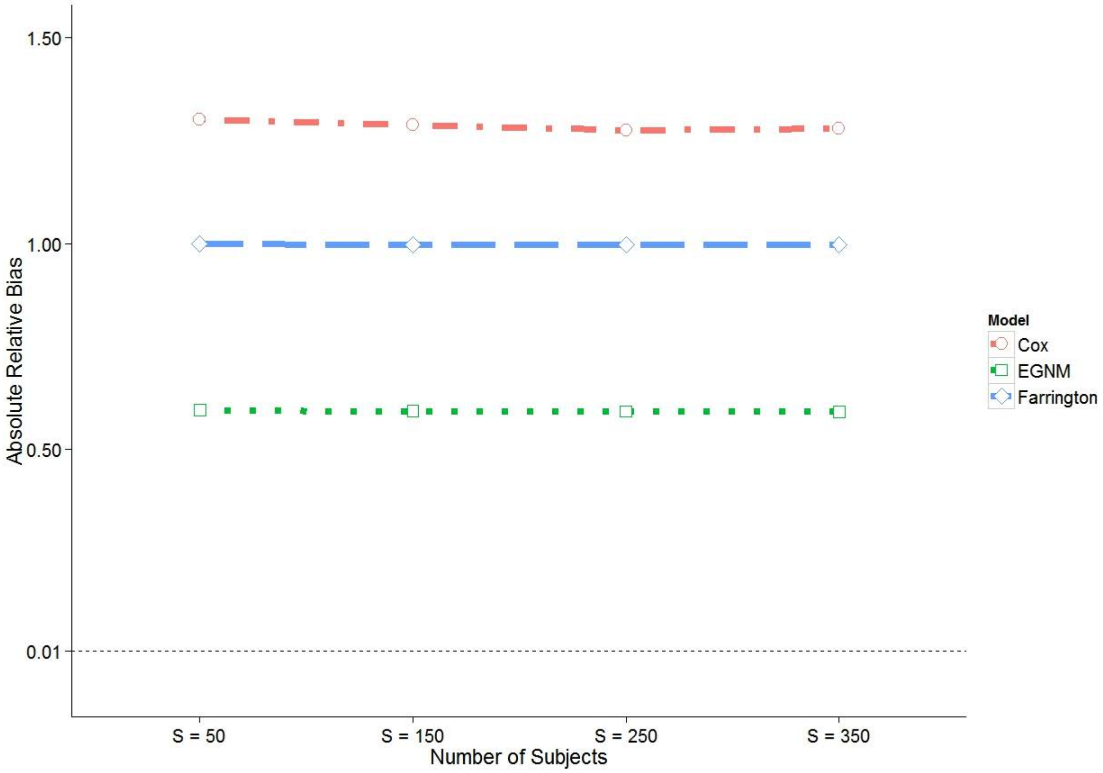

Figure 10. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R500, *c* = 2, ϱ = .3).

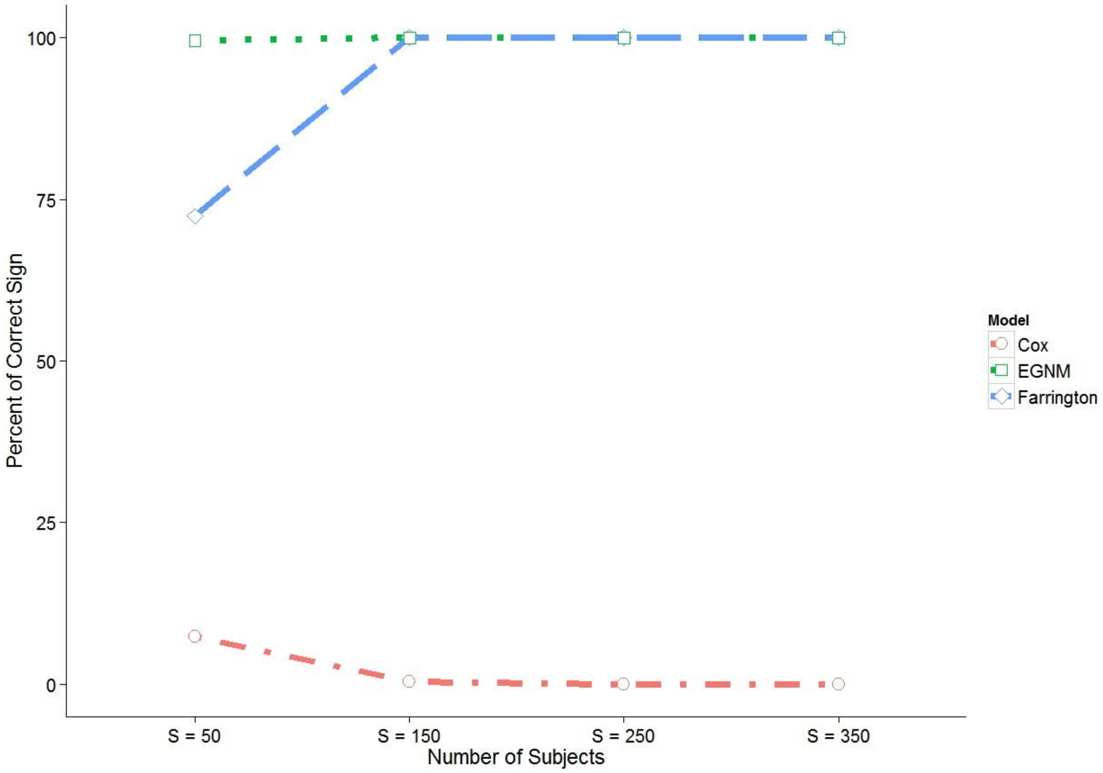

Figure 11. Percent of correct sign of the parameter estimate $\widehat{\beta}_1$ (R500, c = 2, ϱ = .3).

When c = 2 and ϱ = .3, the results from 500 replications were similar to those from 150 replications.

Table 14

Precision of the Parameter Estimates of β_1 and β_2 (R500, c = 5, ϱ =.3)

	M	S = 50	S = 150	S = 250	S = 350
$\bar{\bar{\beta}}_1$(se)	C	1.0017(0.6091)	0.9802(0.3348)	0.9574(0.2578)	0.9703(0.2166)
	F	0.0390(0.3652)	-0.0030(0.2072)	-0.0032(0.1544)	-0.0057(0.1255)
	E	-1.4388(0.6189)	-1.4504(0.3553)	-1.4682(0.2766)	-1.4742(0.2326)
$\bar{\bar{\beta}}_2$(se)	C	-0.0011(0.2605)	-0.0029(0.1437)	-0.0074(0.1100)	0.0031(0.0928)
	F	0.0125(0.2497)	0.0017(0.1406)	-0.0026(0.1056)	-0.0051(0.0857)
	E	-0.0060(0.2364)	0.0009(0.1367)	-0.0030(0.1063)	-0.0034(0.0898)
ARB ($\bar{\bar{\beta}}_1$)	C	1.2782	1.2723	1.2660	1.2695
	F	1.0108	1.0008	0.9991	0.9984
	E	0.6003	0.5971	0.5922	0.5905
% CS ($\hat{\beta}_1$)	C	6.8	0	0	0
	F	9.0	32.0	88.0	98.6
	E	99.6	100	100	100

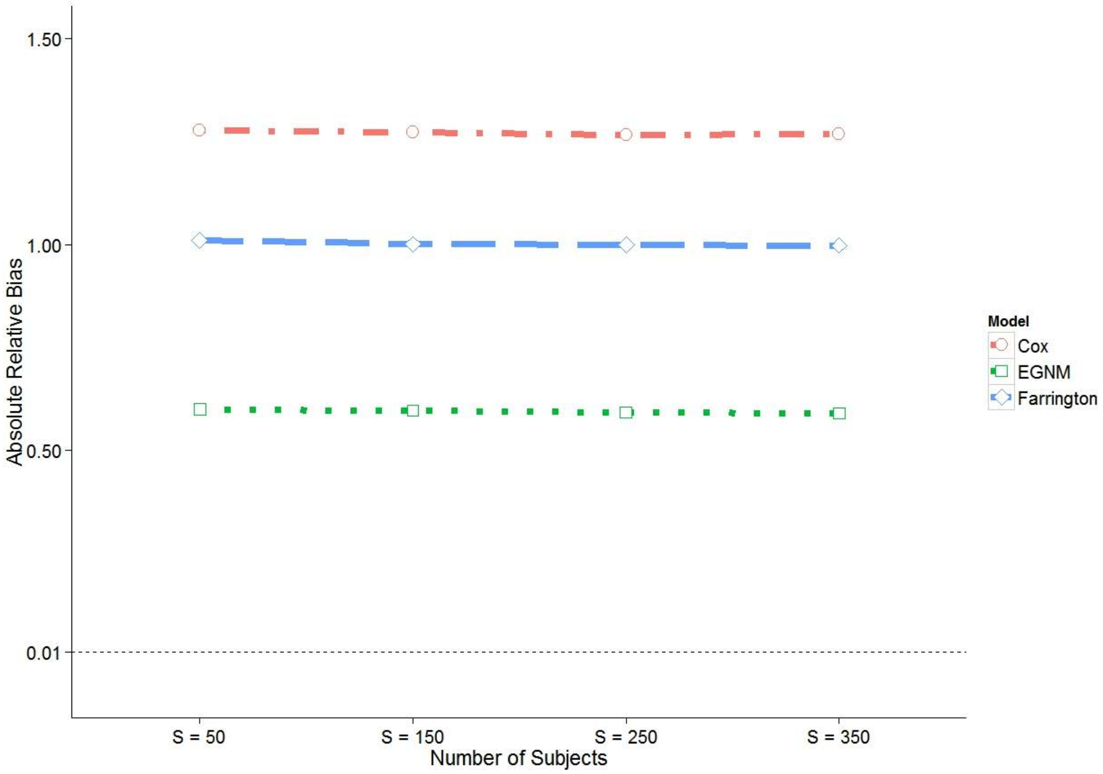

Figure 12. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R500, c = 5, ϱ = .3).

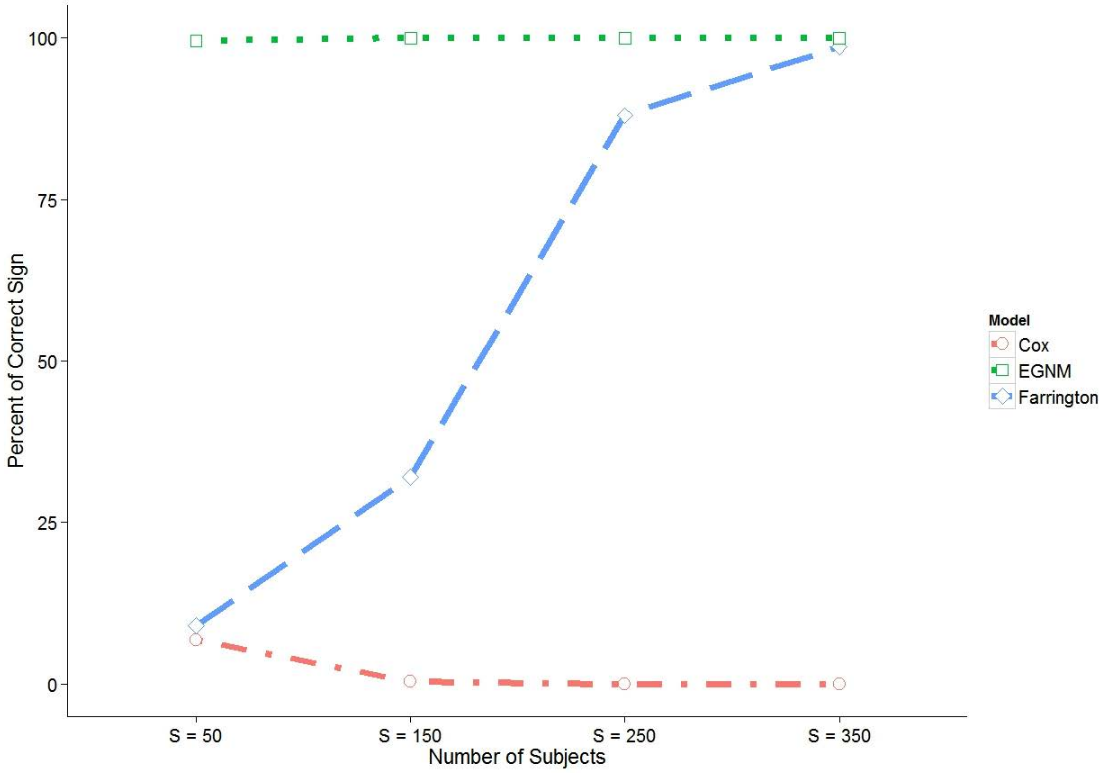

Figure 13. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R500, *c* = 5, *ϱ* = .3).

When the number of replications is 500, and *c* = 5 and *ϱ* = .3, the results were similar to those with 150 replications, *c* = 5 and *ϱ* = .3. It is worth mentioning that although the numbers of subjects used under 500 replications were 50, 150, 250, and 350, % CS($\hat{\beta}_1$) from Farrington's model was acceptable at the 80% level only when the number of subjects was at least 250.

Table 15

Precision of the Parameter Estimates of β₁ and β₂ (R500, c = 2, ϱ =.7)

	M	S = 50	S = 150	S = 250	S = 350
$\bar{\bar{\beta}}_1$(se)	C	1.4141(0.7936)	1.4226(0.4355)	1.3958(0.3333)	1.4020(0.2806)
	F	-0.0008(0.2971)	-0.0115(0.1540)	-0.0139(0.1158)	-0.0154(0.0966)
	E	-3.8364(0.7898)	-3.7318(0.4641)	-3.6928(0.3663)	-3.6758(0.3132)
$\bar{\bar{\beta}}_2$(se)	C	-0.0137(0.2754)	-0.0046(0.1499)	-0.0066(0.1148)	0.0118(0.0970)
	F	-0.0021(0.2412)	-0.0074(0.1238)	-0.0088(0.0927)	-0.0099(0.0771)
	E	-0.0070(0.2437)	0.0025(0.1411)	-0.0014(0.1099)	0.0052(0.0931)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3928	1.3952	1.3877	1.3895
	F	0.9998	0.9968	0.9961	0.9957
	E	0.0657	0.0366	0.0258	0.0211
% CS ($\hat{\beta}_1$)	C	7.2	0	0	0
	F	79.6	100	100	100
	E	100	100	100	100

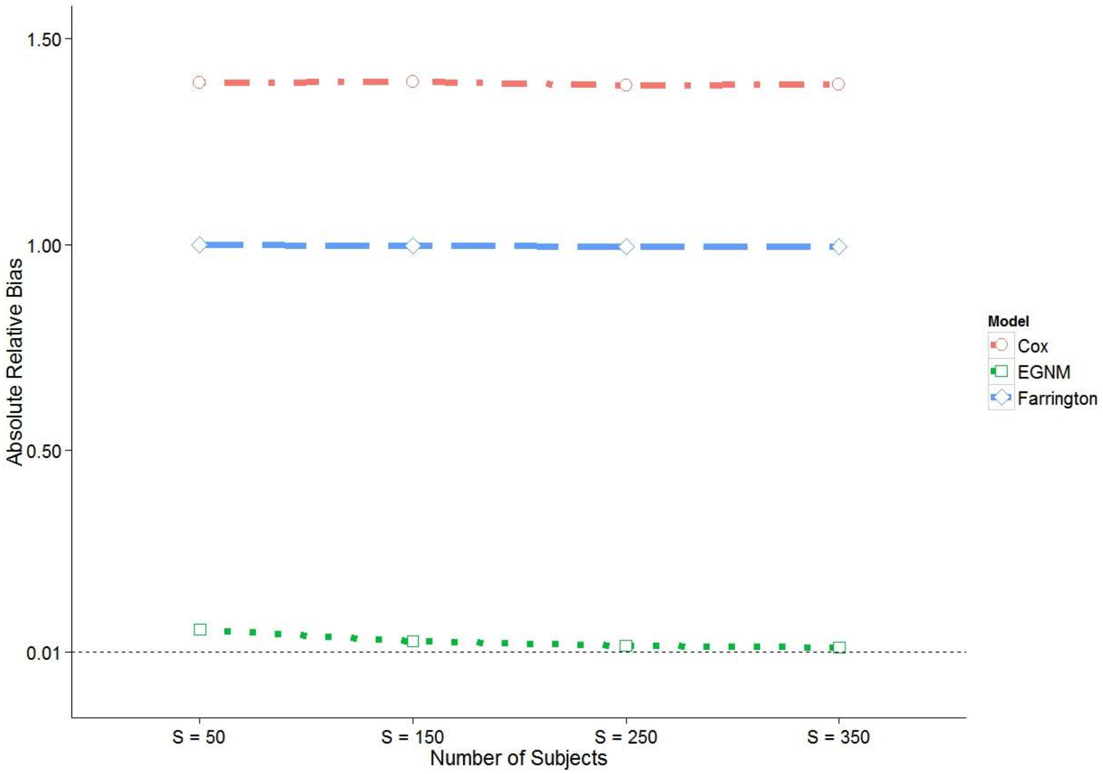

Figure 14. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R500, *c* = 2, ϱ = .7).

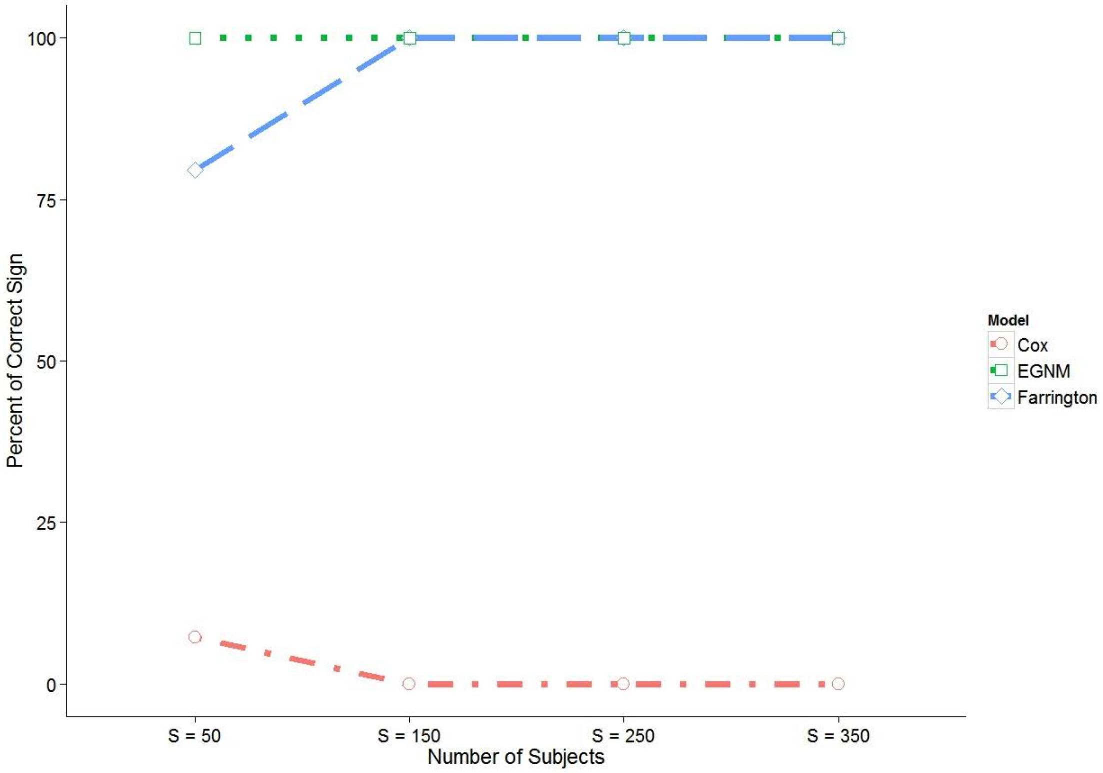

Figure 15. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R500, *c* = 2, *ϱ* = .7).

When the number of replications is 500, and *c* = 2 and *ϱ* = .7, the results were similar to those with 150 replications, *c* = 2 and *ϱ* = .7. It is worth mentioning that although ARB($\bar{\hat{\beta}}_1$) from the EGNM showed a decreasing trend as the number of subjects increased, and approached the acceptable level of .01, for example, 0.0211 from 350 subjects, due to the fact that the largest number of subjects used was 350, none of ARB($\bar{\hat{\beta}}_1$) was acceptable at the .01 level.

Table 16

Precision of the Parameter Estimates of β₁ and β₂ (R500, c = 5, ϱ =.7)

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
$\bar{\bar{\beta}}_1$(se)	C	1.4376(0.7665)	1.4141(0.4175)	1.3774(0.3193)	1.3943(0.2686)
	F	0.0391(0.3095)	0.0032(0.1769)	-0.0032(0.1321)	-0.0072(0.1076)
	E	-3.8364(0.7898)	-3.7318(0.4641)	-3.6928(0.3663)	-3.6764(0.3131)
$\bar{\bar{\beta}}_2$(se)	C	-0.0140(0.2719)	-0.0028(0.1484)	-0.0066(0.1136)	-0.0022(0.0956)
	F	0.0250(0.2509)	-0.0019(0.1420)	-0.0027(0.1057)	-0.0050(0.0861)
	E	0.0070(0.2437)	0.0025(0.1411)	-0.0014(0.1099)	-0.0038(0.0930)
ARB ($\bar{\bar{\beta}}_1$)	C	1.3993	1.3928	1.3826	1.3873
	F	1.0109	1.0009	0.9991	0.9980
	E	0.0657	0.0366	0.0258	0.0212
% CS ($\hat{\beta}_1$)	C	4.4	0	0	0
	F	6.2	32.6	86.2	98.8
	E	100	100	100	100

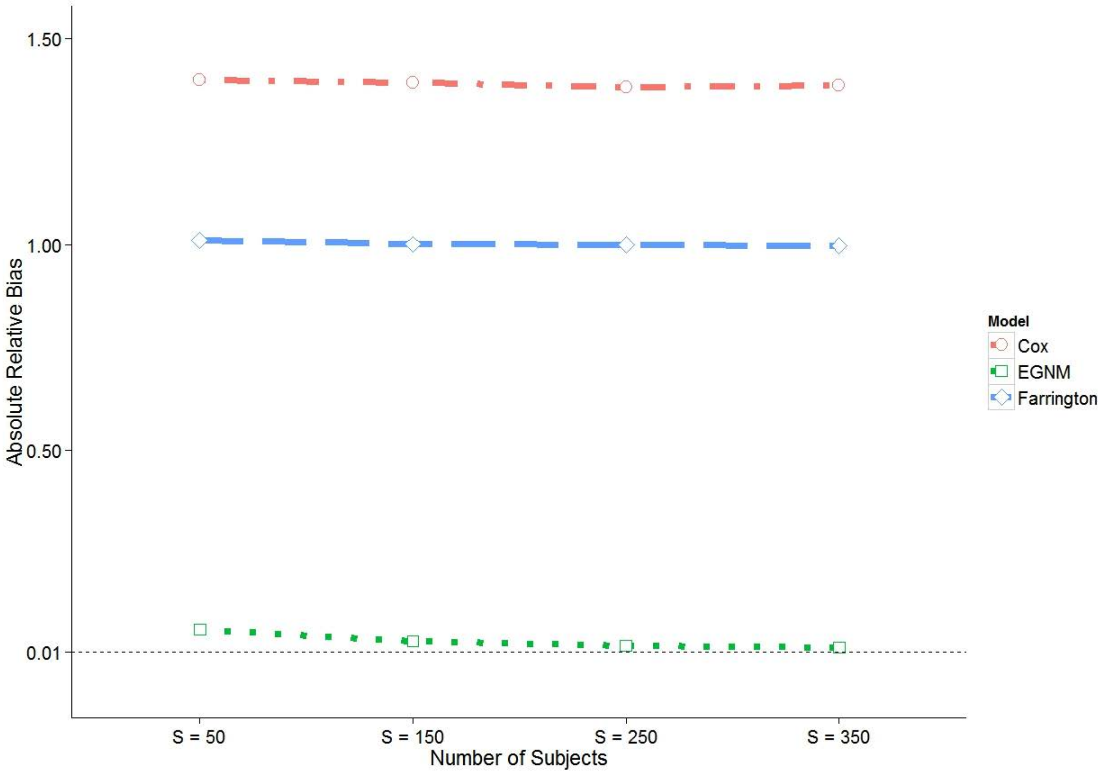

Figure 16. Absolute relative bias (ARB) of the mean parameter estimate $\bar{\bar{\beta}}_1$ (R500, *c* = 5, ϱ = .7).

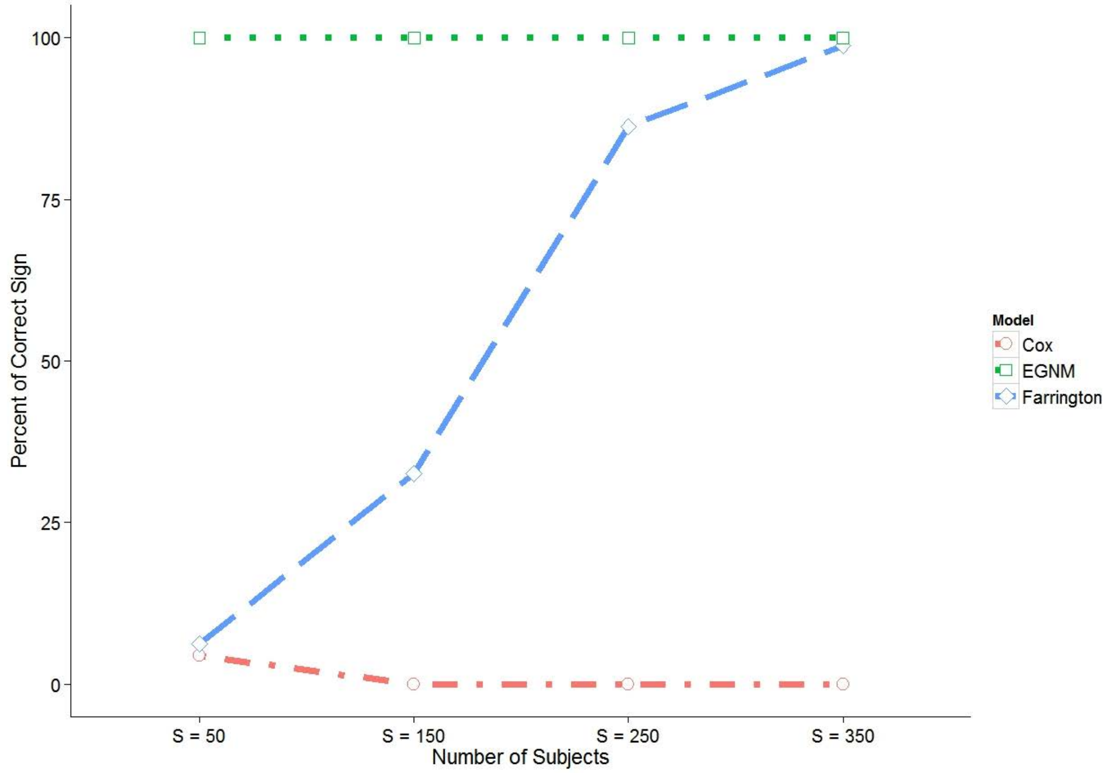

Figure 17. Percent of correct sign of the parameter estimate $\hat{\beta}_1$ (R500, c = 5, ϱ = .7).

When the number of replications is 500, and c = 5 and ϱ = .7, the results were similar to those with 150 replications, for c = 5 and ϱ = .7. Although the numbers of subjects used under 500 replications were 50, 150, 250, and 350, % CS($\hat{\beta}_1$) from Farrington's model was acceptable at the 80% level only when the number of subjects was at least 250, and although ARB($\bar{\bar{\beta}}_1$) from the EGNM showed a decreasing trend as the number of subjects increased, and approached the acceptable level of .01, for example, 0.0212 from 350 subjects, due to the fact that the largest number of subjects used was 350, none of the ARB($\bar{\bar{\beta}}_1$) was acceptable at the .01 level.

In summary, across all eight tables regarding precision of the parameter estimates of β_1 and β_2, the simulation results with the same c and ϱ values were similar.

When c = 2 or c = 5 with ϱ = .3, that is, the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is comparatively low, $\bar{\hat{\beta}}_1$ from any model was far from the true value of β_1, -3.6, but $\bar{\hat{\beta}}_2$ from any model was close to 0; ARB($\bar{\hat{\beta}}_1$) from any model was not acceptable at the .01 level; % CS($\hat{\beta}_1$) from the EGNM was acceptable at the 80% level, % CS($\hat{\beta}_1$) from Farrington's model was acceptable at the 80% level only when the number of subjects was greater than 50 when c = 2 or at least 250 when c = 5, but % CS($\hat{\beta}_1$) from the extended Cox model was not acceptable in any case.

When c = 2 or c = 5 with ϱ = .7, $\bar{\hat{\beta}}_1$ from the EGNM was very close to the true value -3.6, and $\bar{\hat{\beta}}_1$ from the other two models were still far from -3.6, but $\bar{\hat{\beta}}_2$ from any model was close to 0; ARB($\bar{\hat{\beta}}_1$) from the EGNM was acceptable overall, and was acceptable at the .01 level only with at least 500 subjects, and ARB($\bar{\hat{\beta}}_1$) from the other two models were still not acceptable; % CS($\hat{\beta}_1$) from the EGNM was acceptable at the 80% level, % CS($\hat{\beta}_1$) from Farrington's model was not acceptable at the 80% level only when the number of subjects was 50 when c = 5, but % CS($\hat{\beta}_1$) from the extended Cox model was not acceptable in any case. Next, to see whether the confidence intervals constructed for $\bar{\hat{\beta}}_1$ calculated from the three models include the true value of β_1, -3.6, which is the coefficient for the significant covariate X_{1t}, confidence intervals calculations for $\bar{\hat{\beta}}_1$ follow.

Table 17

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R150, c = 2, ϱ =.3)*

M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
	L[l] *U*[m]	*L* *U*	*L* *U*	*L* *U*
C	(-0.3735, 2.8427)	(0.3513, 1.5997)	(0.5518, 1.4984)	(0.6544, 1.3350)
F	(-0.0093, 0.0101)	(-0.0252, -0.0074)	(-0.0306, -0.0094)	(-0.0311, -0.0110)
E	(-2.5357, -0.4423)	(-1.9375, -1.0348)	(-1.7495, -1.1706)	(-1.6952, -1.2583)

[l] *L* refers to the 2.5th percentile of $\bar{\bar{\beta}}_1$, i.e., the lower limit of a confidence interval. [m] *U* refers to the 97.5th percentile of $\bar{\bar{\beta}}_1$, i.e., the upper limit of a confidence interval.

Table 18

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R150, c = 5, ϱ =.3)*

M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.1827, 2.2356)	(0.4617, 1.4962)	(0.5754, 1.3803)	(0.6809, 1.2449)
F	(-0.8066, 0.4520)	(-0.0074, 0.0031)	(-0.0107, -0.0044)	(-0.0128, -0.0080)
E	(-2.5357, -0.4423)	(-1.9299, -1.0311)	(-1.7495, -1.1706)	(-1.6952, -1.2583)

Table 19

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R150, c = 2, ϱ =.7)*

M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.4866, 3.2219)	(0.7353, 2.1808)	(0.8737, 2.0181)	(1.0030, 1.7545)
F	(-0.0109, 0.0053)	(-0.0348, -0.0098)	(-0.0386, -0.0124)	(-0.0396, -0.0140)
E	(-5.4764, -2.5143)	(-4.2692, -3.1614)	(-4.0347, -3.3293)	(-3.9137, -3.3739)

Table 20

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R150, c = 5, ϱ =.7)*

M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.2107, 3.1067)	(0.7425, 1.9840)	(0.9426, 1.8798)	(1.0432, 1.6902)
F	(-0.7676, 0.7950)	(-0.0093, 0.0042)	(-0.0136, -0.0055)	(-0.0164, -0.0102)
E	(-5.4764, -2.5143)	(-4.2692, -3.1614)	(-4.0347, -3.3293)	(-3.9137, -3.3739)

Table 21

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R500, c = 2, ϱ =.3)*

M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.4751, 2.6988)	(0.1953, 1.8425)	(0.3138, 1.6793)	(0.4900, 1.5630)
F	(-0.0079, 0.0266)	(-0.0120, -0.0057)	(-0.0257, -0.0074)	(-0.0279, -0.0085)
E	(-2.5411, -0.4254)	(-2.0789, -0.9164)	(-1.9197, -1.0230)	(-1.8286, -1.0448)

Table 22

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R500, c = 5, ϱ =.3)*

M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.3904, 2.1832)	(0.2834, 1.6427)	(0.4089, 1.5584)	(0.5099, 1.4155)
F	(-0.5056, 0.9095)	(-0.0047, 0.0283)	(-.0075, 0.0030)	(-0.0092, -0.0011)
E	(-2.5697, -0.4689)	(-2.1179, -0.8555)	(-1.9197, -1.0230)	(-1.8403, -1.0772)

Table 23

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R500, c = 2, ϱ =.7)*

M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.3247, 3.2242)	(0.3787, 2.4910)	(0.6787, 2.2182)	(0.8253, 2.0524)
F	(-0.0105, 0.0269)	(-0.0155, -0.0067)	(-0.0338,-0.0099)	(-0.0363, -0.0111)
E	(-5.5395, -2.6961)	(-4.5231, -3.0714)	(-4.2868, -3.1611)	(-4.1663, -3.2026)

Table 24

The Confidence Intervals for $\bar{\bar{\beta}}_1$ *(R500, c = 5, ϱ =.7)*

M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.1932, 3.0695)	(0.5634, 2.3755)	(0.7156, 2.0100)	(0.8769, 1.9915)
F	(-0.6759, 0.8329)	(-0.0052, 0.0305)	(-0.0092, 0.0041)	(-0.0116, -0.0009)
E	(-5.5394, -2.6961)	(-4.5231, -3.0714)	(-4.2868, -3.1611)	(-4.1675, -3.2049)

In summary, across all eight tables the confidence intervals constructed for $\bar{\bar{\beta}}_1$, the simulation results from 150 and 500 replications, with the same *c* and ϱ values, were similar. As the number of subjects increased, the confidence intervals became narrower. However, only the confidence intervals constructed for $\bar{\bar{\beta}}_1$ from the EGNM using $\varrho = .7$ contained the true value of β_1, -3.6, which is the coefficient for the significant covariate X_{1t}. Moreover, when the number of subjects was greater than 50, confidence intervals for $\bar{\bar{\beta}}_1$ across the three models were non-overlapping. Next, to see whether the confidence intervals constructed for $\bar{\bar{\beta}}_2$ calculated from the three models include the true value of β_2, 0, which is the coefficient for X_{2t}, confidence intervals calculations for $\bar{\bar{\beta}}_2$ follow.

Table 25

The Confidence Intervals for $\bar{\bar{\beta}}_2$ *(R150, c = 2,* ϱ *=.3)*

M	$S = 50$	$S = 250$	$S = 500$	$S = 1{,}000$
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.7120, 0.7247)	(-0.2568, 0.2601)	(-0.2201, 0.1992)	(-0.1191, 0.1297)
F	(-0.0094, 0.0120)	(-0.0181, -0.0054)	(-0.0252, -0.0068)	(-0.0268, -0.0086)
E	(-0.5238, 0.4638)	(-0.2477, 0.1982)	(-0.1482, 0.1642)	(-0.1115, 0.1155)

Table 26

The Confidence Intervals for $\bar{\bar{\beta}}_2$ *(R150, c = 5,* ϱ *=.3)*

M	$S = 50$	$S = 250$	$S = 500$	$S = 1{,}000$
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.4732, 0.6561)	(-0.3055, 0.2006)	(-0.1792, 0.1645)	(-0.1071, 0.1237)
F	(-0.2480, 0.1982)	(-0.0065, 0.0028)	(-0.0093, -0.0035)	(-0.0109, -0.0063)
E	(-0.5238, 0.4638)	(-0.2320, 0.2353)	(-0.1482, 0.1642)	(-0.1115, 0.1155)

Table 27

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R150, c = 2, ϱ =.7)

M	S = 50	S = 250	S = 500	S = 1,000
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.5486, 0.5697)	(-0.2474, 0.2253)	(-0.1702, 0.1418)	(-0.1232, 0.1042)
F	(-0.0100, 0.0063)	(-0.0224, 0.0050)	(-0.0254, -0.0068)	(-0.0267, -0.0085)
E	(-0.5090, 0.4434)	(-0.2632, 0.2546)	(-0.1690, 0.1822)	(-0.1134, 0.1099)

Table 28

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R150, c = 5, ϱ =.7)

M	S = 50	S = 250	S = 500	S = 1,000
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.5280, 0.5252)	(-0.2347, 0.1887)	(-0.1753, 0.1360)	(-0.1188, 0.0988)
F	(-0.2400, 0.2967)	(-0.0065, 0.0028)	(-0.0093, -0.0035)	(-0.0109, -0.0063)
E	(-0.5090, 0.4434)	(-0.2632, 0.2546)	(-0.1690, 0.1822)	(-0.1134, 0.1098)

Table 29

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R500, c = 2, ϱ =.3)

M	S = 50	S = 150	S = 250	S = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.6752, 0.6346)	(-0.3396, 0.3542)	(-0.3158, 0.2594)	(-0.2290, 0.2197)
F	(-0.0100, 0.0237)	(-0.0120, -0.0028)	(-0.0205, -0.0050)	(-0.0235, -0.0060)
E	(-0.4534, 0.4468)	(-0.2679, 0.2900)	(-0.2331, 0.2294)	(-0.1653, 0.1960)

Table 30

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R500, c = 5, ϱ =.3)

M	S = 50	S = 150	S = 250	S = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.6185, 0.5858)	(-0.2858, 0.3221)	(-0.2798, 0.2339)	(-0.2043, 0.1935)
F	(-0.2532, 0.2832)	(-0.0048, 0.0254)	(-0.0074, 0.0027)	(-0.0086, -0.0010)
E	(-0.4607, 0.5069)	(-0.2514, 0.2850)	(-0.2331, 0.2294)	(-0.1649, 0.1716)

Table 31

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R500, c = 2, ϱ =.7)

M	S = 50	S = 150	S = 250	S = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.6148, 0.6178)	(-0.3250, 0.3203)	(-0.2391, 0.2449)	(-0.1750, 0.2178)
F	(-0.0096, 0.0220)	(-0.0119, -0.0027)	(-0.0216, -0.0049)	(-0.0249, -0.0059)
E	(-0.5304, 0.4989)	(-0.3006, 0.3184)	(-0.2256, 0.2422)	(-0.1762, 0.1905)

Table 32

The Confidence Intervals for $\bar{\bar{\beta}}_2$ (R500, c = 5, ϱ =.7)

M	S = 50	S = 150	S = 250	S = 350
	L *U*	*L* *U*	*L* *U*	*L* *U*
C	(-0.6042, 0.5801)	(-0.2993, 0.3039)	(-0.2345, 0.2163)	(-0.1991, 0.1977)
F	(-0.2276, 0.3321)	(-0.0048, 0.0226)	(-0.0074, 0.0025)	(-0.0086, -0.0008)
E	(-0.5304, 0.4989)	(-0.3006, 0.3184)	(-0.2256, 0.2422)	(-0.1998, 0.1878)

In summary, across all eight tables the confidence intervals constructed for $\bar{\bar{\beta}}_2$, the simulation results from 150 and 500 replications, with the same *c* and ϱ values, were similar. As the number of subjects increased, the confidence intervals became narrower. However, the confidence intervals constructed for $\bar{\bar{\beta}}_2$ from the extended Cox model and the EGNM included the true value of β_2, 0, which is the coefficient for X_{2t}, in all circumstances. The confidence intervals from Farrington's model sometimes did not include the true value of β_2, especially when the number of subjects was the largest with either 150 or 500 replications. Next, results of precision of the parameter estimate $\bar{\bar{\beta}}_0$ follow.

Table 33

Precision of the Parameter Estimate of β_0 (R150, c = 2, ϱ =.3)

	M	S = 50	S = 250	S = 500	S = 1,000
$\bar{\bar{\beta}}_0$(se)	F	0.0936(0.1569)	0.0701(0.0614)	0.0617(0.0430)	0.0495(0.0313)
	E	-0.9045(0.1735)	-0.8933(0.0756)	-0.8861(0.0532)	-0.8807(0.0373)
ARB	F	-0.9376	-0.9533	-0.9589	-0.9670
($\bar{\bar{\beta}}_0$)	E	-1.6030	-1.5956	-1.5907	-1.5871
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	0	0	0	0

Note. The true value of β_0 is 1.5.

Table 34

Precision of the Parameter Estimate of β_0 (R150, c = 5, ϱ =.3)

	M	S = 50	S = 250	S = 500	S = 1,000
$\bar{\bar{\beta}}_0$(se)	F	0.3643(0.1647)	0.0926(0.0703)	.0788(0.0461)	0.0711(0.0307)
	E	-0.9045(0.1735)	-0.8913(0.0754)	-.8861(0.0532)	-0.8807(0.0373)
ARB	F	-0.7571	-0.9383	-0.9475	-0.9526
($\bar{\bar{\beta}}_0$)	E	-1.6030	-1.5942	-1.5907	-1.5871
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	0	0	0	0

Table 35

Precision of the Parameter Estimate of β_0 (R150, c = 2, ϱ =.7)

	M	S = 50	S = 250	S = 500	S = 1,000
$\bar{\bar{\beta}}_0$(se)	F	0.1028(0.1569)	0.0699 (0.0616)	0.0618 (0.0430)	0.0488 (0.0315)
	E	-0.4259(0.1794)	-0.4451(0.0814)	-0.4436(0.0579)	-0.4461(0.0410)
ARB	F	-0.9315	-0.9534	-.9588	-.9674
($\bar{\bar{\beta}}_0$)	E	-1.2839	-1.2967	-1.2957	-1.2974
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	1.3	0	0	0

Table 36

Precision of the Parameter Estimate of β_0 *(R150, c = 5,* ϱ *=.7)*

	M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
$\bar{\bar{\beta}}_0$(se)	F	0.3556(0.1643)	0.0926(0.0703)	0.0788(0.0461)	0.0712(0.0308)
	E	-0.4259(0.1794)	-0.4451(0.0814)	-0.4436(0.0579)	-0.4461(0.0410)
ARB	F	-0.7629	-0.9383	-0.9475	-0.9526
($\bar{\bar{\beta}}_0$)	E	-1.2839	-1.2967	-1.2957	-1.2974
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	1.3	0	0	0

Table 37

Precision of the Parameter Estimate of β_0 *(R500, c = 2,* ϱ *=.3)*

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
$\bar{\bar{\beta}}_0$(se)	F	0.1108(0.1576)	0.0753(0.0819)	0.0701(0.0615)	0.0676(0.0510)
	E	-0.8942(0.1702)	-0.8860(0.0971)	-0.8818(0.0751)	-0.8820(0.0635)
ARB	F	-0.9261	-0.9498	-0.9532	-0.9549
($\bar{\bar{\beta}}_0$)	E	-1.5961	-1.5907	-1.5879	-1.5880
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	0	0	0	0

Table 38

Precision of the Parameter Estimate of β_0 *(R500, c = 5,* ϱ *=.3)*

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
$\bar{\bar{\beta}}_0$(se)	F	0.3793(0.1646)	0.1177(0.0941)	0.0947(0.0703)	0.0849(0.0574)
	E	-0.8977(0.1697)	-0.8861(0.0974)	-0.8818(0.0751)	-0.8786(0.0633)
ARB	F	-0.7471	-0.9215	-0.9369	-0.9434
($\bar{\bar{\beta}}_0$)	E	-1.5985	-1.5907	-1.5879	-1.5857
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	0	0	0	0

Table 39

Precision of the Parameter Estimate of β_0 (R500, c = 2, ϱ =.7)

	M	$S = 50$	$S = 150$	$S = 250$	$S = 350$
$\bar{\bar{\beta}}_0$(se)	F	0.1078(0.1571)	0.0752(0.0820)	0.0700(0.0616)	0.0668(0.0515)
	E	-0.4217(0.1774)	-0.4347(0.1036)	-0.4412(0.0814)	-0.4467(0.0699)
ARB	F	-0.9282	-0.9498	-0.9533	-0.9555
($\bar{\bar{\beta}}_0$)	E	-1.2811	-1.2898	-1.2941	-1.2978
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	1.0	0	0	0

Table 40

Precision of the Parameter Estimate of β_0 (R500, c = 5, ϱ =.7)

	M	$S = 50$	$S = 150$	$S = 250$	$S = 350$
$\bar{\bar{\beta}}_0$(se)	F	0.3624(0.1640)	0.1053(0.0942)	0.0945(0.0703)	0.0849(0.0573)
	E	-0.4217(0.1774)	-0.4347(0.1036)	-0.4412(0.0814)	-0.4436(0.0696)
ARB	F	-0.7584	-0.9298	-0.9370	-0.9434
($\bar{\bar{\beta}}_0$)	E	-1.2811	-1.2898	-1.2941	-1.2957
% CS	F	100	100	100	100
($\hat{\beta}_0$)	E	1.0	0	0	0

In summary, across all eight tables, precision of the parameter estimates of β_0, across two sets of numbers of subjects and replications, with the same *c* and ϱ values, produced similar results. However, the results were very poor. $\bar{\bar{\beta}}_0$ from the two models were far from the true value of β_0, 1.5, and substantially underestimated; ARB($\bar{\bar{\beta}}_0$) from either model was not acceptable at the .01 level; % CS($\hat{\beta}_0$) from the EGNM was not acceptable in any case, and % CS($\hat{\beta}_0$) from Farrington's model was acceptable at the 80% level.

The reason why the extended Cox model does not estimate an intercept is the parameter is unidentifiable, as the exponentiated intercept term is subsumed by the unknown baseline hazard function, thus any intercept term would simply change the baseline hazard function. As such, only results of precision of the parameter estimate $\bar{\bar{\beta}}_0$ from Farrington's model and the EGNM were included.

Hypothesis Testing of the Parameter Estimates

Regarding each model's capability of detecting the significance of covariates, the results of power analysis and analysis of type I error rate from the three models are presented in Table 41-Table 48. To visually check the simulation results, two figures were created for each table with each combination of the conditions to display the power curves and type I error rate, respectively. The results regarding hypothesis testing of the parameter estimates from 150 replications are shown first, followed by the results from 500 replications.

Table 41

Power and Type I Error Rate for the Three Models (R150, c = 2, ϱ =.3)

	M	S = 50	S = 250	S = 500	S = 1,000
Power	C	0.427	0.933	1.000	1.000
	F	0	0	0	0
	E	0.620	1.000	1.000	1.000
Type I Error Rate	C	0.140	0.147	0.133	0.133
	F	0	0	0	0
	E	0.067	0.053	0.060	0.073

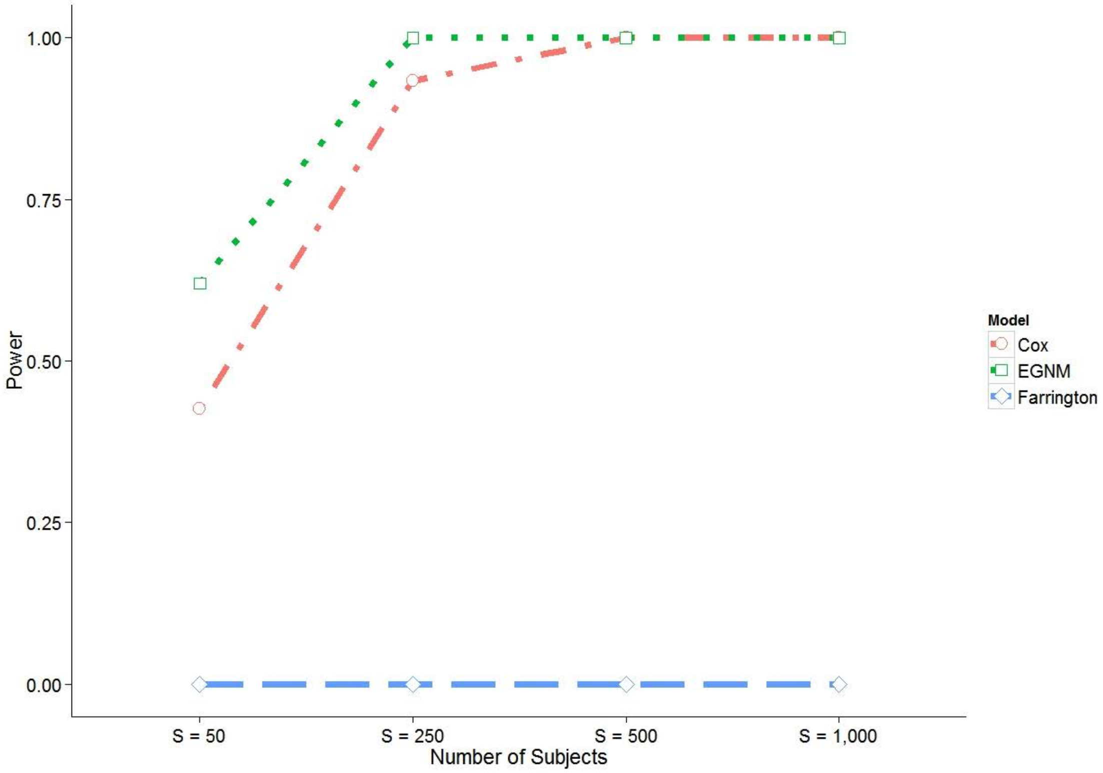

Figure 18. Power curves of the three models (R150, *c* = 2, *ϱ* = .3).

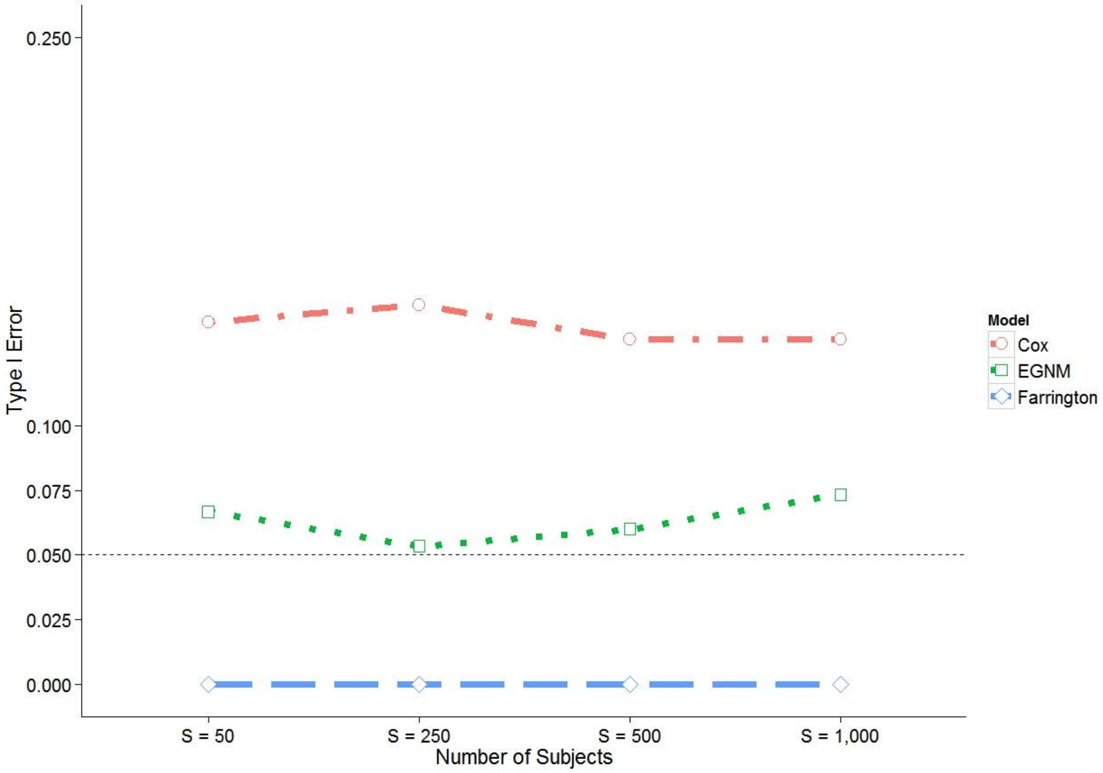

Figure 19. Type I error rates of the three models (R150, *c* = 2, *ϱ* = .3).

When the number of replications is 150, and *c* = 2 and *ϱ* = .3, that is, the widths of simulated censoring intervals are comparatively narrow, and the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is comparatively low, power from the extended Cox model and the EGNM were acceptable at the .90 level when the number of subjects was at least 250, and power from Farrington’s model was not acceptable in any case; the EGNM controlled type I error rate better than the extended Cox model, and type I error rate from Farrington’s model was 0.

Table 42

Power and Type I Error Rate for the Three Models (R150, c = 5, ϱ =.3)

	M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
Power	C	0.453	0.967	1.000	1.000
	F	0.060	0	0	0
	E	0.620	1.000	1.000	1.000
Type I Error Rate	C	0.060	0.093	0.093	0.053
	F	0.013	0	0	0
	E	0.067	0.060	0.060	0.073

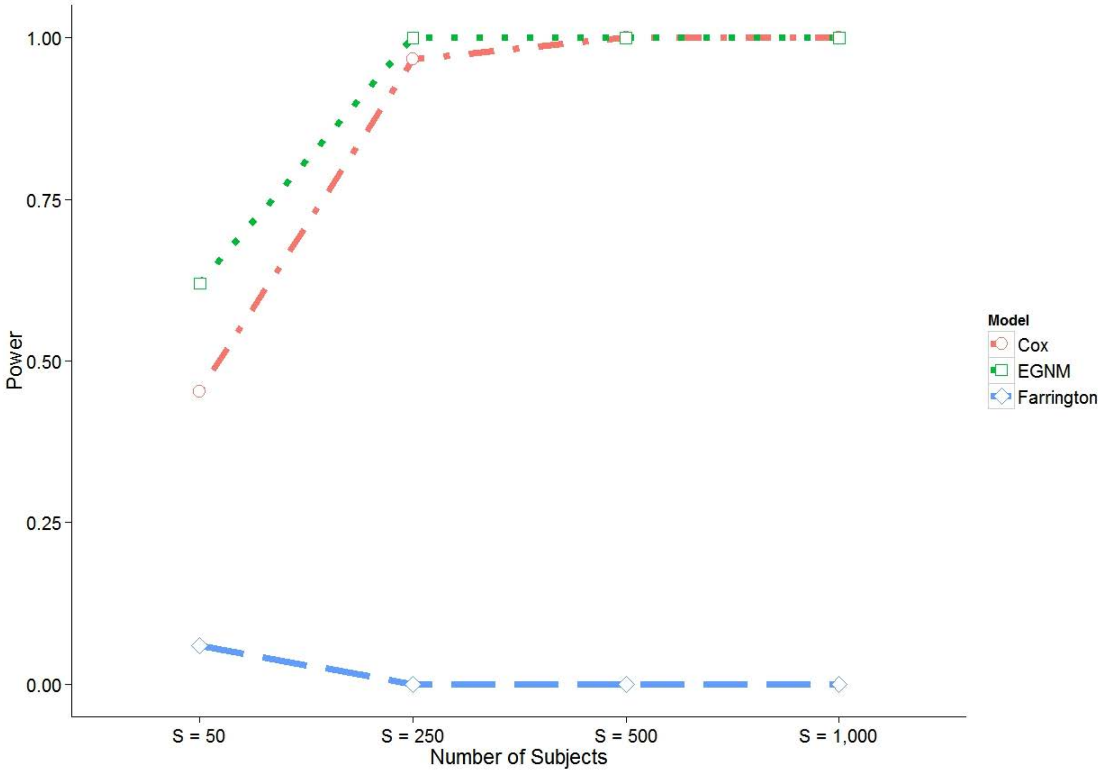

Figure 20. Power curves of the three models (R150, *c* = 5, ϱ = .3).

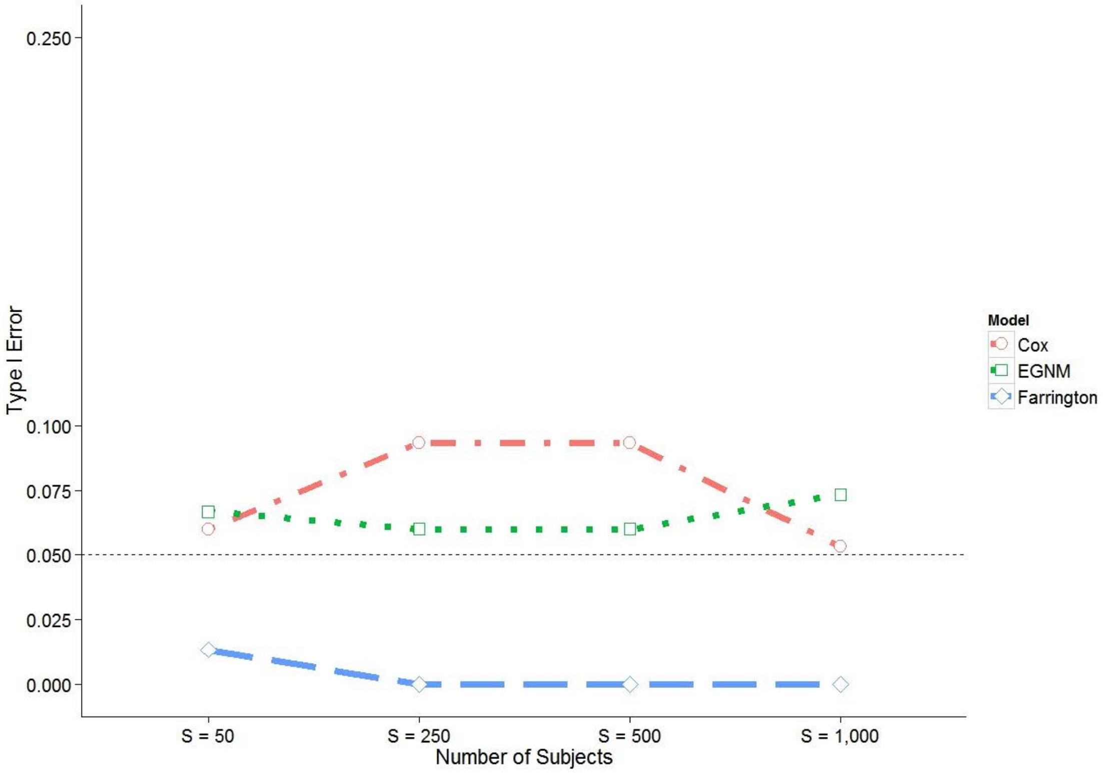

Figure 21. Type I error rates of the three models (R150, *c* = 5, *ϱ* = .3).

When the number of replications is 150, and *c* = 5 and *ϱ* = .3, that is, compared to *c* = 2 and *ϱ* = .3, the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is still comparatively low, but simulated censoring intervals are lengthened, the power from the three models was similar to that with *c* = 2 and *ϱ* = .3. Type I error rate from the EGNM was similar to that with *c* = 2 and *ϱ* = .3, and type I error rate from Farrington's model was almost 0. The EGNM controlled type I error rate slightly better than the extended Cox model when *c* = 2 and *ϱ* = .3.

Table 43

Power and Type I Error Rate for the Three Models (R150, c = 2, ϱ =.7)

	M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
Power	C	0.507	0.980	1.000	1.000
	F	0.007	0	0	0
	E	1.000	1.000	1.000	1.000
Type I Error Rate	C	0.067	0.067	0.067	0.060
	F	0	0	0	0
	E	0.073	0.100	0.093	0.067

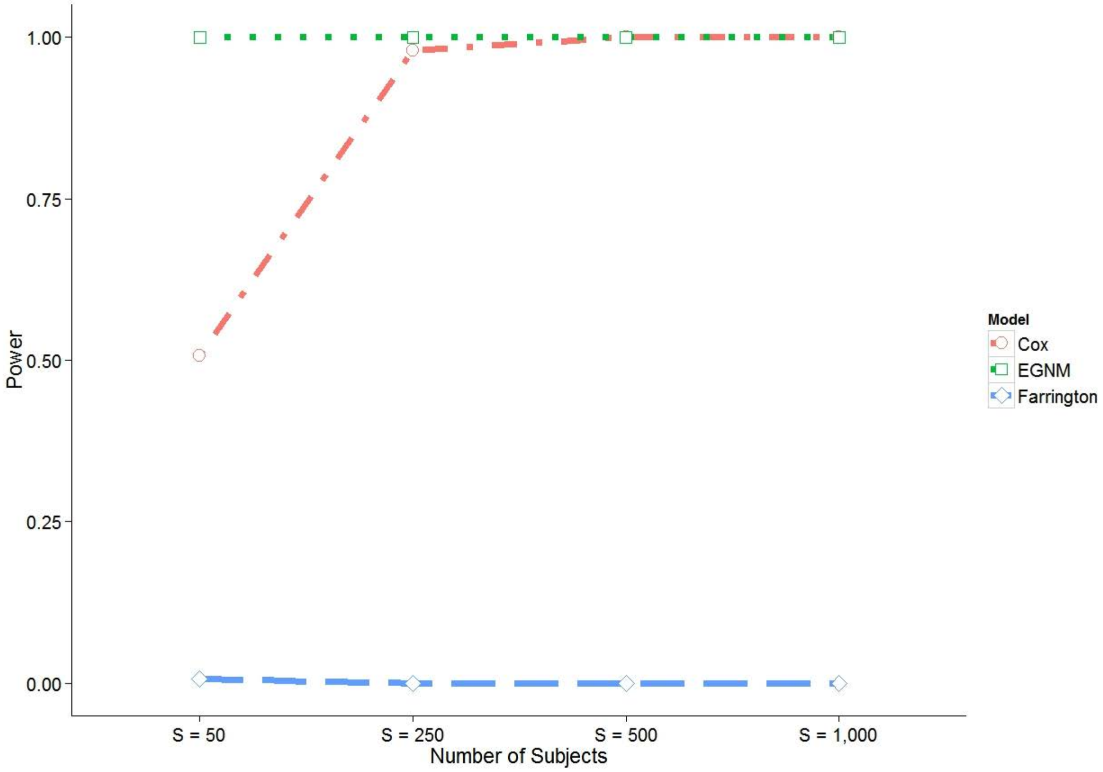

Figure 22. Power curves of the three models (R150, *c* = 2, ϱ = .7).

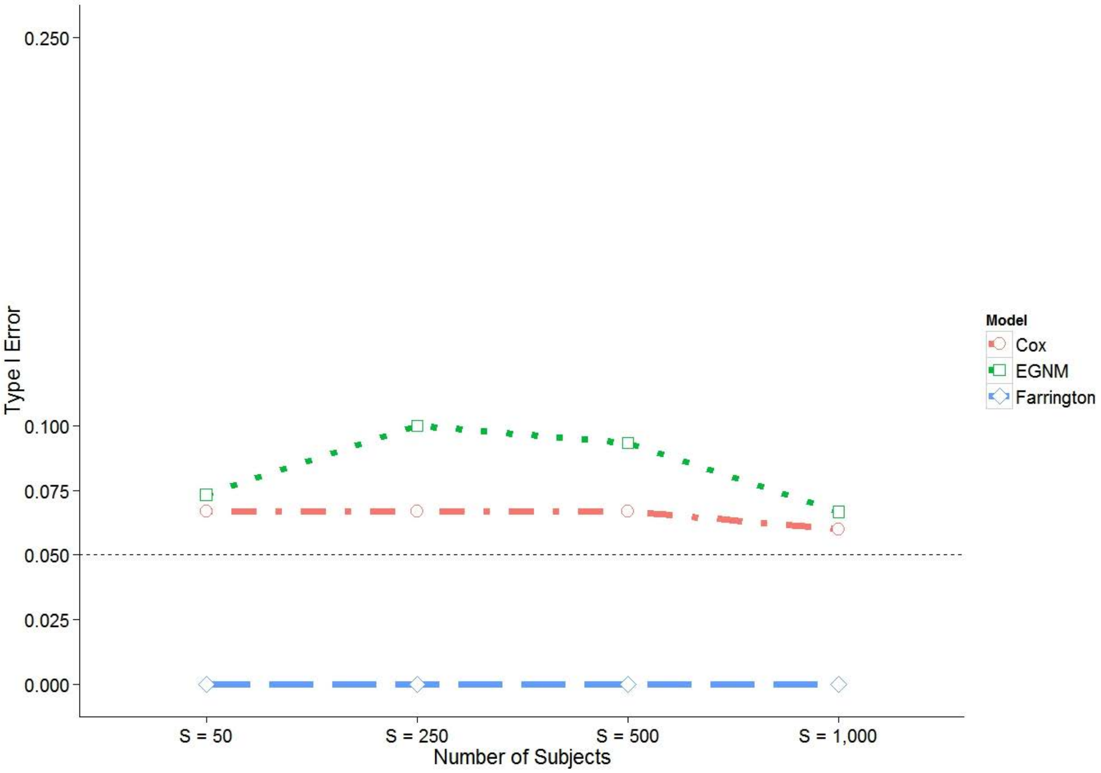

Figure 23. Type I error rates of the three models (R150, *c* = 2, *ϱ* = .7).

When the number of replications is 150, and *c* = 2 and *ϱ* = .7, that is, compared to *c* = 2 and *ϱ* = .3, the widths of simulated censoring intervals are still comparatively narrow, but the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, become high, power from the EGNM was acceptable at the .90 level, and power from the extended Cox model was acceptable when the number of subjects was at least 250, but Farrington's model did not have any power. The extended Cox model controlled type I error rate better than the EGNM, where type I error rate was slightly inflated, and type I error rate from Farrington's model was 0.

Table 44

Power and Type I Error Rate for the Three Models (R150, c = 5, ϱ =.7)

	M	*S* = 50	*S* = 250	*S* = 500	*S* = 1,000
Power	C	0.520	0.987	1.000	1.000
	F	0.087	0	0	0
	E	1.000	1.000	1.000	1.000
Type I Error Rate	C	0.053	0.060	0.047	0.060
	F	0.020	0.006	0	0
	E	0.073	0.100	0.093	0.067

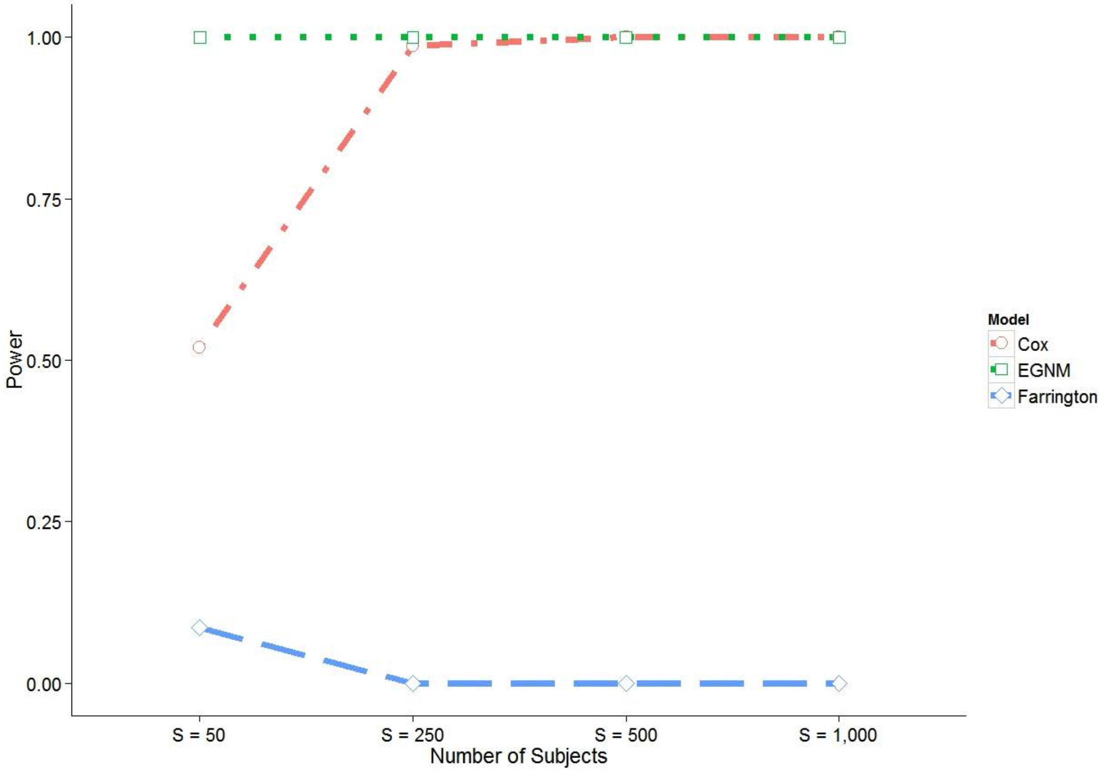

Figure 24. Power curves of the three models (R150, *c* = 5, ϱ = .7).

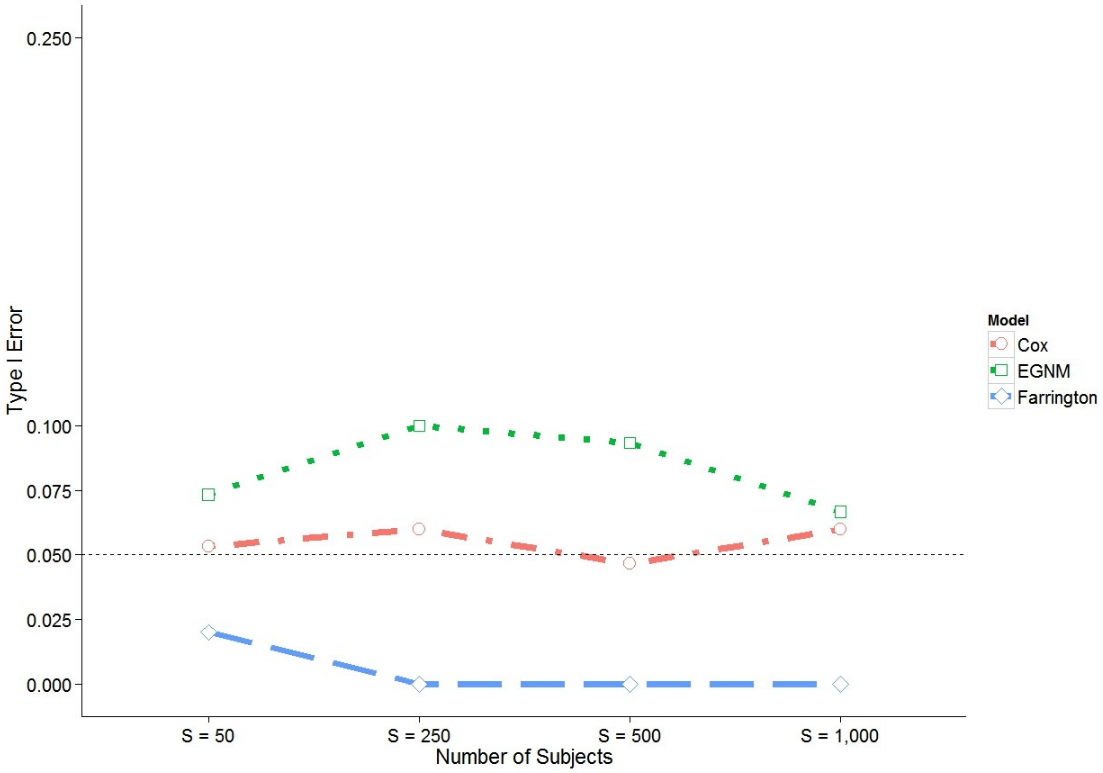

Figure 25. Type I error rates of the three models (R150, *c* = 5, ϱ = .7).

When the number of replications is 150, and *c* = 5 and ϱ = .7, that is, compared to *c* = 2 and ϱ = .7, the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is still comparatively high, but simulated censoring intervals are lengthened, power and type I error rate from the three models was similar to that with *c* = 2 and ϱ = .7. The only difference lay in that the extended Cox model controlled type I error rate was slightly better than when *c* = 2 and ϱ = .7.

Compared to *c* = 5 and ϱ = .3, that is, simulated censoring intervals are still comparatively wide, but the probability of the descending association between X_{1t} and the response probability, or the hazards associated with an occurrence case, is low, the

EGNM became more powerful with 50 subjects than when $c = 5$ and $\varrho = .3$, power from the other two models was similar to that with $c = 5$ and $\varrho = .3$. The extended Cox model controlled type I error rate slightly better than when $c = 5$ and $\varrho = .3$, and type I error rate from the other two models was similar to that with $c = 5$ and $\varrho = .3$.

Table 45

Power and Type I Error Rate for the Three Models (R500, $c = 2$, ϱ =.3)

	M	$S = 50$	$S = 150$	$S = 250$	$S = 350$
Power	C	0.428	0.788	0.920	0.986
	F	0.002	0	0	0
	E	0.676	0.996	1.000	1.000
Type I Error Rate	C	0.110	0.106	0.142	0.114
	F	0.002	0	0	0
	E	0.046	0.062	0.076	0.050

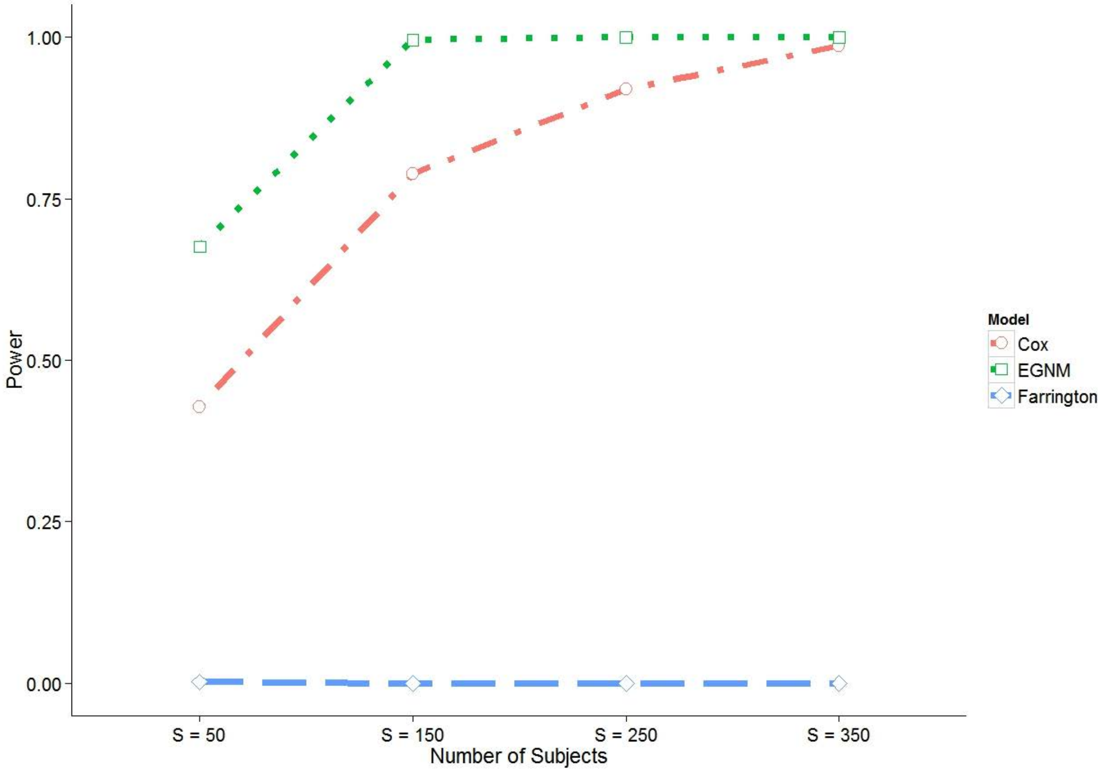

Figure 26. Power curves of the three models (R500, $c = 2$, $\varrho = .3$).

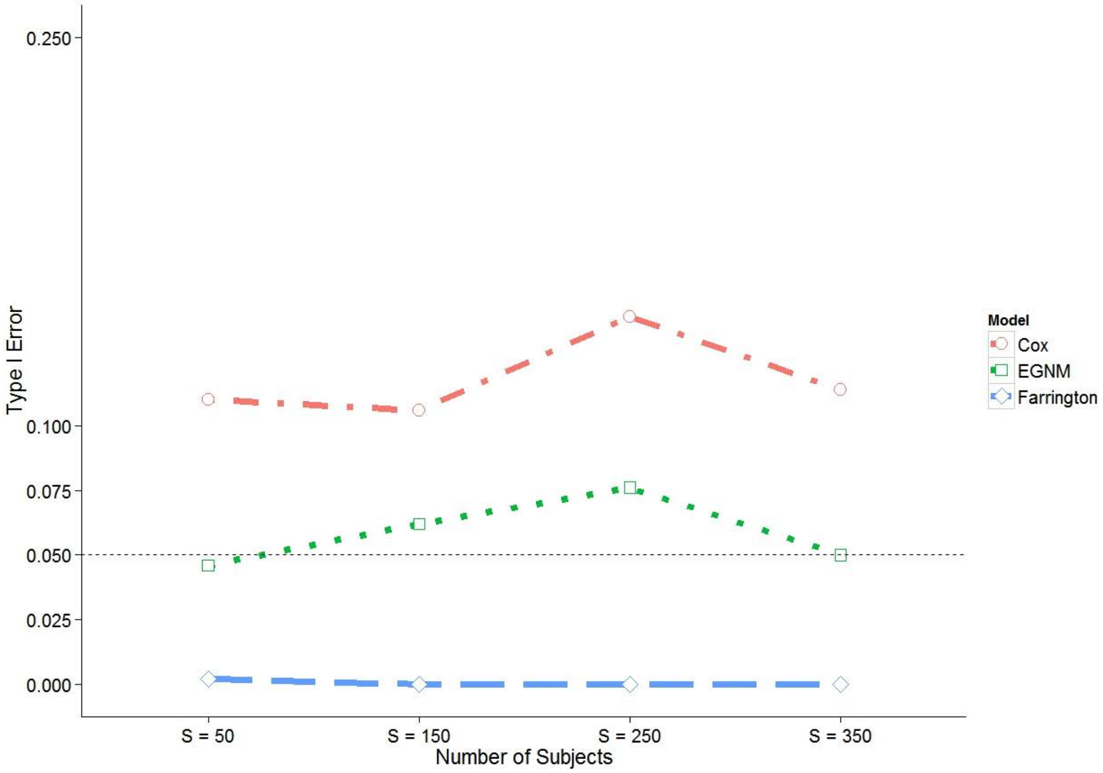

Figure 27. Type I error rates of the three models (R500, *c* = 2, *ϱ* = .3).

When *c* = 2 and *ϱ* = .3, the results regarding hypothesis testing of the parameter estimates from 500 replications were similar to those from 150 replications.

Table 46

Power and Type I Error Rate for the Three Models (R500, c = 5, ϱ =.3)

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
Power	C	0.402	0.814	0.946	0.988
	F	0.058	0.004	0	0
	E	0.654	0.988	1.000	1.000
Type I Error Rate	C	0.086	0.070	0.094	0.068
	F	0.018	0.002	0	0
	E	0.068	0.056	0.076	0.038

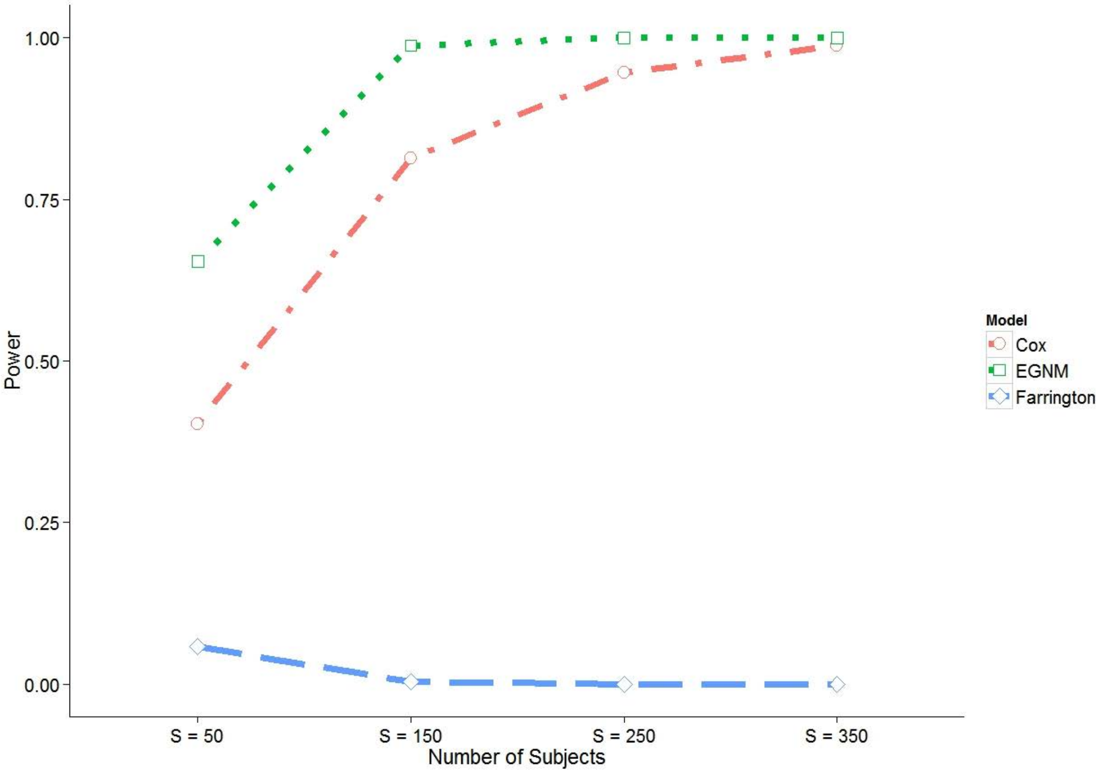

Figure 28. Power curves of the three models (R500, *c* = 5, *ϱ* = .3).

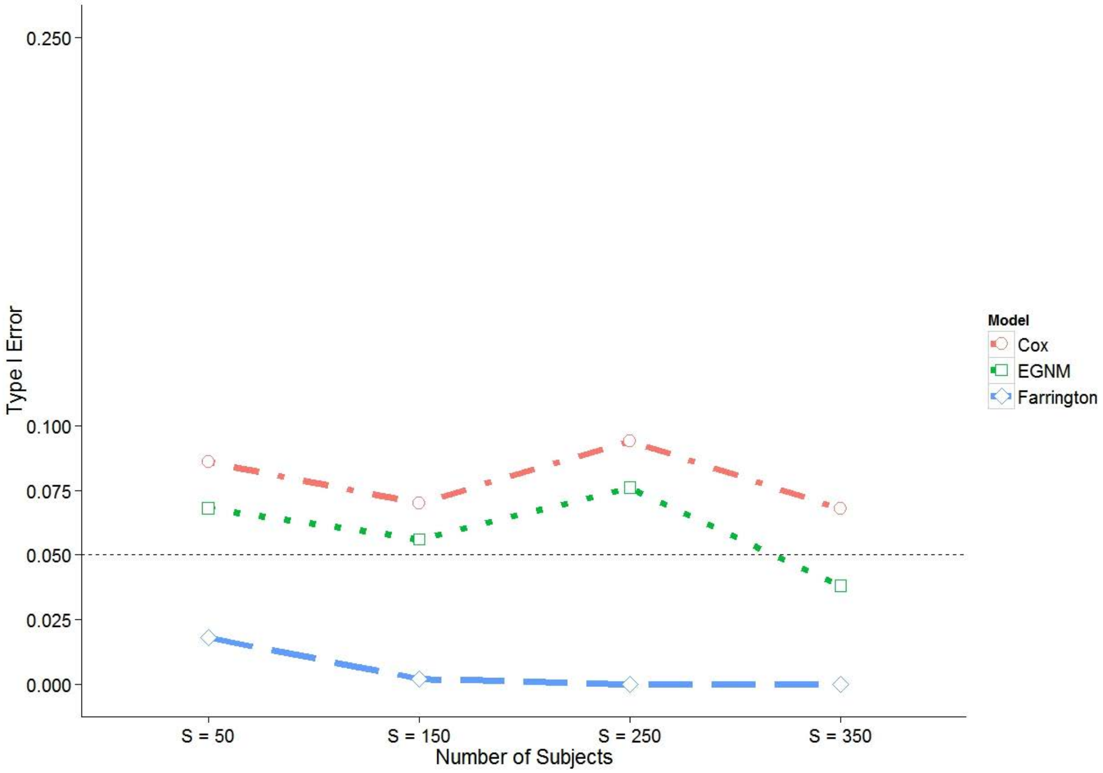

Figure 29. Type I error rates of the three models (R500, *c* = 5, *ϱ* = .3).

When the number of replications is 500, and *c* = 5 and *ϱ* = .3, the power from the EGNM and the extended Cox model was acceptable at the .90 level only when the number of subjects was at least 150 and 250, respectively, and power from Farrington's model was not acceptable in any case. Overall, the EGNM controlled type I error rate slightly better than the extended Cox model, and type I error rate from Farrington's model was not acceptable.

Table 47

Power and Type I Error Rate for the Three Models (R500, c = 2, ϱ =.7)

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
Power	C	0.468	0.874	0.978	0.998
	F	0.006	0	0	0
	E	1.000	1.000	1.000	1.000
Type I Error Rate	C	0.078	0.070	0.074	0.060
	F	0	0	0	0
	E	0.076	0.086	0.072	0.058

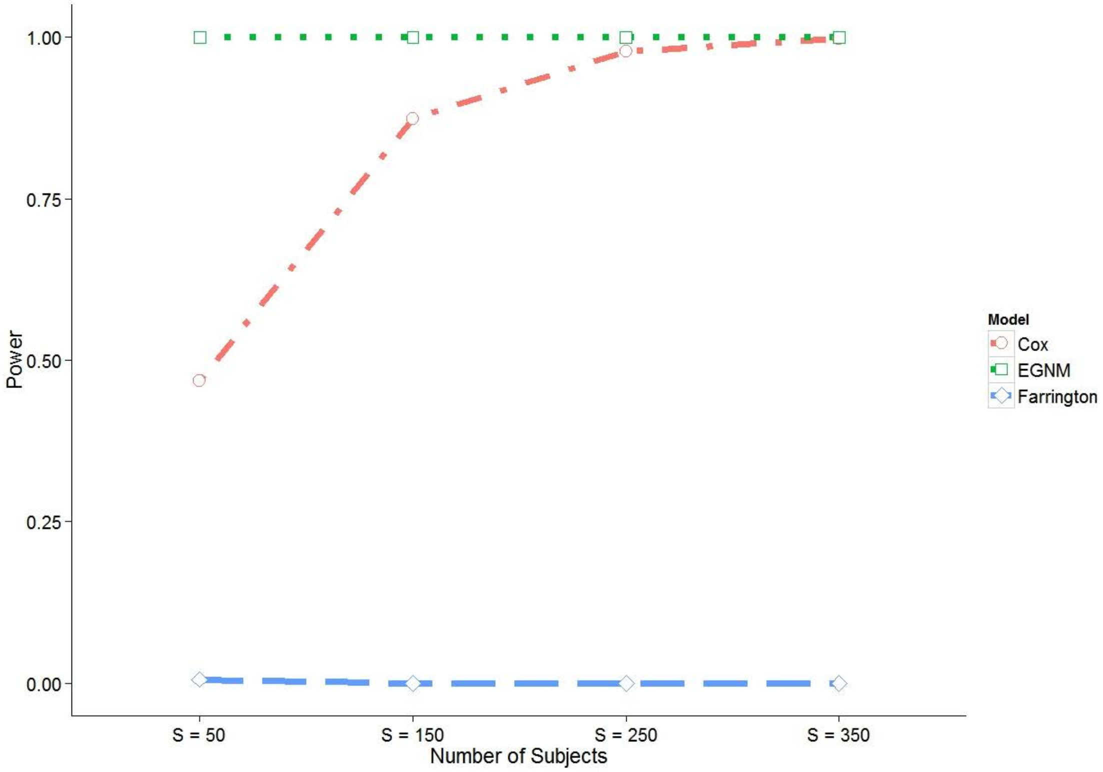

Figure 30. Power curves of the three models (R500, *c* = 2, ϱ = .7).

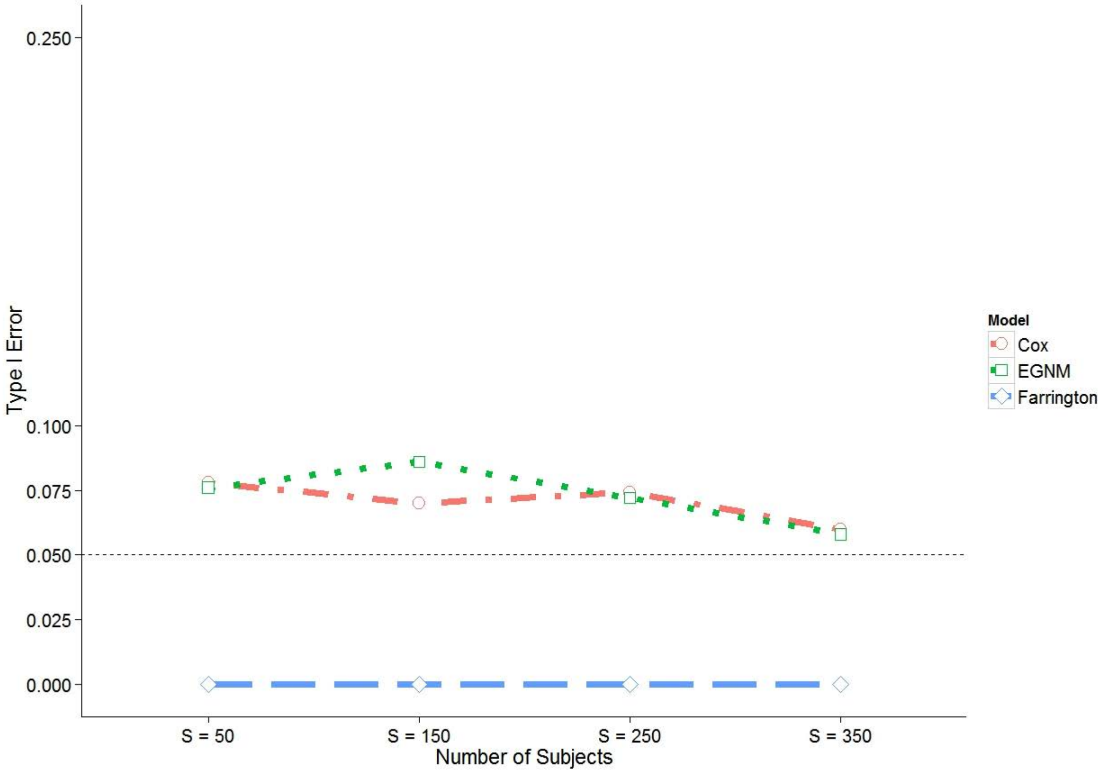

Figure 31. Type I error rates of the three models (R500, *c* = 2, *ϱ* = .7).

When the number of replications is 500, and *c* = 2 and *ϱ* = .7, power from the extended Cox model was acceptable at the .90 level when the number of subjects was at least 250, the EGNM was potentially overpowered even when the number of subjects was 50, and power from Farrington's model was not acceptable. Overall, the EGNM controlled type I error rate slightly better than the extended Cox model, and type I error rate from Farrington's model was 0.

Table 48

Power and Type I Error Rate for the Three Models (R500, c = 5, ϱ =.7)

	M	*S* = 50	*S* = 150	*S* = 250	*S* = 350
Power	C	0.492	0.916	0.990	0.998
	F	0.078	0	0	0
	E	1.000	1.000	1.000	1.000
Type I Error Rate	C	0.068	0.062	0.060	0.062
	F	0.020	0	0	0
	E	0.076	0.086	0.072	0.074

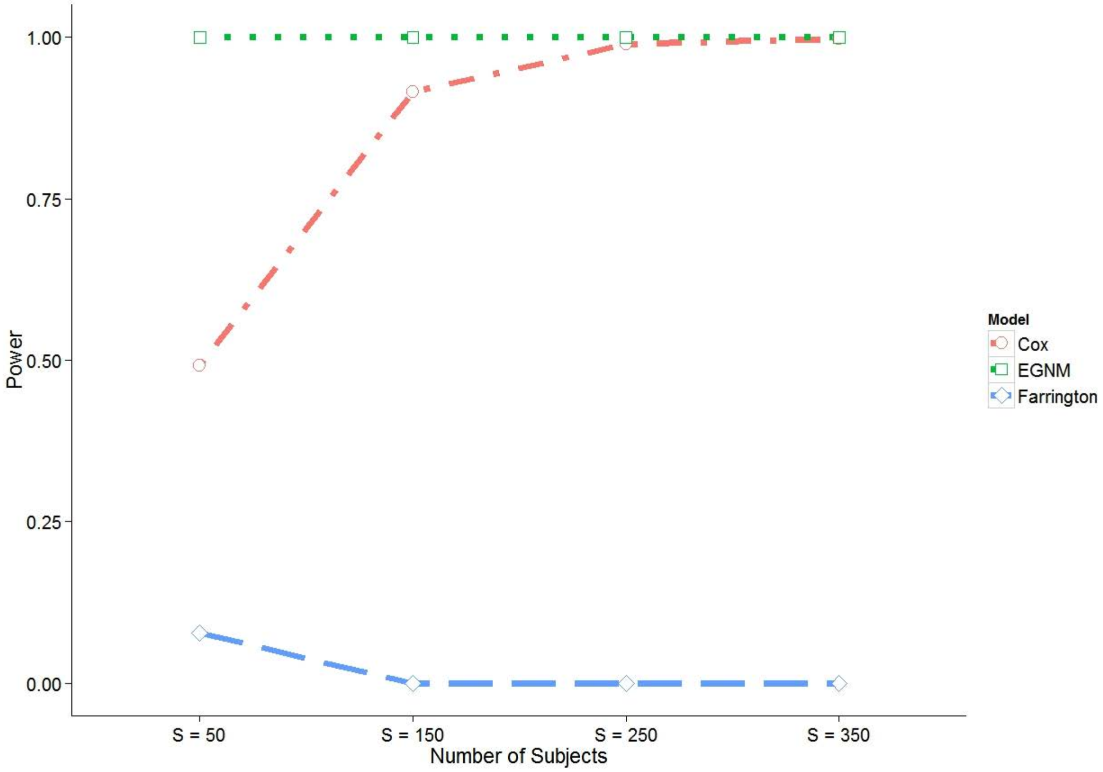

Figure 32. Power curves of the three models (R500, *c* = 5, ϱ = .7).

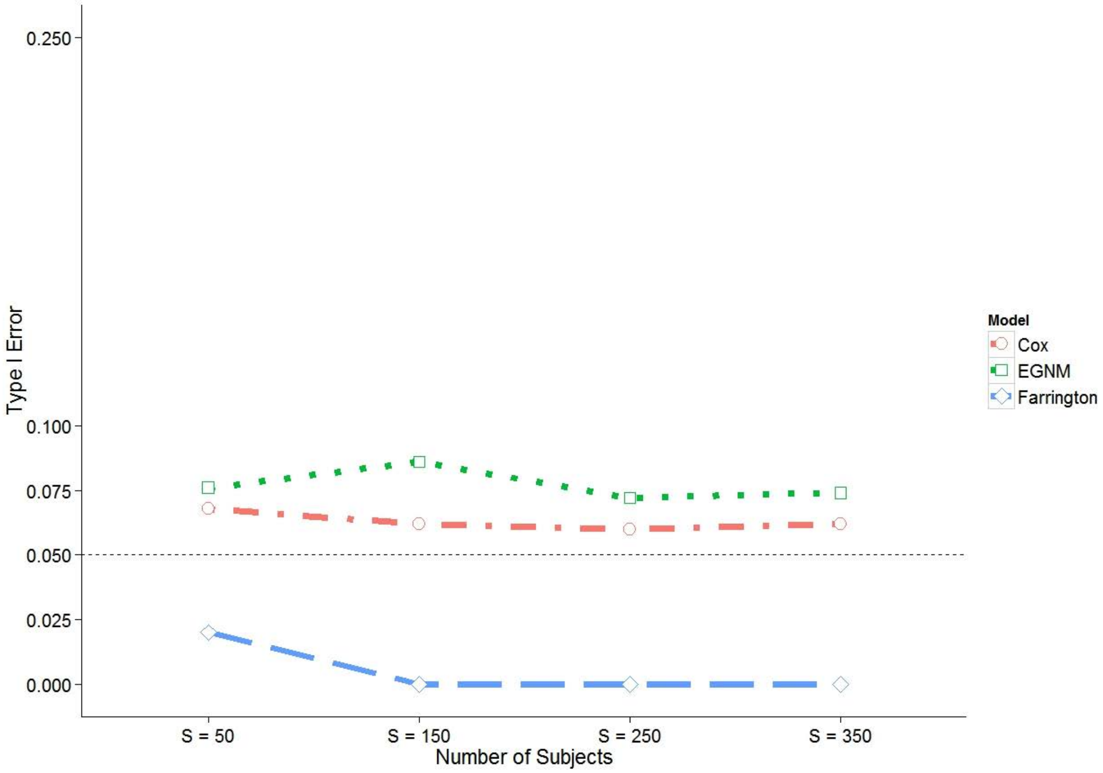

Figure 33. Type I error rates of the three models (R500, *c* = 5, *ϱ* = .7).

When the number of replications is 500, and *c* = 5 and *ϱ* = .7, the results were similar to those with 150 replications, *c* = 5 and *ϱ* = .7. The only difference lay in power from the extended Cox model was acceptable at the .90 level when the number of subjects was at least 150. The extended Cox model controlled type I error rate slightly better than the EGNM, and type I error rate from Farrington's model was almost 0.

In summary, the simulation results regarding hypothesis testing of the parameter estimates from 150 and 500 replications, with the same *c* and *ϱ* values, were similar.

When *c* = 2 or *c* = 5 with *ϱ* = .3, power from the EGNM was acceptable at the .90 level when the number of subjects was at least 150, power from the extended Cox model was acceptable at the .90 level when the number of subjects was at least 250, and power from Farrington’s model was not acceptable. Overall, the EGNM controlled type I error rate slightly better than the extended Cox model, and type I error rate from Farrington’s model was not acceptable.

When $c = 2$ or $c = 5$ with $\varrho = .7$, power from the extended Cox model was acceptable at the .90 level when the number of subjects was at least 150, the EGNM was potentially overpowered even when the number of subjects was 50, and power from Farrington’s model was not acceptable. Overall, the extended Cox model controlled type I error rate better than the EGNM, and type I error rate from Farrington’s model was not acceptable.

It is worth mentioning that overall type I error rate from the EGNM fluctuated around .05, even when the number of subjects was 1,000. There are two possible reasons for this situation. The first possible reason is there was greater variation in the scaled distribution of X_{2t}, N (0.3, 0.36), than would be expected, and thus it was easier to claim that X_{2t} was significant in describing the responsibility. The second possible reason is with repeated measures and nonnormally distributed responses, which were simulated for the EGNM, the EGNM is not robust (Oberfeld & Franke, 2013), that is, type I error rate from the EGNM showed clear deviations from the nominal type I error rate with the simulated data.

Summarizing the Simulation Results

Four comprehensive tables, Table 49-Table 52, were created to summarize the key simulation results under each combination of c and ϱ values and all numbers of subjects, including ARB($\bar{\bar{\beta}}_1$), power, and type I error rate, as the upper bound of a uniform distribution. Hence the width of a censoring interval, the probability of the descending association between X_{1t} and the response, and numbers of subjects were thought to have direct impact on the simulation results.

Tables 49 and 50 show when c is fixed, how the key simulation results behave as ϱ and the number of subjects increase. Tables 51 and 52 show when ϱ is fixed, how the key simulation results behave as c and the number of subjects increase.

Table 49

Comprehensive Table (R150c)

	M		*c* = 2		*c* = 5	
			ϱ = .3	ϱ = .7	ϱ = .3	ϱ = .7
ARB ($\bar{\bar{\beta}}_1$)	C	*S* = 50	1.3135	1.3996	1.3017	1.4050
		S = 250	1.2765	1.3839	1.2677	1.3777
		S = 500	1.2818	1.3894	1.2722	1.3877
		S = 1000	1.2773	1.3965	1.2676	1.3939
	F	*S* = 50	0.9995	0.9980	1.0019	1.0110
		S = 250	0.9970	0.9961	0.9992	0.9990
		S = 500	0.9961	0.9950	0.9978	0.9972
		S = 1000	0.9949	0.9933	0.9971	0.9963
	E	*S* = 50	0.6059	0.0450	0.6059	0.0450
		S = 250	0.6008	0.0218	0.6002	0.0218
		S = 500	0.5977	0.0189	0.5977	0.0189
		S = 1000	0.5931	0.0140	0.5931	0.0140
Power	C	*S* = 50	0.427	0.507	0.453	0.520
		S = 250	0.933	0.980	0.967	0.987
		S = 500	1.000	1.000	1.000	1.000
		S = 1000	1.000	1.000	1.000	1.000
	F	*S* = 50	0	0.007	0.060	0.087
		S = 250	0	0	0.000	0
		S = 500	0	0	0.000	0
		S = 1000	0	0	0.000	0
	E	*S* = 50	0.620	1.000	0.620	1.000
		S = 250	1.000	1.000	1.000	1.000
		S = 500	1.000	1.000	1.000	1.000
		S = 1000	1.000	1.000	1.000	1.000
Type I Error Rate	C	*S* = 50	0.140	0.067	0.060	0.053
		S = 250	0.147	0.067	0.093	0.060
		S = 500	0.133	0.067	0.093	0.047
		S = 1000	0.133	0.060	0.053	0.060
	F	*S* = 50	0	0	0.013	0.020
		S = 250	0	0	0	0.006
		S = 500	0	0	0	0
		S = 1000	0	0	0	0
	E	*S* = 50	0.067	0.073	0.067	0.073
		S = 250	0.053	0.100	0.060	0.100
		S = 500	0.060	0.093	0.060	0.093
		S = 1000	0.073	0.067	0.073	0.067

Note. M = Model.

Table 50

Comprehensive Table (R500c)

	M		$c = 2$		$c = 5$	
			$\rho = .3$	$\rho = .7$	$\rho = .3$	$\rho = .7$
ARB $(\bar{\bar{\beta}}_1)$	C	$S = 50$	1.3015	1.3928	1.2782	1.3993
		$S = 150$	1.2877	1.3952	1.2723	1.3928
		$S = 250$	1.2754	1.3877	1.2660	1.3826
		$S = 350$	1.2802	1.3895	1.2695	1.3873
	F	$S = 50$	1.0003	0.9998	1.0108	1.0109
		$S = 150$	0.9975	0.9968	1.0008	1.0009
		$S = 250$	0.9970	0.9961	0.9991	0.9991
		$S = 350$	0.9968	0.9957	0.9984	0.9980
	E	$S = 50$	0.5958	0.0657	0.6003	0.0657
		$S = 150$	0.5935	0.0366	0.5971	0.0366
		$S = 250$	0.5922	0.0258	0.5922	0.0258
		$S = 350$	0.5918	0.0211	0.5905	0.0212
Power	C	$S = 50$	0.428	0.468	0.402	0.492
		$S = 150$	0.788	0.874	0.814	0.916
		$S = 250$	0.920	0.978	0.946	0.990
		$S = 350$	0.986	0.998	0.988	0.998
	F	$S = 50$	0.002	0.006	0.058	0.078
		$S = 150$	0	0	0.004	0
		$S = 250$	0	0	0	0
		$S = 350$	0	0	0	0
	E	$S = 50$	0.676	1.000	0.654	1.000
		$S = 150$	0.996	1.000	0.988	1.000
		$S = 250$	1.000	1.000	1.000	1.000
		$S = 350$	1.000	1.000	1.000	1.000
Type I Error Rate	C	$S = 50$	0.110	0.078	0.086	0.068
		$S = 150$	0.106	0.070	0.070	0.062
		$S = 250$	0.142	0.074	0.094	0.060
		$S = 350$	0.114	0.060	0.068	0.062
	F	$S = 50$	0.002	0	0.018	0.020
		$S = 150$	0	0	0.002	0
		$S = 250$	0	0	0	0
		$S = 350$	0	0	0	0
	E	$S = 50$	0.046	0.076	0.068	0.076
		$S = 150$	0.062	0.086	0.056	0.086
		$S = 250$	0.076	0.072	0.076	0.072
		$S = 350$	0.050	0.058	0.038	0.074

Table 51

Comprehensive Table (R150ϱ)

	M		$\varrho = .3$		$\varrho = .7$	
			$c = 2$	$c = 5$	$c = 2$	$c = 5$
ARB $(\bar{\bar{\beta}}_1)$	C	$S = 50$	1.3135	1.3017	1.3996	1.4050
		$S = 250$	1.2765	1.2677	1.3839	1.3777
		$S = 500$	1.2818	1.2722	1.3894	1.3877
		$S = 1000$	1.2773	1.2676	1.3965	1.3939
	F	$S = 50$	0.9995	1.0019	0.9980	1.0110
		$S = 250$	0.9970	0.9992	0.9961	0.9990
		$S = 500$	0.9961	0.9978	0.9950	0.9972
		$S = 1000$	0.9949	0.9971	0.9933	0.9963
	E	$S = 50$	0.6059	0.6059	0.0450	0.0450
		$S = 250$	0.6008	0.6002	0.0218	0.0218
		$S = 500$	0.5977	0.5977	0.0189	0.0189
		$S = 1000$	0.5931	0.5931	0.0140	0.0140
Power	C	$S = 50$	0.427	0.453	0.507	0.520
		$S = 250$	0.933	0.967	0.980	0.987
		$S = 500$	1.000	1.000	1.000	1.000
		$S = 1000$	1.000	1.000	1.000	1.000
	F	$S = 50$	0	0.060	0.007	0.087
		$S = 250$	0	0.000	0	0
		$S = 500$	0	0.000	0	0
		$S = 1000$	0	0.000	0	0
	E	$S = 50$	0.620	0.620	1.000	1.000
		$S = 250$	1.000	1.000	1.000	1.000
		$S = 500$	1.000	1.000	1.000	1.000
		$S = 1000$	1.000	1.000	1.000	1.000
Type I Error Rate	C	$S = 50$	0.140	0.060	0.067	0.053
		$S = 250$	0.147	0.093	0.067	0.060
		$S = 500$	0.133	0.093	0.067	0.047
		$S = 1000$	0.133	0.053	0.060	0.060
	F	$S = 50$	0	0.013	0	0.020
		$S = 250$	0	0	0	0.006
		$S = 500$	0	0	0	0
		$S = 1000$	0	0	0	0
	E	$S = 50$	0.067	0.067	0.073	0.073
		$S = 250$	0.053	0.060	0.100	0.100
		$S = 500$	0.060	0.060	0.093	0.093
		$S = 1000$	0.073	0.073	0.067	0.067

Table 52

Comprehensive Table (R500ϱ)

	M		ϱ = .3		ϱ = .7	
			c = 2	c = 5	c = 2	c = 5
ARB ($\bar{\bar{\beta}}_1$)	C	S = 50	1.3015	1.2782	1.3928	1.3993
		S = 150	1.2877	1.2723	1.3952	1.3928
		S = 250	1.2754	1.2660	1.3877	1.3826
		S = 350	1.2802	1.2695	1.3895	1.3873
	F	S = 50	1.0003	1.0108	0.9998	1.0109
		S = 150	0.9975	1.0008	0.9968	1.0009
		S = 250	0.9970	0.9991	0.9961	0.9991
		S = 350	0.9968	0.9984	0.9957	0.9980
	E	S = 50	0.5958	0.6003	0.0657	0.0657
		S = 150	0.5935	0.5971	0.0366	0.0366
		S = 250	0.5922	0.5922	0.0258	0.0258
		S = 350	0.5918	0.5905	0.0211	0.0212
Power	C	S = 50	0.428	0.402	0.468	0.492
		S = 150	0.788	0.814	0.874	0.916
		S = 250	0.920	0.946	0.978	0.990
		S = 350	0.986	0.988	0.998	0.998
	F	S = 50	0.002	0.058	0.006	0.078
		S = 150	0	0.004	0	0
		S = 250	0	0	0	0
		S = 350	0	0	0	0
	E	S = 50	0.676	0.654	1.000	1.000
		S = 150	0.996	0.988	1.000	1.000
		S = 250	1.000	1.000	1.000	1.000
		S = 350	1.000	1.000	1.000	1.000
Type I Error Rate	C	S = 50	0.110	0.086	0.078	0.068
		S = 150	0.106	0.070	0.070	0.062
		S = 250	0.142	0.094	0.074	0.060
		S = 350	0.114	0.068	0.060	0.062
	F	S = 50	0.002	0.018	0	0.020
		S = 150	0	0.002	0	0
		S = 250	0	0	0	0
		S = 350	0	0	0	0
	E	S = 50	0.046	0.068	0.076	0.076
		S = 150	0.062	0.056	0.086	0.086
		S = 250	0.076	0.076	0.072	0.072
		S = 350	0.050	0.038	0.058	0.074

Tables 49 and 50, where different numbers of subjects and replications and the same combination of c and ϱ values were used, produced similar results. In particular, when $c = 2$ or $c = 5$, as ϱ increased from .3 to .7, ARB($\bar{\hat{\beta}}_1$) from the extended Cox model, which was unacceptable, increased by around 0.10; ARB($\bar{\hat{\beta}}_1$) from Farrington's model, which was unacceptable, remained similar; ARB($\bar{\hat{\beta}}_1$) from the EGNM decreased dramatically from around 0.60 to around 0.02. Power from the extended Cox model and the EGNM increased, although the EGNM was potentially overpowered. In other words, the EGNM is very sensitive and possibly would work with even smaller sample sizes and a smaller effect size; power from Farrington's model was negligible. Type I error rate from the extended Cox model became closer to the nominal level .05 overall; type I error rate from the EGNM fluctuated around .05; type I error rate from Farrington's model was negligible.

Tables 51 and 52, where different numbers of subjects and replications and the same combination of c and ϱ values were used, produced similar results. In particular, when $\varrho = .3$ or $\varrho = .7$, ARB($\bar{\hat{\beta}}_1$) and power at $c = 2$ in any model, with slight fluctuations, remained similar to ARB($\bar{\hat{\beta}}_1$) and power at $c = 5$. Type I error rate from the extended Cox model became closer to the nominal level .05 overall; type I error rate from the EGNM fluctuated around .05; type I error rate from Farrington's model was negligible.

Chapter Summary

Key simulation results regarding precision and hypothesis testing of the parameter estimates, including ARB($\bar{\hat{\beta}}_1$), % CS($\hat{\beta}_1$), power, and type I error rate are summarized. Regarding precision of the parameter estimate of β_1, ARB($\bar{\hat{\beta}}_1$) from the EGNM was the smallest among the three models. However, only when the number of subjects was at least 500 and $\varrho = .7$ and, i.e., higher probability of the descending association between X_{1t} and the response probability, regardless of the values c assumed, did the EGNM produce ARB($\bar{\hat{\beta}}_1$) at the .01 level. Otherwise, ARB($\bar{\hat{\beta}}_1$) from the EGNM was not acceptable at the .01 level. % CS($\hat{\beta}_1$) from the EGNM was always acceptable, and % CS($\hat{\beta}_1$) in the extended Cox model was always unacceptable at the 80% level. Only when the number of subjects was at least 250 and $c = 5$, or greater than 50 and $c = 2$, did Farrington's model produce % CS($\hat{\beta}_1$) at the 80% level.

Power from the EGNM was always acceptable at the .90 level either when $\varrho = .7$ or when the number of subjects was at least 150. Power from the extended Cox model was acceptable only when the number of subjects was at least 250, with the exception of .916 power when $c = 5$ and $\varrho = .7$. Power from Farrington's model was negligible.

Type I error rate from the extended Cox model became closer to the nominal level .05 overall as either c or ϱ increased, and outperformed that from the EGNM except when $c = 2$ and $\varrho = .3$. Type I error rate from the EGNM fluctuated around .05. Type I error rate from Farrington's model was negligible.

In conclusion, ϱ and the number of subjects influenced ARB($\bar{\hat{\beta}}_1$) substantially among the three models. The number of subjects and c had only some influence on % CS($\hat{\beta}_1$) of Farrington's model, and ϱ had no influence on % CS($\hat{\beta}_1$) of the EGNM and the extended Cox model. Power from the three models was closely related to ϱ, and the influence from the number of subjects was not obvious. Type I error rate from the three models was loosely related to c and ϱ, and the number of subjects seemed to have no influence on type I error rate.

CHAPTER V

DISCUSSION

This chapter includes a review and discussion of the results, and is organized as follows. First, the simulation results are reviewed and discussed, and limitations of the current research and future research directions are discussed. Then, recommendations of usage among applied researchers are given.

Discussion of the Simulation Results

Summary of the Simulation Results

The motivation for the research stemmed from two facts. The first fact is that the time of the occurrence of an event, as was used in the extended Cox model, is actually inappropriate. In particular, the extended Cox model uses the right-censoring mechanism, where for subjects who have already experienced the event of interest by the end of the study, the last examination time is usually recorded as the exact event time for a subject. The purpose of recording the last examination time as the exact event time is to create risk sets according to ordered exact event times for applying the partial likelihood approach (Cox, 1972). However, chances are slim that subjects would experience an event of interest exactly at the last examination. In other words, an exact event time, as is required in the extended Cox model, does not truly describe when a subject experience an event. The second fact is the association between time-independent covariates, as is used in Farrington's model, and the status of an event is not strong. Thus the EGNM, which accommodates an imprecise, but more appropriate description of the time of the occurrence of an event and external time-dependent covariates, was thought to depict the survival experience of subjects in a follow-up study where subjects are examined intermittently more realistically than either the extended Cox model or Farrington's model. The simulation study supported the supposition.

However, the findings in favor of the EGNM from the simulation study are not unconditional. First, the unique form in Equation 54 for the response probability in the EGNM dictates the descending association between X_{1t} and the response probability, as was illustrated in Figure 1. As such, the probability of the descending association affected the simulation results. Second, the width of a censoring interval dictates the degree of uncertainty about when the event occurs. As such, the upper bound of the uniform distribution c used to create a censoring interval affected the simulation results.

In conclusion, ϱ, i.e., the probability that the smallest X_{1t} value is associated with the event of interest, influenced ARB($\bar{\bar{\hat{\beta}}}_1$) substantially among the three models. ARB($\bar{\bar{\hat{\beta}}}_1$) from the EGNM was acceptable, and Farrington's model was acceptable. ARB($\bar{\bar{\hat{\beta}}}_1$) from the extended Cox model was not acceptable even when stronger association was established between the smallest X_{1t} value and an imputed exact event time. The number of subjects had substantial impact on ARB($\bar{\bar{\hat{\beta}}}_1$) for each model in that as the number of subjects increased, the corresponding ARB($\bar{\bar{\hat{\beta}}}_1$) decreased. The number of subjects and interval width had only some influence on % CS($\hat{\beta}_1$) of Farrington's model, and the probability that the smallest X_{1t} value is associated with the event of interest had no influence on % CS($\hat{\beta}_1$) of the EGNM and the extended Cox model. Power from the three models was closely related to the probability that the smallest X_{1t} value is associated with the event of interest, and the influence from the number of subjects was not obvious. Type I error rate from the three models was loosely related to interval width and the probability that the smallest X_{1t} value is associated with the event of interest. The number of subjects surprisingly seemed to have no influence on type I error rate, as usually as the number of subjects increases, type I error rate tends to get closer to the nominal level .05.

Discussion of the Simulation Results

ARB($\bar{\bar{\beta}}_1$) from the EGNM was acceptable, and Farrington's model was not acceptable. The reason is stronger association between the smallest X_{1t} value and the response probability was established in the EGNM, while association between the X_{1t} value and the response probability was weak in Farrington's model. However, ARB($\bar{\bar{\beta}}_1$) from the extended Cox model was not acceptable, even though stronger association between the smallest X_{1t} value and an exact event time was also established. The reason is exact event times in the extended Cox model were created from the mid-point imputation method (Law & Brookmeyer, 1992), that is, regardless of how two censoring points were created, an exact event time is the middle point of two censoring points. Strong association between the smallest X_{1t} value and the response probability improved the accuracy for the EGNM, but not the extended Cox model and Farrington's model.

% CS($\hat{\beta}_1$) from the EGNM and Farrington's model were acceptable, as association, either strong or weak, between the X_{1t} value and the response probability was established. However, % CS($\hat{\beta}_1$) from the extended Cox model was not acceptable, as the corresponding % CS($\hat{\beta}_1$) pointed to the opposite direction of the effect from the significant covariate X_{1t}, even though stronger association between the smallest X_{1t} value and an exact event time was also established. The reason is as the baseline hazard decreased, the hazard of occurrence of an event of interest increased. As such, the opposite direction of the effect from X_{1t} reflected this inconsistency.

Power from the EGNM and the extended Cox model was acceptable, as stronger association, between the X_{1t} value and the response probability was established in the EGNM and the extended Cox model. However, power from Farrington's model was not acceptable, as association between the X_{1t} value and the response probability was weak in Farrington's model.

Type I error rate from Farrington's model was not acceptable, as association between the X_{2t} value and the response probability was weak in Farrington's model. However, type I error rate from the EGNM and the extended Cox model was not acceptable, as type I error rate from the two models did not stabilize and fluctuated around .05 even at the largest number of subjects, which was found through five simulation studies. The first possible reason is there was greater variation in the scaled distribution of X_{2t}, N (0.3, 0.36), than would be expected, and thus it was easier to claim that X_{2t} was significant in describing the responsibility. The second possible reason is with repeated measures and nonnormally distributed responses, both the EGNM and the extended Cox model are not robust (Oberfeld & Franke, 2013), that is, type I error rate from the two models showed clear deviations from the nominal type I error rate.

The number of subjects influenced ARB($\bar{\bar{\beta}}_1$) subtly, as the ARB($\bar{\bar{\beta}}_1$) from 50 subjects and 1,000 subjects when the probability that the smallest X_{1t} value is associated with the event of interest is low was almost the same. The number of subjects influenced power dramatically for the EGNM and the extended Cox model, but not Farrington's model. The number of subjects did not seem to influence type I error rate for the three models dramatically.

As the upper bound of the uniform distribution c changed, the resulting changing interval widths basically had no influence on ARB($\bar{\hat{\beta}}_1$), power, and type I error rate for the three models.

The probability that the smallest X_{1t} value is associated with the event of interest, influenced ARB($\bar{\hat{\beta}}_1$) substantially among the three models. Strong association, i.e., the probability that the smallest X_{1t} value is associated with the event of interest is high, improved the accuracy for the EGNM, but not the extended Cox model and Farrington's model. With strong association between the smallest X_{1t} value and the response probability, the power for the EGNM increased substantially, but the power for the extended Cox model and Farrington's model was almost the same. Strong association between the smallest X_{1t} value and the response probability basically had no influence on type I error rate among the three models, with the exception that type I error rate for the extended Cox model changed substantially when the probability that the smallest X_{1t} value is associated with the event of interest is low and the interval widths were narrow.

As such, while it is common practice to collect survival data on a regular basis from each subject after entry into a follow-up study, and then apply the Cox model (Cox, 1972), the extended Cox model (Cox, 1972; Therneau & Grambsch, 2000), or Farrington's (1996) model to investigate what factors influence the survival experience of subjects regarding the timing of the occurrence of an event, the EGNM is a promising alternative modeling approach. Suppose in reality the practitioner, such as the medical staff, tracks the occurrence of an event of interest. Instead of recording the last examination time as the exact time, as is used in the Cox model, or employing time-independent covariates, as is used in Farrington's model, the practitioner should record an imprecise event time bound in the last two examinations and employ evolving external time-dependent covariates. With the smallest parameter estimate bias, right direction of the effect, acceptable power, and comparable type I error rate, the EGNM depicts the survival experience of subjects regarding the timing of the occurrence of an event more realistically.

Limitations of the Current Research and Future Research Directions

Although GEE was successfully implemented to the EGNM, and the simulation study supported the supposition conditionally, there are still some limitations to the current research. First, in using Zhang's (2009) naive way of simulating intervals, the upper bound used to generate censoring intervals was $c = 2$ and $c = 5$, respectively. Roughly speaking, the width of the resulting censoring intervals on average was two and five, respectively. Originally $c = 5$ was thought to produce comparatively wider censoring intervals. As the mean number of examinations before a left censoring point in all data situations was around 2.1, and the mean of the simulated left censoring points was around 38, the width of each interval before a left censoring point was around 12. Thus the generated censoring intervals were narrower than the intervals before the left censoring points on average. When narrower censoring intervals created from $c = 2$ or $c = 5$ contained more definite information regarding the time of the occurrence of events, it is of interest to investigate when, for example, c is greater than 12, and hence wider intervals and more uncertainty about when the event occurs, how different the results from the corresponding simulation study would be than those from the current simulation study.

Second, the algorithm for estimating the parameters in the EGNM, i.e., GEE, was very sensitive to the choice of the true values for the parameters and distributions of the two covariates used, due to the unique form of the proposed expression for the response probability. When alternative true values for the parameters and distributions of the two covariates were used, it was found through trial and error that convergence rates for the GEE were below 80%, which is not accepted as satisfactory in a simulation study. The reason was found to be that the values calculated from Equation 54, which was required in GEE, were very close to 0, which in turn produced noninvertible matrices. As such, generalization of the EGNM to applied settings has to be exercised with caution. In addition, the distribution of X_{1t}, either the original N(79, 484), or the scaled N(.3, .06) lacked enough variation and thus caused overpowering and narrow confidence intervals when the number of subjects was greater than 250 for $\bar{\bar{\beta}}_1$. More research is needed on how to modify the EGNM to accommodate more general data situations.

Third, the simulated data sets used in the current research did not authentically mimic the data collection process in reality. For example, only arbitrarily interval-censored data were modeled for the purpose of illustration. However, in practice, both left-censored and right-censored data are collected as well, which the EGNM could not yet accommodate. As such, future research is needed to find a unified approach which is capable of modeling the three types of interval-censored data simultaneously. Moreover, in the current research, information regarding the drop-out rate in each data situation was ignored. Drop-out rates can make a simulation study more authentic account.

Fourth, recall that event times for the EGNM follow the gamma distribution, $\Upsilon \sim$ *GAM* (λ, ρ). The shape parameter ρ took the form in Equation 54,

$$\rho_{it} = 1 - e^{\left[-e^{(\beta_0+\beta_1 * \bar{X}_{1\cdot})}\right]},$$

values of which fall between 0 and 1. As such, the resulting baseline hazard function decreased monotonically, as was described in Chapter III. Consequently, the EGNM applies best to real world examples such as patients' sustainability after organ transplant, survival of burned patients, or incurrence of respiratory disease among newborn infants. In these examples, as time goes on, the hazard of the occurrence of events decreases. Future research is needed to find a modeling approach to accommodate event times with increasing baseline hazard function.

Fifth, the current research concentrated on the role of external time-dependent covariates played in the modeling process based upon the classical Cox model, which relied heavily on the assumption of proportional hazards. In both the extended Cox model and the EGNM, the inclusion of external time-dependent covariates actually violated this assumption. That is, the hazard ratio was no longer constant over time. An alternative approach, which can also accommodate changing hazards over time due to the inclusion of external time-dependent covariates, is the use of additive models (Aalen, 1989; Breslow & Day, 1987). Although additive models have not been used more frequently in applied research, there are times when it may be clinically more meaningful to express survival experience and covariate effects in terms of an additive increase or decrease in the hazard ratio. As such, the additive hazard model might be used to model arbitrarily interval-censored data with external time-dependent covariates in future research.

Sixth, due to the EGNM's inability to accommodate internal time-dependent covariates, the role of internal time-dependent covariates in modeling arbitrarily interval-censored data was not investigated. Future research is needed to find a unified approach to modeling arbitrarily interval-censored data using both external and internal time-dependent covariates together.

Overall Recommendations of Usage

The results of this simulation study were very revealing, and provided guidance on how to choose among the three models included in the current research. Suppose in reality the practitioner, such as the medical staff, tracks the occurrence of certain respiratory disease among newborn infants. In the course of follow-ups, in addition to the status of the disease, information supposed to be associated with the status is collected as well, such as environmental factors. Then, the collected information could be used in various analyses, such as regression analysis of survival data in the current research.

Based on the simulation results in the current research, Farrington's model should not be considered in the first place. Although % CS($\hat{\beta}_1$) is acceptable at the 80% level when the number of subjects was greater than 50 or 150, there is essentially no power from the model, that is, under Farrington's model, time-independent covariates could not explain variation in the response, and the true effect from X_{1t} could not be detected; approximately 100% ARB($\bar{\bar{\beta}}_1$) makes $\bar{\bar{\beta}}_1$ a very inaccurate estimate of the true value of β_1; type I error rate from Farrington's model is essentially zero, which actually becomes a problem, as the rate was far from the nominal level .05.

The extended Cox model should not be considered, either. Although power from the extended Cox is acceptable at the .90 level when the number of subjects was at least

250, and type I error rate is close to the nominal level .05 under certain conditions, at least 125% ARB($\bar{\hat{\beta}}_1$) also makes $\bar{\hat{\beta}}_1$ a very inaccurate estimate of the true value of β_1. Moreover, % CS($\hat{\beta}_1$) most gives the opposite direction of the effect from X_{1t}.

Although type I error rate from the EGNM is slightly inflated, the EGNM should still be adopted for regression analysis of such arbitrarily interval-censored survival data, which is supported by the simulation results. In particular, power from the EGNM is most acceptable at the .90 level, that is, under the EGNM, the time-dependent covariate X_{1t} explains a significant portion of variation in the response, and the true effect from X_{1t} can be detected. Approximately 1%-6% ARB($\bar{\hat{\beta}}_1$) when stronger association between the smallest X_{1t} value and the response probability was established makes $\bar{\hat{\beta}}_1$ a very accurate estimate of the true value of β_1. Moreover, % CS($\hat{\beta}_1$) almost always gives the correct direction of the effect from X_{1t}. As such, the EGNM is capable of depicting the survival experience of subjects regarding the timing of occurrence of an event of interest more realistically.

Overall Summary

In the current research it was supposed that the EGNM, which accommodates an imprecise, but more appropriate description of the time of the occurrence of an event and external time-dependent covariates, depicts survival experience of subjects in a follow-up study where subjects are examined intermittently more realistically than either the extended Cox model or Farrington's model. The simulation study supported the supposition. However, the findings in favor of the EGNM from the simulation study were not unconditional: in addition to the number of subjects, c, the upper bound of a uniform distribution, which dictates the width of a censoring interval, and ϱ, association between the smallest X_{1t} value and the response probability, or the hazards associated with an occurrence case, affected the simulation results directly.

REFERENCES

Aalen, O. O. (1989). A linear regression model for the analysis of life times. *Statistics in medicine*, *8*(8), 907-925.

Aalen, O. O., Borgan, Ø., & Gjessing, H. K. (2008). *Survival and event history analysis: a process point of view*. New York, NY: Springer.

Abrahamowicz, M., Mackenzie, T., & Esdaile, J. M. (1996). Time-dependent hazard ratio: modeling and hypothesis testing with application in lupus nephritis. *Journal of the American Statistical Association*, *91*(436), 1432-1439.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Allison, P. D. (2010). *Survival analysis using the SAS system: a practical guide* (2nd ed.). Cary, NC: SAS Institute Inc..

Andersen, P. K. (1992). Repeated assessment of risk factors in survival analysis. *Statistical Methods in Medical Research*, *1*(3), 297-315.

Bates, D. M., & Watts, D. G. (2007). *Nonlinear regression analysis and its applications*. Hoboken, NJ: Wiley-Interscience.

Betensky, R. A., Lindsey, J. C., Ryan, L. M., & Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine*, *21*(2), 263-275.

Bhatt, V., & Tiwari, N. (2014). A spatial scan statistic for survival data based on Weibull distribution. *Statistics in medicine*, *33*(11), 1867-1876.

Billingsley, P. (1999). *Convergence of probability measures* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Boos, D. D. (1992). On generalized score tests. *The American Statistician*, *46*(4), 327-333.

Brendel, M., Janssen, A., Mayer, C. D., & Pauly, M. (2014). Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, *41*(3), 742-761.

Breslow, N. E. (1972). Discussion on "Regression models and life-tables" by D. R. Cox. *Journal of the Royal Statistical Society. Series B (Methodological), 34*(2), 216-217.

Breslow, N. E., & Day, N. E. (1987). *Statistical methods in cancer research. Volume II: The Design and Analysis of Cohort Studies*. Oxford, U.K.: Oxford University Press.

Cai, T., & Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics*, *59*(3), 570-579.

Calle, M. L., & Gómez, G. (2005). A semi-parametric hierarchical method for a regression model with an interval-censored covariate. *Australian & New Zealand Journal of Statistics*, *47*(3), 351-364.

Carlin, C. S., & Solid, C. A. (2014). An approach to addressing selection bias in survival analysis. *Statistics in medicine*, *33*(23), 4073-4086.

Carstensen, B. (1996). Regression models for interval censored survival data: application to HIV infection in Danish homosexual men. *Statistics in Medicine*, *15*(20), 2177-2189.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

Chen, B., & Zhou, X. H. (2012). A latent-variable marginal method for multi-level incomplete binary data. *Statistics in medicine*, *31*(26), 3211-3222.

Chen, B., & Zhou, X. H. (2013). Generalized Partially Linear Models for Incomplete Longitudinal Data In the Presence of Population-Level Information. *Biometrics*, *69*(2), 386-395.

Chen, L. M., Ibrahim, J. G., & Chu, H. (2011). Sample size and power determination in joint modeling of longitudinal and survival data. *Statistics in medicine*, *30*(18), 2295-2309.

Chen, Q., May, R. C., Ibrahim, J. G., Chu, H., & Cole, S. R. (2014). Joint modeling of longitudinal and survival data with missing and left-censored time-varying covariates. *Statistics in medicine*, *33*(26), 4560-4576.

Cherian, T., Simoes, E., John, T. J., Steinhoff, M., & John, M. (1988). Evaluation of simple clinical signs for the diagnosis of acute lower respiratory tract infection. *The Lancet*, *332*(8603), 125-128.

Christensen, E., Schlichting, P., Andersen, P. K., Fauerholdt, L., Schou, G., Pedersen, B. V., Juhl, E., Poulsen, H., & Tygstrup, N. (1986). Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scandinavian journal of gastroenterology*, *21*(2), 163-174.

Clayton, D. (1994). Some approaches to the analysis of recurrent event data. *Statistical Methods in Medical Research*, *3*(3), 244-262.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Collett, D. (2003). *Modeling survival data in medical research* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.

Combescure, C., Foucher, Y., & Jackson, D. (2014). Meta-analysis of single-arm survival studies: a distribution-free approach for estimating summary survival curves with random effects. *Statistics in medicine*, *33*(15), 2521-2537.

Cortese, G., & Andersen, P. K. (2009). Competing Risks and Time-Dependent Covariates. *Biometrical Journal*, *52*(1), 138-158.

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

Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. New York, NY: Chapman & Hall/CRC Press.

Crowther, M. J., Look, M. P., & Riley, R. D. (2014). Multilevel mixed effects parametric survival models using adaptive Gauss-Hermite quadrature with application to recurrent events and individual participant data meta-analysis. *Statistics in medicine*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Dorey, F. J., Little, R. J., & Schenker, N. (1993). Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, *12*(17), 1589-1603.

Farrington, C. P. (1996). Interval censored survival data: A generalized linear modeling approach. *Statistics in Medicine*, *15*(3), 283-292.

Faucett, C. L., & Thomas, D. C. (1996). Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach. *Statistics in Medicine*, *15*(15), 1663-1685.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*(4), 845-854.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38.

Fisher, L. D., & Lin, D. Y. (1999). Time-dependent covariates in the Cox proportional-hazards regression model. *Annual review of public health*, *20*(1), 145-157.

Goggins, W. B., Finkelstein, D. M., Schoenfeld, D. A., & Zaslavsky, A. M. (1998). A Markov chain Monte Carlo EM algorithm for analyzing interval-censored data under the Cox proportional hazards model. *Biometrics*, *54*(4), 1498-1507.

Grenander, U. (1981). *Abstract inference*. Hoboken, NJ: Wiley-Interscience.

Hartmann, O., Schuetz, P., Albrich, W. C., Anker, S. D., Mueller, B., & Schmidt, T. (2012). Time-dependent Cox regression: Serial measurement of the cardiovascular biomarker proadrenomedullin improves survival prediction in patients with lower respiratory tract infection. *International journal of cardiology*, *161*(3), 166-173.

He, K., & Schaubel, D. E. (2014). Methods for comparing center-specific survival outcomes using direct standardization. *Statistics in medicine*, *33*(12), 2048-2061.

Heinze, G., & Dunkler, D. (2008). Avoiding infinite estimates of time-dependent effects in small-sample survival studies. *Statistics in medicine*, *27*(30), 6455-6469.

Hendry, D. J. (2014). Data generation for the Cox proportional hazards model with time-dependent covariates: a method for medical researchers. *Statistics in medicine*, *33*(3), 436-454.

Hosmer, D. W., Lemeshow, S., & May, S. (2008). *Applied survival analysis: regression modeling of time to event data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., & Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology*, *147*(7), 694-703.

Huang, J., & Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. *Proceedings of the First Seattle Symposium in Biostatistics* (pp. 123-169). New York, NY: Springer.

Kalbfleisch, J. D., & Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, *60*(2), 267-278.

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis: Techniques for censored and truncated data* (2nd ed.). New York, NY: Springer.

Kleinbaum, D., & Klein, M. (2011). *Survival Analysis: A Self-Learning Text* (3rd ed.). New York, NY: Springer.

Kooperberg, C., & Clarkson, D. B. (1997). Hazard regression with interval-censored data. *Biometrics*, *53*(4), 1485-1494.

Lachin, J. M. (2013). Power of the Mantel-Haenszel and other tests for discrete or grouped time-to-event data under a chained binomial model. *Statistics in medicine*, *32*(2), 220-229.

Law, C. G., & Brookmeyer, R. (1992). Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in medicine*, *11*(12), 1569-1578.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13-22.

Lin, D. Y., Oakes, D., & Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, *85*(2), 289-298.

Lin, D. Y., & Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, *81*(1), 61-71.

Lindsey, J. K. (1998). A study of interval censoring in parametric regression models. *Lifetime Data Analysis*, *4*(4), 329-354.

Loewy, J. W. (2015). Novel adaptive designs: aligning drug development and patient incentives. *Clinical Investigation*, *5*(4), 367-371.

Luo, X. (2011). *Longitudinal Data Analysis* [PDF document]. Retrieved from http://www.biostat.umn.edu/~xianghua/note/

Lyman, G. H., Reiner, M., Morrow, P. K., & Crawford, J. (2015). The effect of filgrastim or pegfilgrastim on survival outcomes of patients with cancer receiving myelosuppressive chemotherapy. *Annals of Oncology*, *26*(7), 1452-1458.

Ma, S., & Kosorok, M. R. (2005). Penalized log-likelihood estimation for partly linear transformation models with current status data. *Annals of Statistics*, *33*(5), 2256-2290.

Martinussen, T., & Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, *89*(3), 649-658.

Mehrotra, D. V., Li, X., Liu, J., & Lu, K. (2012). Analysis of longitudinal clinical trials with missing data using multiple imputation in conjunction with robust regression. *Biometrics*, *68*(4), 1250-1259.

McCombie, J., & Thirlwall, T. (Eds.). (2004). *Essays on balance of payments constrained growth: theory and evidence*. Florence, KY: Routledge.

McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*(2), 109-142.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York, NY: Chapman & Hall/CRC Press.

Molyneux, A. J., Birks, J., Clarke, A., Sneade, M., & Kerr, R. S. (2015). The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT). *The Lancet*, *385*(9969), 691-697.

Muggeo, V. M., Attanasio, M., & Porcu, M. (2009). A segmented regression model for event history data: an application to the fertility patterns in Italy. *Journal of Applied Statistics*, *36*(9), 973-988.

Murphy, S. A., & Sen, P. K. (1991). Time-dependent coefficients in a Cox-type regression model. *Stochastic Processes and their Applications*, *39*(1), 153-180.

National Climatic Data Center (2001). *Climatology by state based on climate division data: 1971-2000* [Data file]. Retrieved form http://www.esrl.noaa.gov/psd/data/usclimate/pcp.state.19712000.climo

Oberfeld, D., & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior research methods*, *45*(3), 792-812.

Odell, P. M., Anderson, K. M., & D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, *48*(3), 951-959.

Pan, J., Bao, Y., Dai, H., & Fang, H. B. (2014). Joint longitudinal and survival-cure models in tumour xenograft experiments. *Statistics in medicine*, *33*(18), 3229-3240

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, *56*(1), 199-203.

Paul, S., & Zhang, X. (2014). Small sample GEE estimation of regression parameters for longitudinal data. *Statistics in medicine*, *33*(22), 3869-3881.

Pepe, M. S., & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics-Simulation and Computation*, *23*(4), 939-951.

Pocock, S. J., Geller, N. L., & Tsiatis, A. A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, *43*(3), 487-498.

Prentice, R. L., & Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, *34*(1), 57-67.

Rizopoulos, D. (2012). *Joint models for longitudinal and time-to-event data: with applications in R*. Boca Raton, FL: Chapman & Hall/CRC Press.

Rücker, G., & Messerer, D. (1988). Remission duration: An example of interval-censored observations. *Statistics in Medicine*, *7*(11), 1139-1145.

Salim, A., Ma, X., Fall, K., Andrén, O., & Reilly, M. (2014). Analysis of incidence and prognosis from 'extreme' case-control designs. *Statistics in medicine*. *33*(30), 5388-5398.

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *Journal of the American Statistical Association*, *92*(438), 426-435.

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval censored data. *Biometrika*, *83*(2), 355-370.

Satten, G. A., Datta, S., & Williamson, J. M. (1998). Inference based on imputed failure times for the proportional hazards model with interval-censored data. *Journal of the American Statistical Association*, *93*(441), 318-327.

Schaubel, D. E., Zhang, H., Kalbfleisch, J. D., & Shu, X. (2014). Semiparametric methods for survival analysis of case-control data subject to dependent censoring. *Canadian Journal of Statistics*.

Shen, C. W., & Chen, Y. H. (2012). Model selection for generalized estimating equations accommodating dropout missingness. *Biometrics*, *68*(4), 1046-1054.

Shen, Y., Anderson, A., Sinha, R., & Li, Y. (2014). Joint modeling tumor burden and time to event data in oncology trials. *Pharmaceutical Statistics*.

Shi, X. (2012). *Asymptotic theory* [PDF document]. Retrieved from http://www.ssc.wisc.edu/~xshi/econ715.html

Sun, J. (2006). *The statistical analysis of interval-censored failure time data.* New York, NY: Springer.

Tang, W., He, H., & Tu, X. M. (2012). *Applied categorical and count data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Therneau, T., & Crowson, C. (2015). Using time dependent covariates and time dependent coefficients in the cox model. *Survival Vignettes*.

Therneau, T., & Grambsch, P. (2000). *Modeling survival data: extending the Cox model*. New York, NY: Springer.

Touloumis, A., Agresti, A., & Kateri, M. (2013). GEE for multinomial responses using a local odds ratios parameterization. *Biometrics*, *69*(3), 633-640.

United States Environmental Protection Agency (2013). *National Trends in Nitrogen Dioxide Concentrations in 1980-2012* [Data file]. Retrieved form http://www.epa.gov/cgibin/broker?_service=data&_program=dataprog.aqplot_data_2012.sas&parm=42602&stat=P98V&styear=1980&endyear=2012&pre=val&query=csv&region=99

Van Der Laan, M. J., & Robins, J. M. (1998). Locally efficient estimation with current status data and time-dependent covariates. *Journal of the American Statistical Association*, *93*(442), 693-701.

Vaupel, J. W., Manton, K. G., & Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, *16*(3), 439-454.

Vonesh, E. F., & Chinchilli, V. M. (1996). *Linear and nonlinear models for the analysis of repeated measurements*. New York, NY: Marcel Dekker, Inc..

Wallace, M. L. (2014). Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in Medicine*.

Wang, X., Lee, S., Zhu, X., Redline, S., & Lin, X. (2013). GEE-Based SNP Set Association Test for Continuous and Discrete Traits in Family-Based Association Studies. Genetic epidemiology, *37*(8), 778-786.

Wei, G. C., & Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, *47*(4), 1297-1309.

Westgate, P. M. (2014). Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach. *Biometrical Journal*, *56*(3), 461-476.

Westgate, P. M., & Braun, T. M. (2013). An improved quadratic inference function for parameter estimation in the analysis of correlated data. *Statistics in medicine*, *32*(19), 3260-3273.

Whitehead, J. (2014). One-stage and two-stage designs for phase II clinical trials with survival endpoints. *Statistics in medicine*. *33*(22), 3830-3843.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*(1), 95-103.

Wulfsohn, M. S., & Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, *53*(1), 330-339.

Wynant, W., & Abrahamowicz, M. (2014). Impact of the model-building strategy on inference about nonlinear and time-dependent covariate effects in survival analysis. *Statistics in medicine*, *33*(19), 3318–3337.

Xiao, Y., Abrahamowicz, M., & Moodie, E. E. (2010). Accuracy of conventional and marginal structural Cox model estimators: a simulation study. *The international journal of biostatistics*, *6*(2), 1557-4679

Zeger, S. L., & Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, *42*(1), 121-130.

Zeger, S. L., & Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in medicine*, *11*(14-15), 1825-1839.

Zhang, D. (2005). *Analysis of survival data* [PDF document]. Retrieved from http://www4.stat.ncsu.edu/~dzhang2/st745/

Zhang, X., & Paul, S. (2013). Modified Gaussian estimation for correlated binary data. *Biometrical Journal*, *55*(6), 885-898.

Zhang, Z. (2009). Linear transformation models for interval-censored data prediction of survival probability and model checking. *Statistical Modeling*, *9*(4), 321-343.

Zorn, C. J. (2001). Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, *45*(2), 470-490.

APPENDIX A

R CODE FOR THE SIMULATION STUDY

```
###############################################################################
#                              ONE_THE EGNM                                   #
###############################################################################

###############################################################################
#                          Part I Generate data                               #
###############################################################################

library(foreach)
library(iterators)
library(plyr)
library(dplyr)

NSub = 50
NRep = 20
tcoef_int = 1.5
tcoef_x1 = -3.6
mn = 0.3
std = 0.254
m_n = 0.3
s_td = 0.6
se_a = 3651
se_b = 6323
```

```
######################
# Step I_Event times #
######################


# Generate X1
set.seed(se_a)
tx1 <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do% {
rnorm(100,mn,std)
}


listindex<-matrix(c(1:(NSub*NRep)),NSub, NRep)


tx1list<-list()
foreach(i=1:(NRep*NSub), .combine='list') %do%
{tx1list[[i]]<-tx1[,i]}
tx1list
set.seed(se_a)
rg_c <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do%
{rgamma(1, shape=50, scale=1-exp(-
        exp(tcoef_int+(tcoef_x1)*mean(tx1list[[listindex[j,i]]]))))
}
```

```
rg<-matrix(rg_c, ncol=NRep)

##############################
# Step II_Random quantities #
##############################
set.seed(se_a)
rnn <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do% {runif(2,0,5)}
head(rnn)
rnnn <- foreach(i=1:NRep, .combine='cbind') %do% {rnn[,(NSub*i-(NSub-
        1)):(NSub*i)]}
rnnn_1 <- matrix(rnnn[1,],NSub,NRep)
head(rnnn_1)
rnnn_2 <- matrix(rnnn[2,],NSub,NRep)
head(rnnn_2)

######################
# Step III_Intervals #
######################
intervals_left <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do%
{c(max((rg-rnnn_1)[j,i],(rg+rnnn_2-5)[j,i]))}
left<-matrix(intervals_left,NSub,NRep)
```

```
left<-round(left)
min(left)
left<-ifelse(left>=(6+1),left,(6+1))


intervals_right <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do%
{c(min((rg+rnnn_2)[j,i],(rg-rnnn_1+5)[j,i]))}
right<-matrix(intervals_right,NSub,NRep)
right<-round(right)


e_zero<-foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='cbind') %do% {
ifelse(left[j,i]-right[j,i]==0,right[j,i]<-right[j,i]+1,right[j,i])
}
head(right)


############################################################
# STEP IV. Generating examination times for each individual #
############################################################
P_avg <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='c') %do%{
        1-exp(-exp((tcoef_int+(tcoef_x1)*mean(tx1list[[listindex[j,i]]]))))}
head(P_avg)
```

```
set.seed(se_a)
numberofet <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:NSub, .combine='c') %do% {rbinom(1,(6-1),1-P_avg[j,i])+1}


set.seed(se_a)
tdcs_long <- do.call(rbind.fill,
          lapply(1:NRep, function(i)
                do.call(rbind.fill,
                   lapply(1:NSub, function(j)
                     data.frame(rbind(sample(tx1list[[listindex[j,i]]],
                                 (numberofet[j,i]+2), replace=F)))))))


colnames(tdcs_long) <- paste("x", 1:ncol(tdcs_long), sep="")


## Generate X2.
set.seed(se_b)
stdcs_long <- do.call(rbind.fill,
          lapply(1:NRep, function(i)
                do.call(rbind.fill,
                   lapply(1:NSub, function(j)
                      data.frame(rbind(rnorm((numberofet[j,i]+2),m_n,s_td)))))))
```

```
colnames(stdcs_long) <- paste("s", 1:ncol(stdcs_long), sep="")

head(stdcs_long)

## A series of event times for each subject.

set.seed(se_a)

numberofet_fill_long <- do.call(rbind.fill,

          lapply(1:NRep, function(i)

                do.call(rbind.fill,

                   lapply(1:NSub, function(j)

                     data.frame(cbind(t(sort(c(0,sample(1:(left[j, i]-1), numberofet[j, i],

        replace=F),

                        c(left[j, i],right[j,i]))))))))))


colnames(numberofet_fill_long) <- paste("E", 1:ncol(numberofet_fill_long), sep="")


numberofet1 <- foreach(i=1:NRep, .combine='cbind') %do%

        {numberofet_fill_long[(NSub*i-(NSub-1)):(NSub*i),]}


################################

# STEP VI. Generate responses #

################################

responses_long <- do.call(rbind.fill,

          lapply(1:NRep, function(i)

                do.call(rbind.fill,
                  lapply(1:NSub, function(j)
                   data.frame(matrix(c(rep(0,(numberofet+1)[j,i]),1),nrow=1))))))


colnames(responses_long) <- paste("y", 1:ncol(responses_long), sep="")


#######################################
# STEP VII. Putting things together   #
#######################################
final_df <- foreach(i=1:NRep, .combine='cbind') %do%
      {cbind(numberofet_fill_long[(NSub*i-(NSub-1)):(NSub*i),],
            tdcs_long[(NSub*i-(NSub-1)):(NSub*i),],stdcs_long[(NSub*i-(NSub-
      1)):(NSub*i),],
                  responses_long[(NSub*i-(NSub-1)):(NSub*i),])
}


mylist <- list()
listofdfs <- foreach(i=1:NRep, .combine='list') %do%
      {mylist[[i]]=cbind(numberofet_fill_long[(NSub*i-(NSub-1)):(NSub*i),],
                        tdcs_long[(NSub*i-(NSub-
      1)):(NSub*i),],stdcs_long[(NSub*i-(NSub-1)):(NSub*i),],
                              responses_long[(NSub*i-(NSub-1)):(NSub*i),])
}
```

```
gete_long <- foreach(i=1:NRep, .combine='rbind') %do%
{mylist[[i]][,grepl( "E" , names(mylist[[i]]))]}
head(gete_long)

# k = NSub*NRep
examinations_long <- foreach(k=1:(NSub*NRep), .combine='rbind') %do%
{data.frame(cbind(rep(k,length(gete_long[k,][!is.na(gete_long[k,])])-1),
gete_long[k,][!is.na(gete_long[k,])][1:length(gete_long[k,][!is.na(gete_long[k,])])-1],
gete_long[k,][!is.na(gete_long[k,])][2:length(gete_long[k,][!is.na(gete_long[k,])])]))}

colnames(examinations_long) <- c("id","e1","e2")

############################################################
getx_long <- foreach(i=1:NRep, .combine='rbind') %do%
{mylist[[i]][,grepl( "x" , names(mylist[[i]]))]}

timedcs_long <- foreach(k=1:(NSub*NRep), .combine='rbind') %do%
{cbind(getx_long[k,][!is.na(getx_long[k,])][1:length(getx_long[k,][!is.na(getx_long[k,])]
        )])}
timedcs_long<-data.frame(timedcs_long)

colnames(timedcs_long) <- c("x1")
```

```
###############################################################################
getx2_long <- foreach(i=1:NRep, .combine='rbind') %do%
{mylist[[i]][,grepl( "s" , names(mylist[[i]]))]}


timedcs2_long <- foreach(k=1:(NSub*NRep), .combine='rbind') %do%
{cbind(getx2_long[k,][!is.na(getx2_long[k,])][1:length(getx2_long[k,][!is.na(getx2_long[
        k,])])])}
timedcs2_long<-data.frame(timedcs2_long)

colnames(timedcs2_long) <- c("x2")

###############################################################################
gety_long <- foreach(i=1:NRep, .combine='rbind') %do%
{mylist[[i]][,grepl( "y" , names(mylist[[i]]))]}


res_long <- foreach(k=1:(NSub*NRep), .combine='rbind') %do%
{cbind(gety_long[k,][!is.na(gety_long[k,])][1:length(gety_long[k,][!is.na(gety_long[k,])]
        )])}
res_long<-data.frame(res_long)

colnames(res_long) <- c("y")

finaldataframe<-cbind(examinations_long,timedcs_long,timedcs2_long,res_long)
```

```
###########################################################################
finaldataframe_1 <- foreach(i=1:NRep, .combine='rbind') %do%
{finaldataframe[finaldataframe$id %in% c((NSub*i-(NSub-1)):(NSub*i)),]}


###########################################################################
newlist <- list()
newlistofdfs <- foreach(i=1:NRep, .combine='list') %do%
{newlist[[i]]=finaldataframe_1[finaldataframe_1$id %in% c((NSub*i-(NSub-
      1)):(NSub*i)),]}
newlist
################################
# Clean the generated data    #
################################
atdlist_c <- list()
po <- foreach(i=1:NRep, .combine='list') %do% {
atdlist_c[[i]]<-group_by(newlist[[i]], id) %>%
mutate(check = ifelse(any(e1 == e2 | e1 > e2) == TRUE, 1, 0)) %>%
filter(check == 0) %>%
ungroup %>%
mutate(id = cumsum(c(TRUE, diff(id) != 0))) %>%
select(-check)
}
atdlist_c
```

```
atdlist_d<-list()
foreach(i=1:NRep, .combine='list') %do% {
dc12<-atdlist_c[[i]][,c(1,4)]
atdlist_d[[i]]<-dc12%>%
  group_by(id) %>%
      arrange(desc(x1))%>%
         filter(id<= round(0.7*max(dc12[,1])))
}

atdlist_e<-list()
foreach(i=1:NRep, .combine='list') %do% {
dc15<-atdlist_c[[i]][,c(1,4)]
atdlist_e[[i]]<-dc15%>%
  group_by(id) %>%
         filter(id > round(0.7*max(dc15[,1])))
}

atdlist_f<-list()
foreach(i=1:NRep, .combine='list') %do% {
atdlist_f[[i]]<-rbind(atdlist_d[[i]],atdlist_e[[i]])
}

atdlist<-list()
```

```
foreach(i=1:NRep, .combine='list') %do% {
atdlist[[i]]<-cbind(atdlist_c[[i]][,c(1:3)],atdlist_f[[i]][,2],
atdlist_c[[i]][,c(5:6)])
}
alh<-atdlist

###############################################################################
#                                      Part II Create IV's                                                        #
###############################################################################
ivlistdf<-list()
foreach(i=1:NRep, .combine='list') %do%
{ivlistdf[[i]]<-data.frame(cbind(atdlist[[i]]$id, atdlist[[i]]$e1,atdlist[[i]]$e2))
colnames(ivlistdf[[i]])<-c("subjectID","left", "right")
}
ivlistdf

pl<-list()
a<-foreach(i=1:NRep, .combine='c') %do% {
df<-ivlistdf[[i]]
foo <- df[order(df$right),]
stop=1
res <- c()
while(stop>0){
```

```
x <- min(foo$right)

res <- c(res, x)

pl[[i]]<-res

foo2 <- subset(foo, left >= x)

foo <- foo2

if(length(foo$right)==0)

stop=-1

}

}

pl


# Create iv's for each list and then combine the iv's.

ivlist<-list()

io<- foreach(i=1:NRep, .combine='list') %do% {

zxlist<-list()

ivlist[[i]]<-foreach(m=1:nrow(ivlistdf[[i]]), .combine='rbind') %do% {

zxlist[[m]]<-ifelse(pl[[i]] <= ivlistdf[[i]][m,3], 1, 0)

}

zxlist

}


nhm<-ivlist
```

```
n <- list()

foreach(i=1:NRep, .combine='list') %do% {

n[[i]]=cbind(ivlistdf[[i]]$subjectID,ivlist[[i]])

}

mmaxid <- c()

foreach(i=1:NRep, .combine='c') %do%

{mmaxid[i]=max(ivlistdf[[i]]$subjectID)

}


# Create iv's with all possible 1's.

pw<-list()

b<-foreach(i=1:NRep, .combine='list') %do% {

newlist1 <- list()

newlist1ofn <- foreach(m=1:mmaxid[i], .combine='list') %do%

{newlist1[[m]]=n[[i]][which(n[[i]][,1]==m),]

}

pw[[i]]=newlist1

}


# Create NA's.

pd<-list()

b<-foreach(i=1:NRep, .combine='list') %do% {

newlist2 <- list()
newlist2ofn <- foreach(m=1:mmaxid[i], .combine='list') %do%
{newlist2[[m]]<-matrix(,nrow=nrow(pw[[i]][[m]])-1, ncol=length(pl[[i]]))}
pd[[i]]=newlist2
}


pe<-list()
c<-foreach(j=1:NRep, .combine='list') %do% {
mdlist<-list()
md <- foreach(i=1:mmaxid[j], .combine='list')  %do% {
foreach(m=2:nrow(pw[[j]][[i]]), .combine='rbind')  %do% {
foreach(w=2:(length(pl[[j]])+1), .combine='rbind') %do% {
   ifelse((pw[[j]][[i]][m,w]-pw[[j]][[i]][(m-1),w])==0, pd[[j]][[i]][(m-1),(w-1)]<-0,
        pd[[j]][[i]][(m-1),(w-1)]<-pw[[j]][[i]]  [m,w])
mdlist[[i]]<-pd[[j]][[i]]
pe[[j]]=mdlist
}
}
}
pe
}
pe


pf<-list()
```

```
d<-foreach(i=1:NRep, .combine='list') %do% {
newlist3<-list()
n1<- foreach(m=1:mmaxid[i], .combine='list')  %do% {
newlist3[[m]] <-rbind(pw[[i]][[m]][1,2:(length(pl[[i]])+1)],pd[[i]][[m]])
}
pf[[i]]=newlist3
}


pg<-list()
e<-foreach(i=1:NRep, .combine='list') %do% {
newlist4<-list()
j <- foreach(m=1:mmaxid[i], .combine='list')  %do% {
newlist4[[m]] <-cbind(pw[[i]][[m]][,1],pf[[i]][[m]])
}
pg[[i]]=newlist4
}


ph<-list()
f<-foreach(i=1:NRep, .combine='list') %do% {
ph[[i]]<-foreach(m=1:mmaxid[i], .combine='rbind') %do% {
pg[[i]][[m]]
}
}
```

```
##########################################################################
#                          Part III Combining data                       #
##########################################################################
y_binary<-list()
f<-foreach(i=1:NRep, .combine='list') %do% {
y_binary[[i]]<-cbind(atdlist[[i]]$id,matrix(data.frame(atdlist[[i]])[,6],ncol=1))
colnames(y_binary[[i]])<-c("subjectID","y")
}


##
intercept<-list()
g<-foreach(i=1:NRep, .combine='list') %do% {
intercept[[i]]<-matrix(rep(1,nrow(atdlist[[i]])),ncol=1)
}


## The second column is the intercept.
X_l<-list()
h<-foreach(i=1:NRep, .combine='list') %do% {
X_l[[i]] <- cbind(atdlist[[i]]$id,intercept[[i]],atdlist[[i]]$x1,atdlist[[i]]$x2, ph[[i]][,-1])
colnames(X_l[[i]])<-c("subjectID","int","x1","x2",paste("d", 1:ncol(ph[[i]][,-1]),
        sep=""))
}
```

```
############################################################################
#                         Part IV Analysis                                 #
############################################################################
# Final full data set.
# Two "subjectID"'s.
newdat2<-list()
foreach(i=1:NRep, .combine='list') %do% {
newdat2[[i]]<-cbind(X_l[[i]],y_binary[[i]])
}
colnames(newdat2[[2]])

blh<-newdat2

ssubjectID<-list()
foreach(i=1:NRep, .combine='c') %do% {
ssubjectID[[i]]<-as.vector(newdat2[[i]][,1])
}



# Data containing d's and y alone.
newdat3<-list()
foreach(i=1:NRep, .combine='list') %do% {
newdat3[[i]]<-newdat2[[i]][,-c(1:4,(ncol(newdat2[[i]])-1))]
}


############################################################################################################
#                              TWO_FARRINGTON'S MODEL                                                      #
############################################################################################################


# after "colnames(newdat2[[2]])"
fdata_1<-list()
f<-foreach(i=1:NRep, .combine='list') %do% {
fdata_1[[i]]<-cbind(ivlistdf[[i]],newdat2[[i]][,c(2:4, ncol(newdat2[[i]]))])
}


g_b<-list()
foreach(i=1:NRep, .combine='list') %do% {
g_b[[i]]<-group_by(fdata_1[[i]], subjectID)%>%
filter(left==0|right==max(right))
}


ug_b<-list()
foreach(i=1:NRep, .combine='list') %do% {
ug_b[[i]]<-ungroup(g_b[[i]])
}
```

```
dhg<-ug_b


change_right<-list()
foreach(i=1:NRep, .combine='list') %do% {
for (e in 1:mmaxid[i]) {
dhg[[i]][(2*e-1),3]<-dhg[[i]][(2*e),2]
}
change_right[[i]]<-dhg[[i]]
}
change_right
change_tdc<-list()
foreach(i=1:NRep, .combine='list') %do% {
for (e in 1:mmaxid[i]) {
dhg[[i]][(2*e),c(5:6)]<-dhg[[i]][(2*e-1),c(5:6)]
}
change_tdc[[i]]<-dhg[[i]]
}
change_tdc


newfdata<-list()
foreach(i=1:NRep, .combine='list') %do% {
newfdata[[i]]<-data.frame(change_tdc[[i]])
}
```

```
ivflist<-list()
ifo<- foreach(i=1:NRep, .combine='list') %do%{
zxflist<-list()
ivflist[[i]]<-foreach(m=1:mmaxid[i], .combine='rbind') %do% {
zxflist[[m]]<-rbind(ifelse(pl[[i]] <= newfdata[[i]][(2*m-1),3], 1, 0),
ifelse(pl[[i]] <= newfdata[[i]][(2*m),3], 1, 0))}
zxflist
}
ivflist

fnhm<-ivflist

freplace_row2 <- foreach(i=1:NRep, .combine='c') %:%
foreach(m=1:mmaxid[i], .combine='c') %do% {ivflist[[i]][2*m,]<-
replace(ivflist[[i]][(2*m),], ivflist[[i]][(2*m-1),]>=1 & ivflist[[i]][2*m,]>=1, 0)
}

aghlist<-ivflist

foreach(i=1:NRep, .combine='list') %do% {
colnames(aghlist[[i]]) <- paste("d", 1:ncol(aghlist[[i]]), sep="")
}
```

```
# d's are from the extended method.
fatdlist<-list()
foreach(i=1:NRep, .combine='list') %do% {
fatdlist[[i]]<-cbind(newfdata[[i]],aghlist[[i]])
}
########################################################################
library(bbmle)
library(optimx)

f_pe_se_pvalues_c<-list()
foreach(i=1:NRep, .combine='list') %do% {
a<-mle2(y~dbinom(prob=1-(exp(-p)^exp(d+b*x1+c*x2)),size=1),
        parameters=list(update(as.formula(paste("p ~ ", paste(paste("d", 1:(ncol(fatdlist[[i]]
           [,c(7:ncol(fatdlist[[i]]))])-1),sep=""), collapse= "+"))), ~ .-1)),start=list(p=0.1,
          d=0.1, b=0, c=0),
              lower = c(rep(0,ncol(fatdlist[[i]][,c(7:ncol(fatdlist[[i]]))])-1),-Inf,-Inf,-Inf),
                   upper = c(rep(Inf,ncol(fatdlist[[i]][,c(7:ncol(fatdlist[[i]]))])-1),Inf,Inf,Inf),
                      optimizer="optimx",method="bobyqa",
                         data=fatdlist[[i]])
f_pe_se_pvalues_c[[i]] <-c(coef(a),tail(sqrt(1/diag(a@details$hessian)),3),1-
          pchisq((tail(coef(a),3)/
                                          tail(sqrt(1/diag(a@details$hessian)),3))^2,1))
}
```

```
f_dcoef<-list()

foreach(i=1:NRep, .combine='list') %do% {

f_dcoef[[i]] = f_pe_se_pvalues_c[[i]][-(length(f_pe_se_pvalues_c[[i]])-(8:0))]

}

f_dcoef


pl_f<-list()

a1<-foreach(i=1:NRep, .combine='c') %do% {

df1<-newfdata[[i]]

foo <- df1[order(df1$right),]

stop=1

res <- c()

while(stop>0){

x <- min(foo$right)

res <- c(res, x)

pl_f[[i]]<-res

foo2 <- subset(foo, left >= x)

foo <- foo2

if(length(foo$right)==0)

stop=-1

}

}

pl_f
```

```
# Create iv's for each list and then combine the iv's.
ivlist_f<-list()
io1<- foreach(i=1:NRep, .combine='list') %do% {
zxlist_f<-list()
ivlist_f[[i]]<-foreach(m=1:nrow(newfdata[[i]]), .combine='rbind') %do% {
zxlist_f[[m]]<-ifelse(pl_f[[i]] <= newfdata[[i]][m,3], 1, 0)
}
zxlist_f
}

nhm_f<-ivlist_f

n_f <- list()
foreach(i=1:NRep, .combine='list') %do% {
n_f[[i]]=cbind(newfdata[[i]]$subjectID,ivlist_f[[i]])
}

mmaxid_f <- c()
foreach(i=1:NRep, .combine='c') %do%
{mmaxid_f[i]=max(newfdata[[i]]$subjectID)
}

# Create iv's with all possible 1's.
```

```
pw_f<-list()

b1<-foreach(i=1:NRep, .combine='list') %do% {

newlist11 <- list()

newlist11ofn <- foreach(m=1:mmaxid_f[i], .combine='list') %do%

{newlist11[[m]]=n_f[[i]][which(n_f[[i]][,1]==m),]

}

pw_f[[i]]=newlist11

}



# Create NA's.

pd_f<-list()

b3<-foreach(i=1:NRep, .combine='list') %do% {

newlist22 <- list()

newlist22ofn <- foreach(m=1:mmaxid_f[i], .combine='list') %do%

{newlist22[[m]]<-matrix(,nrow=nrow(pw_f[[i]][[m]])-1, ncol=length(pl_f[[i]]))}

pd_f[[i]]=newlist22

}



pe_f<-list()

c<-foreach(j=1:NRep, .combine='list') %do% {

mdlist<-list()
md <- foreach(i=1:mmaxid_f[j], .combine='list')  %do% {
foreach(m=2:nrow(pw_f[[j]][[i]]), .combine='rbind')  %do% {
foreach(w=2:(length(pl_f[[j]])+1), .combine='rbind') %do% {
   ifelse((pw_f[[j]][[i]][m,w]-pw_f[[j]][[i]][(m-1),w])==0, pd_f[[j]][[i]][(m-1),(w-1)]<-0,
         pd_f[[j]][[i]][(m-1),(w-1)]<-pw_f[[j]][[i]]  [m,w])
mdlist[[i]]<-pd_f[[j]][[i]]
pe_f[[j]]=mdlist
}
}
}
pe_f
}



pf_f<-list()
d1<-foreach(i=1:NRep, .combine='list') %do% {
newlist33<-list()
n2<- foreach(m=1:mmaxid_f[i], .combine='list')  %do% {
newlist33[[m]] <-rbind(pw_f[[i]][[m]][1,2:(length(pl_f[[i]])+1)],pd_f[[i]][[m]])
}
pf_f[[i]]=newlist33
}
```

```
pg_f<-list()
e<-foreach(i=1:NRep, .combine='list') %do% {
newlist44<-list()
j <- foreach(m=1:mmaxid_f[i], .combine='list')  %do% {
newlist44[[m]] <-cbind(pw_f[[i]][[m]][,1],pf_f[[i]][[m]])
}
pg_f[[i]]=newlist44
}


ph_f<-list()
f<-foreach(i=1:NRep, .combine='list') %do% {
ph_f[[i]]<-foreach(m=1:mmaxid_f[i], .combine='rbind') %do% {
pg_f[[i]][[m]]
}
}


aghlist_f<-list()
foreach(i=1:NRep, .combine='list') %do% {
aghlist_f[[i]] = matrix(ph_f[[i]][,-1], nrow=2*NSub)
}


foreach(i=1:NRep, .combine='list') %do% {
colnames(aghlist_f[[i]]) <- paste("d", 1:ncol(aghlist_f[[i]]), sep="")
}


fatdlist_f<-list()
foreach(i=1:NRep, .combine='list') %do% {
fatdlist_f[[i]]<-cbind(newfdata[[i]],aghlist_f[[i]])
}


f_pe_se_pvalues_f<-list()
foreach(i=1:NRep, .combine='list') %do% {
a15<-mle2(y~dbinom(prob=1-(exp(-p)^exp(d+b*x1+c*x2)),size=1),
        parameters=list(update(as.formula(paste("p ~ ", paste(paste("d",
          1:(ncol(fatdlist_f[[i]]
           [,c(7:ncol(fatdlist_f[[i]]))])-1),sep=""), collapse= "+"))), ~ .-1)),start=list(p=0.1,
          d=0.1, b=0, c=0),
              lower = c(rep(0,ncol(fatdlist_f[[i]][,c(7:ncol(fatdlist_f[[i]]))])-1),-Inf,-Inf,-Inf),
                   upper = c(rep(Inf,ncol(fatdlist_f[[i]][,c(7:ncol(fatdlist_f[[i]]))])-
          1),Inf,Inf,Inf),
                      optimizer="optimx",method="bobyqa",
                         data=fatdlist_f[[i]])
f_pe_se_pvalues_f[[i]] <-c(coef(a15),tail(sqrt(1/diag(a15@details$hessian)),3),1-
          pchisq((tail(coef(a15),3)/
                                          tail(sqrt(1/diag(a15@details$hessian)),3))^2,1))
}
```

```
f_dcoef_f<-list()

foreach(i=1:NRep, .combine='list') %do% {

f_dcoef_f[[i]] = f_pe_se_pvalues_f[[i]][-(length(f_pe_se_pvalues_f[[i]])-(8:0))]

}


###################################################################


f_pe_se_pvalues<-list()

foreach(i=1:NRep, .combine='list') %do% {

f_pe_se_pvalues[[i]] = f_pe_se_pvalues_f[[i]][length(f_pe_se_pvalues_f[[i]])-(8:0)]

}

f_pe_se_pvalues


f_pe_se_pvalues_mat<-matrix(unlist(f_pe_se_pvalues), ncol=9, byrow=T)


f_pe_mat<-f_pe_se_pvalues_mat[,1:3]

f_pese_mat<-f_pe_se_pvalues_mat[,4:6]

f_pvalues_mat<-f_pe_se_pvalues_mat[,7:9]


mean(f_pe_mat[,1])

mean(f_pe_mat[,2])

mean(f_pe_mat[,3])
```

```
f_int_bias<-(mean(f_pe_mat[,1])-tcoef_int)/(abs(tcoef_int))
f_b1_bias<-(mean(f_pe_mat[,2])-tcoef_x1)/(abs(tcoef_x1))


count_f_int_sign<-sum(f_pe_mat[,1] > 0)
count_f_int_sign
corrsign_f_int_percent<-count_f_int_sign/length(f_pe_mat[,1])


count_f_b1_sign<-sum(f_pe_mat[,2] < 0)
count_f_b1_sign
corrsign_f_b1_percent<-count_f_b1_sign/length(f_pe_mat[,2])


mean(f_pese_mat[,1])
mean(f_pese_mat[,2])
mean(f_pese_mat[,3])


count_f_int_pvalues=sum(f_pvalues_mat[,1]<=0.05)
power_f_int_percent<-count_f_int_pvalues/length(f_pvalues_mat[,1])
count_f_b1_pvalues=sum(f_pvalues_mat[,2]<=0.05)
power_f_b1_percent<-count_f_b1_pvalues/length(f_pvalues_mat[,2])
count_f_b2_pvalues=sum(f_pvalues_mat[,3]<=0.05)
typeI_f_b2_percent<-count_f_b2_pvalues/length(f_pvalues_mat[,3])


#################################################
```

```
#   Step II_obtain the coefficient for the tdc  #
##################################################
yy_binary<-list()
foreach(i=1:NRep, .combine='list') %do% {
yy_binary[[i]] <- cbind(newdat2[[i]][,1],newdat2[[i]][,ncol(newdat2[[i]])])
}


XX_l<-list()
foreach(i=1:NRep, .combine='list') %do% {
XX_l[[i]] <- newdat2[[i]][,1:(ncol(newdat2[[i]])-2)]
}


##########################################################################
fx<-list()
foreach(z=1:NRep, .combine='list', .errorhandling=c('pass')) %do% {


y_binary<-yy_binary[[z]]
X_E<-XX_l[[z]]
subjectID<-ssubjectID[[z]]
maxid<- mmaxid[z]
```

```
# INVERSE LINK FUNCTION #
g_inv = function(x){1-exp(-exp(x))}

# NORM: Euclidean distance #
norm = function(x){sqrt(t(x)%*%x)}

# MINIMIZE EE USING ITERATIVE METHOD OF LIANG / ZEGER / QAQISH #
betaHat    = rep(0,3)
deltaBeta  = rep(10,3)
epsilon    = 0.0001

while(norm(deltaBeta) > epsilon)
{
        # INITIALIZE INDEX, VALUE #
        index = 1
        N = maxid
        sumA = matrix(0,3,3)
        sumB = rep(0,3)

        # CONSTRUCT DELTABETA COMPONENTS BY SUBJECT #
        for(i in 1:N)
        {
                # UPDATE RESPONSE, PREDICTORS, INDEX #
```

```
  y_binary_i = as.vector(y_binary[,-1][which(subjectID == subjectID[index])])
  X_E_i      = as.matrix(X_E[,2:4][which(subjectID == subjectID[index]),])
  index      = max(which(subjectID == subjectID[index]))+1


        # SYSTEMATIC COMPONENT #
        eta_i = as.vector(X_E_i[,1:3] %*% betaHat[1:3])


        # ESTIMATED VALUES #
        p_i = as.vector(g_inv(eta_i))
        cat("Predicted probability:")
        cat("\n")
        print(p_i)


        # RESIDUAL VECTOR #
        b_i = y_binary_i - p_i
        cat("Residual:")
        cat("\n")
        print(b_i)



        # WORKING COVARIANCE STRUCTURE #
        V_i = diag(p_i*(1-p_i))          # V_i is a diagonal matrix, with each
diagonal element being the variance
```

```
                  # of the mean. #
        cat("WCS:")
        cat("\n")
        print(V_i)

        # DERIVATIVE MATRIX #
        D_i   = log(1/(1-p_i))*(1-p_i)*X_E_i[,1:3]
        cat("d_beta:")
        cat("\n")
        print(D_i)



        # UPDATE VALUES #
        sumA = sumA + t(D_i) %*% solve(V_i) %*% D_i
        cat("sumA:")
        cat("\n")
        print(sumA)
        sumB = sumB + t(D_i) %*% solve(V_i) %*% b_i
        cat("sumB:")
        cat("\n")
        print(sumB)
}
```

```
        # UPDATE BETAHAT #
        deltaBeta = solve(sumA) %*% sumB
        cat("deltaBeta:")
        cat("\n")
        print(deltaBeta)
        betaHat   = betaHat + deltaBeta
        cat("betaHat:")
        cat("\n")
        print(betaHat)
}

        deltaBeta
        fx[[z]]<-betaHat
}
fx
length(unlist(fx))

##
e_pe_mat<-matrix(unlist(fx), ncol=3, byrow=T)
mean(e_pe_mat[,1])
mean(e_pe_mat[,2])
mean(e_pe_mat[,3])
```

```
e_int_bias<-(mean(e_pe_mat[,1])-tcoef_int)/(abs(tcoef_int))

e_b1_bias<-(mean(e_pe_mat[,2])-(tcoef_x1))/(abs(tcoef_x1))

count_e_int_sign<-sum(e_pe_mat[,1] > 0)

count_e_int_sign

corrsign_e_int_percent<-count_e_int_sign/length(e_pe_mat[,1])


count_e_b1_sign<-sum(e_pe_mat[,2] < 0)

count_e_b1_sign

corrsign_e_b1_percent<-count_e_b1_sign/length(e_pe_mat[,2])





##

covEst<-list()

foreach(r=1:NRep, .combine='list', .errorhandling=c('pass')) %do% {


y_binary<-yy_binary[[r]]

X_E<-XX_l[[r]]

subjectID<-ssubjectID[[r]]

maxid<- mmaxid[r]

betaHat<-fx[[r]]


# INVERSE LINK FUNCTION #
```

```
g_inv = function(x){1-exp(-exp(x))}


# OBTAIN STANDARD ERRORS #

##      USE BETAHAT      ##


index=1

N = maxid

sumJ = matrix(0,3,3)

sumK = matrix(0,3,3)


for(i in 1:N)

{

        # UPDATE RESPONSE, PREDICTORS, INDEX #

         y_binary_i = as.vector(y_binary[,-1][which(subjectID == subjectID[index])])

         X_E_i      = as.matrix(X_E[,2:4][which(subjectID == subjectID[index]),])

    index      = max(which(subjectID == subjectID[index]))+1


        # SYSTEMATIC COMPONENT #

    eta_i = as.vector(X_E_i[,1:3] %*% betaHat[1:3])




        # ESTIMATED VALUES #

        p_i = as.vector(g_inv(eta_i))
        # RESIDUAL VECTOR #
        b_i = y_binary_i - p_i

        # WORKING COVARIANCE STRUCTURE #
        V_i = diag(p_i*(1-p_i))

        # DERIVATIVE MATRIX #
        D_i   = log(1/(1-p_i))*(1-p_i)*X_E_i[,1:3]

        # UPDATE VALUES #
        sumJ = sumJ + t(D_i) %*% solve(V_i) %*% D_i
        sumK = sumK + t(D_i) %*% solve(V_i) %*% b_i %*% t(b_i) %*%
        t(solve(V_i)) %*% D_i
}
covEst[[r]]<- solve(sumJ) %*% sumK %*% solve(sumJ)
}
covEst

seEst<-list()
foreach(D=1:NRep, .combine='list', .errorhandling=c('pass')) %do% {
cE<-covEst[[D]]
seEst[[D]]<-sqrt(diag(cE))
```

```
}
seEst


e_pese_mat<-matrix(unlist(seEst), ncol=3, byrow=T)
e_pese_mat
mean(e_pese_mat[,1])
mean(e_pese_mat[,2])
mean(e_pese_mat[,3])


##
e_int_pvalues<-1-pchisq(((e_pe_mat[,1]/e_pese_mat[,1])^2),1)
e_b1_pvalues<-1-pchisq(((e_pe_mat[,2]/e_pese_mat[,2])^2),1)
e_b2_pvalues<-1-pchisq(((e_pe_mat[,3]/e_pese_mat[,3])^2),1)


count_e_int_pvalues<-sum(e_int_pvalues<=0.05)
power_e_int_percent<-count_e_int_pvalues/length(e_pe_mat[,1])
count_e_b1_pvalues<-sum(e_b1_pvalues<=0.05)
power_e_b1_percent<-count_e_b1_pvalues/length(e_pe_mat[,2])


count_e_b2_pvalues<-sum(e_b2_pvalues<=0.05)
typeI_e_b2_percent<-count_e_b2_pvalues/length(e_pe_mat[,3])
```

```
############################################################################
#                          THREE_THE EXTENDED COX MODEL                          #
############################################################################
ncet_minx<-list()
foreach(i=1:NRep, .combine='list') %do% {
dr1<-alh[[i]][,c(1,4)]
ncet_minx[[i]]<-dr1%>%
  group_by(id) %>%
      filter(x1 == min(x1))  %>%
           filter(id<= round(0.7*max(dr1[,1])))
}

set.seed(se_a)
rgc_c <- foreach(i=1:NRep, .combine='cbind') %:%
foreach(j=1:(round(0.7*NSub)), .combine='cbind') %do%
{rgamma(1, shape=50, scale=1-exp(-
          exp(tcoef_int+(tcoef_x1)*as.numeric(ncet_minx[[i]][j,2]))))
}

rgc<-matrix(rgc_c, ncol=NRep)

rgclist<-list()
foreach(i=1:NRep, .combine='list') %do% {
```

```
rgclist[[i]]<-rgc[,i]
}

cdat1<-list()
foreach(i=1:NRep, .combine='list') %do% {
dr2<-alh[[i]]
cdat1[[i]]<-dr2%>%
  group_by(id) %>%
        filter(id<= round(0.7*max(dr2[,1])))
}

ncet_r<-list()
foreach(i=1:NRep, .combine='list') %do% {
tu<-cdat1[[i]]
ti<-rgclist[[i]]
tu$e2[cumsum(table(tu$id))]= c(ti)
ncet_r[[i]]<-tu
}

cdat2<-list()
foreach(i=1:NRep, .combine='list') %do% {
dr3<-alh[[i]]
cdat2[[i]]<-dr3%>%
  group_by(id) %>%
        filter(id > round(0.7*max(dr3[,1])))
}

cdata<-list()
foreach(i=1:NRep, .combine='list') %do% {
cdata[[i]]<-rbind(ncet_r[[i]],cdat2[[i]])
}

cdata_c <- list()
po1 <- foreach(i=1:NRep, .combine='list') %do% {
cdata_c[[i]]<-group_by(cdata[[i]], id) %>%
mutate(check = ifelse(any(e1 == e2 | e1 > e2) == TRUE, 1, 0)) %>%
filter(check == 0) %>%
ungroup %>%
mutate(id = cumsum(c(TRUE, diff(id) != 0))) %>%
select(-check)
}

library(survival)
cdatacoef<-list()
foreach(i=1:NRep, .combine='list') %do% {
cdatacoef[[i]] <-coef(coxph(Surv(e1,e2,y) ~ x1+x2, data=cdata_c[[i]]))
```

```
}


##

c_pe_mat<-matrix(unlist(cdatacoef), ncol=2, byrow=T)

mean(c_pe_mat[,1])

mean(c_pe_mat[,2])


c_b1_bias<-(mean(c_pe_mat[,1])-(tcoef_x1))/(abs(tcoef_x1))


count_c_b1_sign<-sum(c_pe_mat[,1] < 0)

count_c_b1_sign

corrsign_c_b1_percent<-count_c_b1_sign/length(c_pe_mat[,1])

##

cdatacoefse<-list()

foreach(i=1:NRep, .combine='list') %do% {

cdatacoefse[[i]]<-diag((coxph(Surv(e1,e2,y) ~ x1+x2, data=cdata_c[[i]]))$var)^0.5

}


c_pese_mat<-matrix(unlist(cdatacoefse), ncol=2, byrow=T)

c_pese_mat

mean(c_pese_mat[,1])

mean(c_pese_mat[,2])
```

```
## coef(summary(coxph(Surv(e1,e2,y) ~ x1+x2, data=cdata_c[[35]])))[,1:5]

coxpvalues<-list()
foreach(i=1:NRep, .combine='list') %do% {
coxpvalues[[i]]<-coef(summary(coxph(Surv(e1,e2,y) ~ x1+x2, data=cdata_c[[i]])))[,5]
}
c_pvalues_mat<-matrix(unlist(coxpvalues), ncol=2, byrow=T)
c_b1_pvalues<-c_pvalues_mat[,1]
c_b2_pvalues<-c_pvalues_mat[,2]

count_c_b1_pvalues<-sum(c_b1_pvalues<=0.05)
power_c_b1_percent<-count_c_b1_pvalues/length(c_pe_mat[,1])

count_c_b2_pvalues<-sum(c_b2_pvalues<=0.05)
typeI_c_b2_percent<-count_c_b2_pvalues/length(c_pe_mat[,2])
###############################################################################################
#                               FOUR_SIMULATION RESULTS                                       #
###############################################################################################
c_pesebscs<-c(NA, NA,
mean(c_pe_mat[,1]),mean(c_pese_mat[,1]),mean(c_pe_mat[,2]),mean(c_pese_mat[,2]),N
A, c_b1_bias,NA,corrsign_c_b1_percent)
```

```
f_pesebscs<-c(mean(f_pe_mat[,1]), mean(f_pese_mat[,1]),
        mean(f_pe_mat[,2]),mean(f_pese_mat[,2]),
mean(f_pe_mat[,3]),mean(f_pese_mat[,3]),f_int_bias,f_b1_bias,corrsign_f_int_percent,c
        orrsign_f_b1_percent)
e_pesebscs<-c(mean(e_pe_mat[,1]),mean(e_pese_mat[,1]),
        mean(e_pe_mat[,2]),mean(e_pese_mat[,2]),
mean(e_pe_mat[,3]),mean(e_pese_mat[,3]),e_int_bias,e_b1_bias,corrsign_e_int_percent,
        corrsign_e_b1_percent)
pesebscs_results<-rbind(c_pesebscs,f_pesebscs,e_pesebscs)
##
c_pt<-c(NA, power_c_b1_percent,typeI_c_b2_percent)
f_pt<-c(power_f_int_percent,power_f_b1_percent,typeI_f_b2_percent)
e_pt<-c(power_e_int_percent,power_e_b1_percent,typeI_e_b2_percent)
pt_results<-rbind(c_pt,f_pt,e_pt)
##
Simulation_results<-cbind(pesebscs_results,pt_results)
colnames(Simulation_results) <-
        c("pe_int","pese_int","pe_b1","pese_b1","pe_b2","pese_b2","int_bias",
                    "b1_bias","int_cs(%)", "b1_cs(%)",
        "int_power(%)","b1_power(%)","b2_typeI(%)")
rownames(Simulation_results) <- c("Cox","Farrington", "Extended")
Simulation_results
```