2018

Bayesian Analysis of Survival Data with Missing Censoring Indicators and Simulation of Interval Censored Data

Veronica Bunn

FLORIDA STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

BAYESIAN ANALYSIS OF SURVIVAL DATA WITH MISSING CENSORING INDICATORS AND SIMULATION OF INTERVAL CENSORED DATA

By

VERONICA BUNN

A Dissertation submitted to the
Department of Statistics
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2018

Veronica Bunn defended this dissertation on July 10, 2018.
The members of the supervisory committee were:

Debajyoti Sinha

Professor Co-Directing Dissertation

Naomi Brownstein

Professor Co-Directing Dissertation

Richard Nowakowski

University Representative

Elizabeth Slate

Committee Member

Antonio Linero

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

To my sister, who has done more for me than she will ever know.

ACKNOWLEDGMENTS

Special thanks are due to my advisors, Dr. Debajyoti Sinha and Dr. Naomi Brownstein, for guiding me through the research process. Additional thanks to my committee members Dr. Elizabeth Slate, Dr. Antonio Linero, and Dr. Richard Nowakowski for their time and valuable feedback. I would also like to thank my peers in the Department of Statistics at Florida State University for their camaraderie and lunch discussions. Lastly, I'd like to thank Rumy for his unwavering belief and support in my goals – here's to us, and the next chapter in our lives.

CONTENTS

Appendices

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In some large clinical studies, it may be impractical to give physical examinations to every subject at his/her last monitoring time in order to diagnose the occurrence of an event of interest. This challenge creates survival data with missing censoring indicators where the probability of missing may depend on time of last monitoring. We present a fully Bayesian semi-parametric method for such survival data to estimate regression parameters of Cox's proportional hazards model [Cox, 1972]. Simulation studies show that our method performs better than competing methods. We apply the proposed method to data from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study.

Clinical studies often include interval censored data. We present a method for the simulation of interval censored data based on Poisson processes. We show that our method gives simulated data that fulfills the assumption of independent interval censoring, and is more computationally efficient that other methods used for simulating interval censored data.

CHAPTER 1

INTRODUCTION

Time-to-event analyses arise in many settings, including medicine, engineering, and other biological sciences. In medical and epidemiological applications, researchers may be interested in the time to mortality, or time until diagnosis of some disease of interest. It may be of interest to estimate the survival functions from the data, compare the survival functions between some groups, or to assess the effect of associated risk factors or covariates on survival time.

Time-to-event analyses require a different approach than standard statistical analyses because some subjects will not have experienced the event of interest by the end of the study. These subjects will have unknown, or censored, survival times. Methods of analysis for time-to-event data are well-developed and implementable in standard statistical software packages. Testing of differences in survival time between groups is possible with the logrank test. Users of SAS or R can easily estimate and plot the survival distribution using non-parametric methods. The Cox proportional hazards model allows for a semi-parametric method for estimation of the hazard function, in the presence of covariates.

All of the methods given above require that both event status and event or censoring time be known for all subjects. In some studies, specifically large scale prospective studies concerning a complex diagnosis, event status may be unknown for a subset of patients. To handle this problem, investigators may employ delayed event adjudication. This method allows investigators to evaluate between possible and non-cases using simple methods. Then, to decide event status of patients that are deemed "possible cases", a more lengthy or in-depth examination is given. This is the case for a study where the event of interest is onset of temporomandibular disorder (TMD). Diagnosis of TMD requires a specialized and invasive dental examination. Thus, it is infeasible to give this examination to all subjects across all time-points. Instead, patients are given a relatively simple assessment to evaluate their possibility of having TMD. Patients who screen positively from this assessment are deemed "possible cases" are then brought to be evaluated and diagnosed with TMD. However, not all patients receive the in-depth examination, due to inability or unwillingness to travel to the

research center to receive an examination. Thus, a time-to-event analysis of this data will have a subset of subjects that are deemed "possible cases", but are never diagnosed, resulting in missing censoring indicators. This data requires addressing the statistical challenges of missing censoring indicators using proper and careful methods to avoid a biased analysis.

In the method above, we assume that if a subject's event status is censored, it is right-censored, meaning that we never observe this subject fail. Another type of censoring that commonly occurs is interval censored data. In this type of data, we do not observe an exact failure time. Instead, we our knowledge of a subject's failure time is that it occurred between two time-points. This scenario often arises in medical data, since a subject is not continuously observed. They may have an appointment at some time-point A, in which they were observed to not experience the event of interest, but by their next follow-up appointment at time-point B, they have experienced the event of interest. Thus, a subject is observed to fail, but their exact failure time is unknown. The only information available is that the subject failed at some time between A and B.

A common assumption in the analysis of interval censored data is that the interval censoring is independent, meaning that the censoring time is independent of the failure time. Under this assumption, it is not necessary to model the censoring mechanism while analyzing the interval censored data. This will give a much more simplified likelihood function, yielding an overall simpler and more efficient way to model time until failure.

Data simulation is necessary when performing statistical research. Studies using simulated data are important to explore potential biases, behavior of estimators, and to compare the results of various statistical procedures under difference scenarios. Standard software available for the analysis of interval censored data makes the assumption of independent interval censoring. Therefore, it is imperative to make sure that the simulated data used for simulation studies follows this same assumption of independence.

Additionally, not all methods for simulating interval censored data are equally efficient. When simulating interval censored data, the goal should be to replicate the pieces of information known from a real world interval censored dataset: the bounds of the observed censoring interval. We found that the available methods for the simulation of interval censored data depend on a simulating an extra piece of information: the number of appointments, or inspections, for each patient. To make simulations more computationally efficient, and less statistically complex, we advocate for a method

of simulating independent interval censored data that does not depend on generating the number of inspection times.

We review general concepts of survival analysis and survival models in Chapter 2, followed by overviews of methods for handling missing data in survival analysis and semi-parametric Bayesian methods for data analysis.

In Chapter 3, we outline a theoretical problem with previous research effort in the field of survival data with missing censoring indicators. We give a solution to this problem, and give a fully Bayesian method for parameter estimation in the case of missing censoring indicators in a Cox proportional hazards model.

In Chapter 4, we introduce assumptions made when simulating and analyzing interval censored data, as well as review common methods for the simulation of interval censored data. We propose a novel methodology for simulating interval censored data with independent censoring and give a proof that shows this simulation method fulfills the assumption of independent interval censoring. We extend this method to allow for the simulation of generalized type of interval censored data, where the observed interval boundaries may depend on time.

CHAPTER 2

LITERATURE REVIEW

Our methodology relies on an understanding of the basics of survival analysis, semi-parametric Bayesian modeling, and missing data. In this chapter, we give a review of literature pertaining to these topics.

2.1 Survival Analysis

In many studies, both clinical and otherwise, the main outcome of interest is the time to a particular event. Survival analysis measures the time until the occurrence of an event of interest, where the risk of occurrence changes over time. Let T be a non-negative, continuous random variable denoting the failure time of a subject. The probability of a subject surviving beyond a certain time t is called the survival function, defined as

$$S(t) = \mathrm{P}(T > t), \tag{2.1}$$

where $S(t)$ is a right-continuous, monotonically decreasing function with $S(0) = 1$ and $lim_{t\to\infty}S(t) = 0$). The elapsed time from $t = 0$ to T is referred to as the survival time. Another characterization of the distribution of T is given by the hazard function $\lambda(t)$, or the instantaneous rate of occurrence of the event of interest. This is given as

$$\lambda(t) = lim_{dt\to 0}\frac{\mathrm{P}(t \leq T < t + dt|T \geq t)}{dt} \tag{2.2}$$

The expression on the right in (2.2) gives the conditional probability that the event will occur in the time interval $[t, t + dt)$, given that it has not yet occurred, dividing over the width of the interval. Difficulties in survival analysis stem from the fact that in a finite time interval, not all subjects will experience the event of interest. Thus, these subjects have survival times that are unknown, or censored. The three main types of censoring are right, left, and interval. The most commonly encountered form of censoring is right censoring. A subject is said to be right censored if they do not experience the event of interest during the duration of the study, i.e, the survival time is greater

than the censoring time C. Left censoring occurs when the event of interest occurs prior to entry into a study, such that the survival time is less than the censoring time C. If the precise time of the event of interest is known only to have occurred within an interval, then the subject has an interval censored observation such that $L \leq T < U$, where L is the last last observed time before failure occurred, and U is the first observed time after the failure occurred. This is a common scenario in medical data, where researchers will not know the exact time that a patient experienced the event of interest. Instead, the researchers only know that a subject experienced the event between two successive appointments or surveys.

When dealing with censored data, researchers may take a simple approach, such as setting the censored observations as missing, or replacing the unobserved failure time with a minimum, maximum, or mean value. These simple solutions cause serious bias in both the estimates and the standard error of the estimates obtained in subsequent statistical analysis. If subjects with censored failure times are discarded altogether, this both discards potentially important information and can create a non-representative sample of the population studied. Therefore, data that contain censored observations must be dealt with in a careful and appropriate manner to avoid introducing bias into the statistical analysis.

The goals of an analysis of survival data are varied, and may include estimation of the survival function or comparing the survival functions between different groups. Kaplan and Meier [1958] give an estimate of the survival distribution using a non-parametric maximum likelihood estimate. The main appeal of estimation of the hazard function is that it can be used to visualize details in failure risk patterns, and can be used to identify periods of elevated risk of failure in a population [Aalen and Gjessing, 2001].

Another insight of interest in a survival analysis is modeling the relationship between a (set of) predictor variable(s) and the time until failure. Feigl and Zelen [1965] introduced a single-covariate regression model under the assumption of an exponentially distributed failure time. Zippin and Armitage [1966] extended this work to include subjects with censored data. The single covariate model defined above is limited, but the general framework has been extended to other parametric models that allow for multiple covariates, and different distributions for the failure times. These models have been widely discussed and used in practice, [Collett, 2015, Rodriguez, 2010]. Parametric models such as these all have the stringent assumption of a fixed hazard distribution.

2.1.1 Cox's Proportional Hazards Model

Cox [1972] introduced a now well-known model, commonly referred to as Cox's proportional hazard model. This model allows the baseline hazard function to be flexible over time, instead of fixed as in a parametric models given above. The Cox model estimates both the baseline hazard function as well as the regression covariate(s). The proportional hazards assumption is an important property of the Cox proportional hazards model in which the hazard ratio is required to be constant over time. In Cox [1975], Cox defines a partial likelihood, and shows that the large-sample properties of the maximum likelihood estimates of the regression parameters apply when using a partial likelihood.

Efron [1977] shows that the Cox partial likelihood method is asymptotically efficient, given that the Fisher information matrix for the regression parameter is (under broad conditions) asymptotically equal to information based on the partial likelihood. Additionally, Efron connects the estimate of the hazard rate to the estimate of the survival function given by Kaplan and Meier [1958]. Over time, the methodology for Cox's model has been extended to cover competing risks,

2.1.2 Semi-parametric Bayesian Methods in Cox's Proportional Hazards Model

Piecewise Constant Hazard Models. The piecewise exponential model, first proposed by Friedman [1982], extends the basic exponential model by partitioning the time axis into intervals, and assuming that the hazard rate is constant within each interval. Extending the survival model to accommodate various shapes of the baseline hazard while keeping the assumption of piecewise constant hazards yields the piecewise constant hazard model, one of the most convenient and popular models for semi-parametric survival analysis. An overview of the construction of the piecewise constant hazard model is given in [Ibrahim et al., 2001]. Piecewise constant hazard models are often used within a Bayesian framework, given that they reduce computational complexity by limiting the number of parameters required to model the baseline hazard function.

Prior Elicitation. The gamma process prior is the most commonly used prior process for Cox's model. This prior was first introduced by Kalbfleisch [1978], in which a gamma process prior is used for the baseline cumulative hazard function. For more on this prior process, see Burridge [1981], Ibrahim et al. [2001]. Alternatively, a gamma process prior can be specified on

the hazard rate itself, as introduced by Dykstra and Laud [1981]. Other prior processes on the cumulative baseline hazard include the correlated gamma process [Mezzetti and Ibrahim, 2000], beta process [Hjort, 1990], and the Dirichlet process [Ferguson and Phadia, 1979, Susarla and Van Ryzin, 1976]. When using a piecewise constant baseline hazard model, we use a discretization of the non-parametric Gamma process prior on the hazard rate λ_j, where $\frac{a}{b}$ is the prior mean ("guess") and b is the "precision" of the unknown baseline hazard λ_j in some interval j. For details on the likelihood construction for this model, see Ibrahim et al. [1999].

2.2 Missing Data

Missing data, both covariate and response, is a problem encountered in nearly all data collection scenarios. There are well developed methods of handling missing data, as well as commonly used naive methods that may result in biased analyses. In this section, we introduce some common terminology in the field of missing data, as well as review the literature on common techniques for handling missing data.

2.2.1 General Theory and Methods

When missing data is a concern, assumptions need to be made about the mechanism that generated the missing data. Potential reasons for missingness can include lost data, patient refusal, study dropout. Rubin [1976] gives the classic definitions for three missing data mechanisms:

1. Missing Completely at Random (MCAR): The probability of having a missing value is independent of both the observed data and the missing data
2. Missing at Random (MAR): The probability of having a missing value is independent of any missing values, but may depend on observed covariates or values
3. Missing Not at Random (MNAR): The probability of having a missing value depends on the missing value itself

The missing-data mechanism is described as "non-ignorable" if the failure to observe a value depends on the value that would have been observed. Therefore, MCAR and MAR are said to be "ignorable" missingness, whereas if data are missing MNAR, the missingness is "non-ignorable". More information is given in Sun [2006].

Missing data is a major issue in many applied settings, particularly in the biomedical sciences. There are three major approaches to handling missing data: deletion based methods, simple replacement methods, and model based or conditional methods. Deletion based methods, especially complete case analyses, are the most common way of handling missing data (either covariate or response). A complete case analysis is a naive method that simply entails omitting any patients with missing data. When using a Cox regression model to analyze data in SAS, STATA, or R, all software default to omitting subjects with missing covariate or response data. Particularly when the missing data mechanism is non-ignorable, complete case analysis is not an appropriate tool to use, and will lead to biased and inefficient parameter estimates.

The most common type of missing data is missing covariates, where some subjects may have one or more unobserved covariates. When modeling survival data with covariates, another potential type of missing data arrive: missing censoring indicators. This is a less studied missing data problem. When dealing with survival data, the most common way of dealing with missing censoring indicators is to use a simple replacement method, such as treating all patients with missing indicators as censored, or treating all patients with missing indicators as failures. Simple replacement methods such as these, as well as complete case analyses, will often lead to biased and/or inefficient parameter estimation. Previous research efforts in this area [Brownstein et al., 2015, Cook and Kosorok, 2004] have shown that when some portion of censoring indicators is unknown, regression parameters can be more accurately estimated when these censoring indicators are augmented using model based or conditional methods. We review a selection of such methodologies below.

2.2.2 Model Based Methods for Missing Data in Survival Analysis

EM Algorithm. The Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is one of the most ubiquitous methods for handling missing data. This algorithm entails three distinct steps:

1. Calculation of the expected value of the full log-likelihood given the observed data and current parameter estimate,
2. Updating the parameter estimate by maximizing the expectation,
3. Iterating until convergence is reached.

In Herring and Ibrahim [2001], the authors give an EM algorithm method for missing covariates in Cox proportional hazards models. Dinse [1982] estimate the survival function under data with partially observed competing risks using the EM algorithm.

Multiple Imputation. Multiple imputation is a popular approach when dealing with missing data. This technique includes creating multiple complete datasets by filling in values for the missing data. Each filled-in data set is then analyzed as if it were a complete data set. Rubin [1996] gives a thorough summary for multiple imputation for covariate data, including comparing multiple imputation to competing methods, such as single imputation, jackknife, and bootstrapping. The motivation and basis of multiple imputation is Bayesian. The "imputation" step in multiple imputation can be viewed as sampling from the posterior predictive distribution. The process of multiple imputation contains three distinct phases: the imputation phase, the analysis phase, and the pooling phase.

In the imputation phase, m copies of the data set are created, where each copy contains different imputed valued for the missing variable. The values are imputed by using regression equations to predict for the observation with incomplete data, using the complete observations. Thus, it is very important that the imputation model is correctly specified, and fits the distributional assumptions of the data.

in the second phase, or the analysis phase, the prescribed statistical analysis is carried out on each 'completed' data set. Each completed data set consists of the imputed observations and the raw data. For each imputation, the parameter estimate $\hat{\beta}_j$ is recorded along with covariance matrix $\hat{U}_j$. This step is repeated m times, where m is the desired number of imputations.

Lastly, the multiple parameter estimates are pooled to form a single set of results. Following Rubin [2004], the pooled parameter estimate of the regression coefficient $\bar{\beta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_j$, the within-imputation variance estimate is $\bar{U} = \frac{1}{m}\sum_{j=1}^{m}\hat{U}_j$, the between-imputation variance is $\hat{B} = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\beta}_j - \bar{\beta}_j)(\hat{\beta}_j - \bar{\beta}_j)'$, and the estimated covariance matrix is $\hat{\text{Var}}(\bar{\beta}) = \bar{U} + (1+\frac{1}{m})\hat{B}$. Since it can be shown that $\bar{\beta}$ / $\hat{\text{Var}}(\bar{\beta})$ follows a t distribution with degrees of freedom $(m-1)(1+\frac{m\bar{U}}{(m+1)\hat{B}})$, confidence intervals can be computed for the multiply imputed parameter estimate $\bar{\beta}$.

Fully Bayesian. Fully Bayesian methods for missing data involve specifying prior distributions for all parameters, as well as specifying the distribution of the missing data. Connections between maximum likelihood, multiple imputation, and fully Bayesian procedures are discussed in

detail by Ibrahim et al. [2005]. Details of the fully Bayesian methodology in the presence of missing censoring indicators are discussed in 3.3.2.

CHAPTER 3

BAYESIAN ANALYSIS OF SURVIVAL DATA WITH MISSING CENSORING INDICATORS

3.1 Introduction

There is a long history of Bayesian and frequentist methods for right censored survival data with known censoring status of each subject [Ibrahim et al., 2001, Kalbfleisch and Prentice, 2002]. However, for some large cohort studies, censoring status may not be available for all subjects, that is, some subjects may not be known to either fail or be censored at their last observation time. Our motivating study is the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study, a large prospective cohort study to identify factors affecting time to first-onset of temporomandibular disorder (TMD). Confirming diagnosis of TMD requires a specialized and invasive dental exam. Given the size and duration of OPPERA study, it is infeasible to administer physical examinations to all subjects at every time-point. Each subject is instead given a screening questionnaire at pre-specified time-intervals, and every subject who screens positively through these questionnaires is then examined by clinicians. However, subjects may drop out after screening positively, due to their inability or unwillingness to travel to a research center to receive an invasive test. The result is a subset of subjects with reported positive screening outcome, but missing event status at their last monitoring time.

In the context of OPPERA, "survival time" is the time to first onset of TMD (the event of interest). Thus, subjects who had completed questionnaires, but never had a positive screening outcome, are considered right-censored at their last observation time. Subjects who had screened positively and were then subsequently examined and diagnosed with TMD have uncensored (observed) survival times. A subject who screens positively, is subsequently examined, and is determined to be free from TMD is considered right-censored at the last observation time. On the other hand, a subject who screened positively but subsequently left the study before being physically examined is considered to have a missing censoring indicator at the last observation time.

One way of dealing with missing censoring indicators is to treat all missing indicators as censored. Another option is to omit data from all subjects with missing censoring indicators. Previous research efforts in this area have shown that inference about regression parameters can be substantially biased and inefficient for complete case analysis as well as for other ad-hoc methods [Brownstein et al., 2015]. In the OPPERA study, the likelihood of a positively screened subject completing the physical examination was weakly associated with demographic variables [Bair et al., 2013], indicating that the censoring indicators may not be Missing Completely At Random (MCAR) [Rubin, 1976]. Thus, a complete case analysis of the data would result in biased estimates. In addition, the probability of a subject being diagnosed with TMD depends on a subject's response to the questionnaires. These above considerations require addressing the statistical challenges of missing censoring indicator using proper and careful methods.

Cook and Kosorok [2004] suggest consistent and asymptotically normal estimates of regression parameters of Cox's proportional hazards model [Cox, 1972] via weighting observations with missing censoring indicators with estimated probabilities of event occurrence. The standard error of their estimates are computed using a re-sampling method. Brownstein et al. [2015] use a multiple imputation method to obtain the regression parameter estimated of Cox's model [1972] while assuming a logistic regression model for the probability of failure for a subject with a missing censoring indicator. These estimated probabilities are used to generate multiple imputations of failure status for each subject with a missing censoring indicator. Both the above multiple imputation and probability weighting method depend heavily on the validity of the model for probability of failure given a missing censoring indicator. In Section 3, we explain that the methods of Brownstein et al. and Cook and Kosorok result in the underestimation of the variance of the estimator.

In this paper, we propose a model for fully Bayesian estimation of parameters in Cox's regression models [1972] for survival data with missing censoring indicators. We provide reasoning for and description of our survival model and likelihood for right-censored survival data (without missing censoring indicator) in Section 2. In Section 3, we outline the theoretical shortcomings of the previous work, and and extend the model in section 2 to present a solution in a Bayesian method for survival data with in the presence of missing censoring indicators. In Section 4, we present a simulation study to examine the performance of the proposed method and compare it with existing

methods. We introduce apply our method to the data from the OPPERA study in Section 5, followed by a discussion in Section 6.

3.2 Survival Model for Right-Censored Data

We model the effect of a $(p \times 1)$ vector of fixed covariates Z_i (measured at baseline) on the survival time T_i for subject $i = 1, \ldots, n$ using Cox's model [Cox, 1972] with hazard

$$\lambda(t|Z_i) = \lambda_0(t) \exp(\beta' Z_i) \tag{3.1}$$

and corresponding survival function $S(t|Z_i) = \exp\{-H_0(t) \exp(\beta' Z_i)\}$, where $\lambda_0(t)$ is the unspecified baseline hazard function with corresponding baseline cumulative hazard $H_0(t) = \int_0^t \lambda_0(u)du$ and $\beta = (\beta_1, \ldots \beta_p)'$ is the unknown $(p \times 1)$ vector of regression parameters. For each subject $i = 1, \ldots, n$, we denote $V_i = \min(T_i, C_i)$, and censoring indicator $\Delta_i = I(T_i \leq C_i)$, where C_i is the non-informative [Kalbfleisch and Prentice, 2002] censoring time.

We partition the time axis into $I_j = [\tau_{j-1}, \tau_j)$ for $j = 1, \ldots, J$ intervals, with cut-points $0 = \tau_0 < \tau_1 < \ldots < \tau_{J-1} < \max(V_i) \leq \tau_J$ and width $w_j = \tau_j - \tau_{j-1}$. We assume that the baseline hazard $\lambda_0(t)$ is constant within each time-interval I_j such that $\lambda_0(t) = \lambda_j$ for $t \in [\tau_{j-1}, \tau_j)$, where $\lambda_0 = (\lambda_1, \ldots, \lambda_j)$ is an unknown parameter vector.

In the scenario that all censoring indicators $\Delta_1, \ldots, \Delta_n$ are observed, the complete data likelihood is written as

$$L_C(\beta, \lambda_0|\underset{\sim}{\Delta}, \underset{\sim}{V}, \underset{\sim}{Z}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n} \{\lambda_0(V_i) \exp(\beta Z_i)\}^{\Delta_i} \exp\{-H_0(V_i) \exp(\beta Z_i)\}. \tag{3.2}$$

We now introduce an "at risk" indicator R_{ij} for a subject i in I_j as $R_{ij} = 1$ if $V_i \geq \tau_{j-1}$ and $R_{ij} = 0$ if $V_i < \tau_{j-1}$. We also introduce an interval specific failure indicator as

$$N_{ij} = \begin{cases} 1 & \text{if } V_i \in I_j \text{ and } \Delta_i = 1 \\ 0 & \text{otherwise.} \end{cases}$$

When all censoring indicators Δ_i are observed, the complete data likelihood of (3.2) can be rewritten as

$$L_C(\beta, \lambda_0|N, \underset{\sim}{V}, \underset{\sim}{Z}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n} (\mu_{ij})^{N_{ij}} \exp(-\mu_{ij}), \tag{3.3}$$

where $\mu_{ij} = \lambda_j G_{ij} \exp(\beta Z_i)$ when $R_{ij} = 1$, and $\mu_{ij} = 1$ when $R_{ij} = 0$. We set $G_{ij} = w_j$ when $V_i > \tau_j$ and $G_{ij} = V_i - \tau_{j-1}$ if $V_i \in I_j$. Thus, G_{ij} is the length of observation for subject i in interval j, and μ_{ij} can be thought of as the hazard accumulated for subject i during the length of observation within interval j. We note that the likelihood of (3.3) is proportional to the likelihood obtained via treating N_{ij} as independent observations with distribution $N_{ij} \sim Poi(\mu_{ij})$. We use independent priors

$$\lambda_j \overset{ind}{\sim} Gamma(a, b) \text{ for } j = 1, \ldots, J \text{ and } \beta \sim N(\mu_0, \Sigma_0), \tag{3.4}$$

where a, b are user-specified positive values, μ_0 is an a priori specified vector, and Σ_0 is a positive definite matrix. The prior of λ_j in (3.4) is a discretization of the non-parametric Gamma process prior [Ibrahim et al., 2001], where $\frac{a}{b}$ is the prior mean ("guess") and b is the "precision" of the unknown baseline hazard λ_j in interval I_j. We set the hyperparameters a and b to ensure a reasonable prior predictive mean and variance of the survival time T within the context of the OPPERA study. There is no a priori reason to expect the distribution of regression coefficients to be skewed, therefore we use a multivariate $N(\mu_0, \Sigma_0)$ prior for β. In practice, we often assume Σ_0 to be diagonal, thus using independent $N(\mu_{0k}, \sigma^2_{0k})$ priors for each regression coefficient β_k for $k = 1, \ldots, K$ [Ibrahim et al., 2001].

3.3 Method for Missing Censoring Indicators

In the OPPERA study, the observed time V_i for subject i is the time since the entry into the study to either diagnosis of TMD or loss to follow-up.

If subject i screens positively on the questionnaire given at V_i and is subsequently examined and diagnosed with TMD, then missingness indicator $\zeta_i = 1$ and $\Delta_i = 1$, indicating that $V_i = T_i$. If subject i screens positively at V_i, is examined, and is not diagnosed to have TMD, then $\zeta_i = 1$ and $\Delta_i = 0$, indicating right censoring of T_i and $V_i < T_i$. If subject i screens positively at V_i and is not given a physical examination, then the status of Δ_i is unknown and $\zeta_i = 0$. If a subject i does not screen positively for TMD on their final questionnaire at V_i then $Q_i = 0$, and the subject is considered censored with $\Delta_i = 0$ at $V_i < T_i$. Hence, only a subject with $Q_i = 1$ is examined and has the chance to be diagnosed with TMD at V_i. Figure 1 shows the flowchart for possible subject outcomes within the OPPERA study.

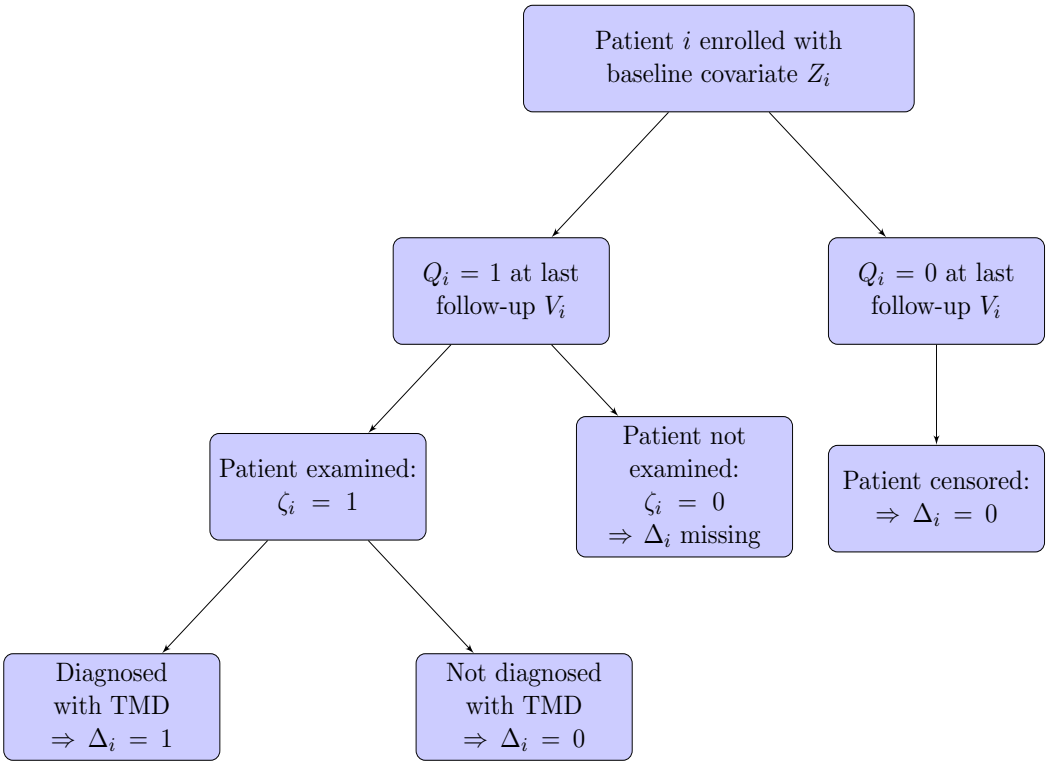

Figure 3.1: Depiction of elapsed and waiting time in relation to inspection times and failure time in interval censored data.

3.3.1 Theoretical Results

If we denote the hazard function of the non-informative censoring time C_i of the survival time T_i at time t as $\tilde{h}(t|Z_i, X_i)$, we can show that

$$\mathrm{P}(\Delta_i = 1|V_i = t, \zeta_i = 0; Z_i, X_i) = \frac{\lambda(t|Z_i)}{\lambda(t|Z_i) + \tilde{h}(t|Z_i, X_i)}, \tag{3.5}$$

where the derivation of (3.5) is given in the Appendix. Both the multiple imputation method from Brownstein et al. [2015] and the probability weighting method from Cook and Kosorok [2004] assume a logistic regression model for the conditional distribution of Δ_i given $\zeta_i = 0$ as

$$\mathrm{P}(\Delta_i = 1|V_i = t, \zeta_i = 0; Z_i, X_i) = \frac{\exp(\alpha X_i + \gamma Z_i + \eta t)}{1 + \exp(\alpha X_i + \gamma Z_i + \eta t)}, \tag{3.6}$$

while treating $\{V_i, Z_i, X_i\}$ as covariates. For (3.6) to be true, additional restrictive assumptions must be made about the relationship between the hazard $\tilde{h}(t|Z_i, X_i)$ of censoring and the hazard

function $\lambda(t|Z_i)$ in (3.1) to ensure that equations (3.5) and (3.6) are equal. More details are provided in Appendix A.

3.3.2 Bayesian Method

To overcome the need for additional assumptions on the censoring hazard and the hazard function as discussed in Section 3.1, we present a fully Bayesian method in which the likelihood function incorporates the missing censoring indicators. We assume a piecewise constant baseline hazard function as outlined in Section 2. Thus, we must find the probability of failure given a subject with a missing censoring indicator with $V_i \in I_j$, i.e., $\mathrm{P}(\Delta_i = 1|V_i \in I_j, \zeta_i = 0, Z_i, X_i)$.

To model the likelihood contribution of the data with missing censoring indicators, where $\zeta = 0$, we provide two results.

Result 1. When the censoring indicator Δ_i for subject i is not observed ($\zeta_i = 0$), the probability of subject i failing at time t depends on the censoring distribution.

Result 2. To find the probability p_{ij} that subject i with a missing censoring indicator is a failure in I_j, i.e., $\mathrm{P}(\Delta_i = 1|V_i \in I_j, \zeta_i = 0, Z_i, X_i)$, we must model both the hazard function, $\lambda(t|Z_i)$ and the censoring hazard $\tilde{h}(t|Z_i, X_i)$. This probability p_{ij} is approximated by

$$p_{ij} = \frac{\lambda(V_i|Z_i)}{\lambda(V_i|Z_i) + \tilde{h}(V_i|Z_i, X_i)} \tag{3.7}$$

The derivations for Results 1 and 2 are in Appendix B. Now, we define an indicator variable M_{ij} of a subject i being not missing in interval I_j as

$$M_{ij} = \begin{cases} 1 & \text{if } V_i \notin I_j \\ 1 & \text{if } V_i \in I_j \ \&\ Q_i = 0 \\ 1 & \text{if } V_i \in I_j \ \&\ Q_i = 1 \ \&\ \zeta_i = 1 \\ 0 & \text{if } V_i \in I_j \ \&\ Q_i = 1 \ \&\ \zeta_i = 0. \end{cases}$$

Thus for subject i, $M_{ij} = 0$ for at most in one interval I_j, and only if $V_i \in I_j$ and Δ_i is missing ($\zeta_i = 0$). When the observed data include some subjects with missing censoring indicator Δ_i, there

are four following cases for a subject's contribution to the likelihood function $L(\beta, \lambda_0|N, M, \underset{\sim}{V}, \underset{\sim}{Z})$:

Cases	Likelihood Contribution
1. Subject i is not "at risk" in I_j $R_{ij} = N_{ij} = 0, M_{ij} = 1$	$\exp(-1)$
2. Subject i is observed to survive I_j $R_{ij} = 1, N_{ij} = 0, M_{ij} = 1$	$\exp(-\mu_{ij})$
3. Subject i is observed to fail in I_j $R_{ij} = 1, N_{ij} = 1, M_{ij} = 1$	$\mu_{ij} \exp(-\mu_{ij})$
4. Subject i is at risk in I_j, but Δ_i is missing $R_{ij} = 1, N_{ij} = 0, M_{ij} = 0$	$\{(1 - p_{ij}) + p_{ij}\mu_{ij}\} \exp(-\mu_{ij})$,

where p_{ij} is defined in (3.7), and μ_{ij} is defined as

$$\mu_{ij} = \begin{cases} \lambda_j G_{ij} \exp(\beta Z_i) & \text{Cases 2 and 3} \\ 1 & \text{for Case 1} \\ 1 & \text{for Case 4.} \end{cases}$$

As defined in Section 2 after equation (3.3), G_{ij} is the width of observation for subject i in interval I_j. We note that the resultant likelihood can be expressed by

$$L(\beta, \lambda_0|N, M, \underset{\sim}{V}, \underset{\sim}{Z}) \propto \prod_{j=1}^{J} \prod_{i=1}^{n} \Big\{ (\mu_{ij})^{N_{ij}} \exp(-\mu_{ij}) \times M_{ij} + (1 - M_{ij}) \times \{(1 - p_{ij}) + p_{ij}\mu_{ij}\} \exp(-\mu_{ij}) \Big\}. \tag{3.8}$$

3.3.3 Incorporating the Likelihood into Standard Software

The likelihood function in (3.8) is a non-standard likelihood function not available in any standard statistical software. Therefore, to incorporate the likelihood of (3.8) within programs such as WinBUGS, OpenBUGS, Proc MCMC or JAGS, we introduce a new set of random variables and show that the likelihood based on pseudo values of these variables produce a likelihood equivalent to (3.8).

Consider the random variable $D_{ij} \overset{indep.}{\sim} Poi(\tau_{ij})$ where τ_{ij} is defined subsequently. Suppose that we create pseudo observations of D_{ij}, denoted D_{ijO}, where D_{ijO} and τ_{ij} are defined as follows:

	D_{ijO}	τ_{ij}	Likelihood Contribution
Case 1:	0	1	$\exp(-1)$
Case 2:	0	μ_{ij}	$\exp(-\mu_{ij})$
Case 3:	1	μ_{ij}	$\mu_{ij}\exp(-\mu_{ij})$
Case 4:	0	$-\log\{(1-p_{ij})+p_{ij}\mu_{ij}\}+\mu_{ij}$	$\{(1-p_{ij})+p_{ij}\mu_{ij}\}\exp(-\mu_{ij})$

Then likelihood contributions for each of the four possible cases of pseudo data D_{ijO} and associated parameter τ_{ij} are equivalent to the likelihood contributions from the targeted likelihood in (3.8). The resulting likelihood based on the pseudo observations D_{ijO} is

$$L(\beta, \lambda_0|D) \propto \prod_{j=1}^{J}\prod_{i=1}^{n}(\tau_{ij})^{D_{ijO}}\exp(-\tau_{ij}). \tag{3.9}$$

The likelihood of (3.9) can now be used in standard software to incorporate the targeted likelihood of (3.8). We use the priors as defined in (3.4), and include the additional prior specification such that $\tilde{h}(t) \sim \text{Gamma}(1,1)$ for $j = 1, \ldots, J$.

3.4 Simulation Study

We simulated data with missing censoring indicators. The results of several different methods were used to compare the bias, coverage, and confidence interval width. Survival times were generated for 250 individuals under a proportional hazards model, with covariates as proposed by Bender et al. [2005], and similar to the structure of simulations in Brownstein et al. [2015]. This gave the survival time for each subject as being distributed according to (1), with the baseline hazard $\lambda_0(t) = 1$. A single baseline covariate Z_i was set to a normal distribution with mean 2 and unit variance. This gives the failure times T_i, conditioned on Z_i, to be exponentially distributed with a hazard $\exp(\beta' Z_i)$, where $\beta \in \{-0.5, -1.5, -3\}$. The censoring times, C_i, were exponentially

distributed with mean 5. For $\beta = -0.5$, $\beta = -1.5$, and $\beta = -3$, this gave approximately 35%, 75%, and 90% censoring, respectively. We let $\Delta_i = I(T_i \leq C_i)$. If $\Delta_i = 0$, then we say that subject i was censored at time C_i.

The two simulated covariates are Z_i, a covariate measured at baseline, and X_i, a covariate measured on the last observed questionnaire. For each observation, X_i is generated from a normal distribution with mean Δ_i, and standard deviation 0.3. This variable is then used to generate $Q_i = I(X_i > 0.5)$, which determines if a subject screened positively on their final questionnaire. Thus, X_i is dependent on whether a subject is diagnosed with TMD, and Q_i is dependent on X_i, because the outcome of the questionnaire depends on the symptoms reported. In these simulations, δ_i is substituted as the censoring indicator in place of Δ_i, where $\delta_i = \Delta_i$ if $Q_i = 1$ and $\delta_i = \Delta_i$ if $Q_i = 0$. Therefore, the censoring indicator is set to zero if the final questionnaire is negative. This is done to more closely reflect the OPPERA study in that subjects whose final questionnaire screened negatively were never examined, and were thus never given the change to be positively diagnosed with TMD. The censoring indicators from the simulated data are set to be MAR per the definitions of Rubin [1976]:

1. Missing Completely at Random (MCAR): The probability of having a missing censoring indicator does not depend on any data, observed or missing.

2. Missing at Random (MAR): The probability of having a missing censoring indicator may depend on observed covariate(s), but does not depend on the missing censoring indicators.

3. Missing Not at Random (MNAR): The probability of having a missing censoring indicator depends on the outcome of the censoring indicator itself, which may or may not be missing. MAR is referred to as non-ignorable missingness.

3.4.1 Censoring Indicators Missing at Random

For our simulations, we assume that the data are MAR. The simulations follow the study protocol Bair et al. [2013], in that censoring indicators can only be missing for subjects who have a positive questionnaire, or where $Q_i = 1$. The censoring indicators are set to be missing with probability

$$\mathrm{P}(\zeta = 0 | X_i, Z_i, V_i, Q_i = 1) = \frac{\exp(-0.2 - 0.3Z_i + 0.1V_i)}{1 + \exp(-0.2 - 0.3Z_i + 0.1V_i)}$$

This results in a missing rate of approximately 50%, which is similar to the rate of missing censoring indicators in the OPPERA study [Bair et al., 2013].

Our method is compared with the method introduced by Brownstein et al.[2015], the method proposed by Cook and Kosorok [2004], and several naive methods described below. Following the method of Brownstein et al., for each observation with a missing censoring indicator, we estimate the probability $\hat{p}_i$ that an event occurred by using a logistic regression model based on the covariates measured at time of the putative event and baseline (X_i and Z_i) and the survival time (V_i). The coefficients in the logistic regression model were estimated using a Bayesian model with all coefficients having Cauchy(0, 2.5) priors. A Bernoulli random variable with success probability $\hat{p}_i$ is used to generate a censoring indicator for observations with a missing censoring indicator. The raw and imputed data are combined, a Cox proportional hazards model fit to the completed data set, and the resulting regression coefficient $\hat{\beta}_j$ and the variance recorded. This imputation process was repeated 10 times, to calculate 10 imputed estimates of β, where observations with missing censoring indicators have censoring indicators Δ_{ij} estimated independently for each imputation.

To obtain the estimates of Cook and Kosorok we estimate the probability $\hat{p}_i$ for each potentially unobserved event for a subject i was an event, as previously described. Then, we fit a weighted Cox proportional hazards model to the data set with the following weights. If an observation has a missing censoring indicator, that observation is deleted and replaced with two new observations, where the two new observations had identical covariates and failure times, but different censoring indicators and weights. The first new observation has weight $\hat{p}_i$, and $\hat{\Delta}_i = 1$, and the second observation is given weight $1 - \hat{p}_i$ and $\hat{\Delta}_i = 0$. Subjects with non-missing censoring indicators retained their original observation in the data set and are given a unit weight. The estimated regression coefficient, $\hat{\beta}$, is recorded.

To estimate the variance of the regression coefficient for the method of Cook and Kosorok, we generate 1,000 bootstrap replicates of each simulated dataset and refit the model for each bootstrap replicated. Within each bootstrap iteration, 250 subjects are selected from the data by sampling with replacement. The estimated probability $\hat{p}_i^*$ that each subject was a true failure is calculated for each bootstrap replicate. We calculate a bootstrap estimate $\hat{\beta}^*$ of β from the estimated $\hat{p}_i^*$'s using the weighted Cox model described in the paragraph above. This variance estimate and the average parameter estimate $\bar{\hat{\beta}}$ are used to construct percentile intervals $(\beta_{0.025}, \beta_{0.0975})$.

Additionally, we compared our method to a Bayesian analysis of the ideal data, where all values of Δ_i are observed, a Bayesian analysis of the complete case data (where all data with missing censoring indicators is excluded from analysis), and Bayesian analysis of two *ad hoc* methods where all missing censoring indicators are treated either as all censored or as all failures. The bias of each method was estimated by calculating the mean difference between the estimated Cox regression coefficient and the true coefficient over 250 simulations. We calculated the average standard error of the estimate of the Cox regression coefficient over the 250 simulations. The mean square error of the estimated regression coefficient is given, and the mean width of the credible or confidence intervals for each method above was also found. The empirical coverage probability was also calculated for the credible and confidence intervals produced by each method by dividing the number of times that the credible intervals contained the true value of the parameter by the number of simulations.

Simulations for all Bayesian methods were performed using `OpenBUGS` software via the `R` package `R2OpenBUGS` [Sturtz et al., 2005]. The bootstrap estimates of the standard error of the method introduced by Cook and Kosorok were calculated using the `boot` and `boot.ci` functions within the `boot R` package [Canty and Ripley, 2016]. The `mi.binary` function of the `mi R` package [Gelman and Hill, 2011] was used to generate imputed values of the missing censoring indicators for the method used by Brownstein et al. Cox proportional hazards models were fit using the `survfit` function in the `survival R` package [Therneau, 2015].

Results for the simulations when censoring indicators are missing at random are given in Table 3.1. The fully Bayesian method, the multiple imputation method [Brownstein et al., 2015] and the method of Cook and Kosorok [2004] produced approximately unbiased estimates and reasonable coverage in all considered scenarios. The estimates produced by the ad-hoc methods gave a larger amount of bias. Our method consistently gave the narrowest, or very close to the narrowest credible intervals for each scenario, showing that using a fully Bayesian method is better able to quantify associations. In addition, using a fully Bayesian method produces the smallest mean squared error (MSE) for the estimation of the Cox regression coefficient in all scenarios. The multiple imputation method proposed by Brownstein et al. and the probability weighting method proposed by Cook and Kosorok result in the MSE being greater than the average posterior variance estimate, which shows that these methods underestimate the variance of the regression coefficient. In contrast, the

square root of the MSE for the fully Bayesian method is approximately equal to or less than the average posterior standard error estimate in all simulated scenarios.

3.5 Analysis of the Orofacial Pain: Prospective Evaluation and Risk Assessment Study

The Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) is a large prospective cohort study designed to identify risk factors for first-onset temporomandibular disorder (TMD). Between 2006 and 2008, 3,263 subjects without TMD were recruited at four study sites. After enrolling in the study, each subject was evaluated for possible risk factors of TMD. Thereafter, each subject was asked to complete a quarterly questionnaire to evaluate recent orofacial pain. The intention of the questionnaire was to identify subjects who had likely developed TMD since their previous questionnaire. Subjects who screened positively based on their questionnaire were slated to receive a follow-up examination by a clinical expert to diagnose the presence or absence of TMD.

A small number of subjects in the OPPERA study were examined multiple times. This happened if the previous examination resulted in a non-diagnosis of TMD, and that subject had subsequently screened positively again. In this analysis, we discarded all but the most recent questionnaire and resulting diagnosis. Thus, each subject has at most one positive screening result and can be diagnosed with TMD only once.

Clinical examinations performed by one examiner resulted in a much higher percentage of subjects diagnosed with TMD compared to the other examiners, as reported by Bair et al. [2013]. We set all of the clinical examinations by this examiner as missing. We applied our method to the OPPERA data to adjust for the effect of subjects with missing failure indicators, which in this context are subjects who screened positively on the questionnaire, but have a missing clinical examination outcome.

The result of the clinical examination was not strongly associated with the majority of the variables measured on the questionnaire. The strongest predictor of being positively diagnosed with TMD was a count of non-specific orofacial symptoms in the previous three months. Other important covariates, as shown by Bair et al. [2013] were the time elapsed since enrollment and the OPPERA study site. Thus, the estimate of the probability of being diagnosed with TMD was based on the count of non-specific orofacial symptoms, time since enrollment, and the OPPERA study site.

Bair et al. [2013] examined univariate relationships between attendance at the clinical examinations and possible predictor variables. A few differences between examined and non-examined subjects were statistically significant, indicating that the data are not MCAR.

Tables 3.2, 3.3 and 3.4 gives the results of applying our method to a subset of the risk factors for TMD measured by the questionnaire in the OPPERA study. All continuous variables were normalized to have mean 0 and standard deviation 1, prior to fitting Cox models. Hence, the hazard ratios for continuous variables represent the hazard ratios corresponding to a one standard deviation increase in the predictor variable. In the subset of variables given in Table **??**, all quantitative sensory testing and psychosocial variables were continuous, and the clinical variables were dichotomous. The few missing values in these predictor variables were (singly) imputed using the expectation-maximization (EM) algorithm. For a more detailed description of the OPPERA study, methodology, and initial findings, see Maixner et al. [2011], Slade et al. [2011], and Bair et al. [2013].

The Bayesian analysis of the data treating all missing indicators as censored and our fully Bayesian method were run with 2 chains of 1500 burn-in iterations and 5000 iterations. (Note: Three models took longer to converge, these were run with 2 chains of 45,000 burn-in iterations and 5000 additional iterations.) Convergence and autocorrelation for all models were checked using diagnostic plots. The Bayesian p-values for the OPPERA analysis are calculated per Lin et al. [2017] as $2 \times \min\{\mathrm{P}(\theta > 0|Data), \mathrm{P}(\theta < 0|Data)\}$.

When treating all missing indicators as censored (a naive analysis), being a "current smoker", as opposed to a "never smoker", is not significantly associated with first-onset TMD, but more substantial evidence for an association is apparent when using a multiple imputation and fully Bayesian approach. However, the hazard ratio for being a former smoker is attenuated after using the Bayesian approach, leading to a much weaker association with first-onset TMD. In addition, after Bayesian parameter estimation, the hazard ratio for the history of five respiratory conditions decreased enough that having a history of five respiratory conditions no longer gives substantial evidence of being associated with first-onset TMD.

All psychosocial variables listed in Table 3 are significant after applying the fully Bayesian method, which is consistent with the methods compared. Yet, hazard rations for all psychosocial

variables slightly increased after applying the fully Bayesian method, as compared to multiple imputation.

A similar effect was seen in the Quantitative Sensory Testing (QST) variables. Of particular interest, Brownstein et al. [2015] found that the three variables related to pressure pain threshold became weakly (or not) associated with first-onset TMD after applying their multiple imputation method, as compared to the *ad hoc* method of treating all missing censoring indicators as censored. After implementing the fully Bayesian method, we found that the pressure pain threshold variables for temporalis, masseter, and TM joint all give substantial evidence of being associated with first-onset TMD.

3.6 Discussion

We have developed a fully Bayesian method for the analysis of time-to-event data with missing censoring indicators. Our method can be implemented using readily available software such as WinBUGS, Proc MCMC, or JAGS. The ability to use these standard software packages is highly attractive given both the reliability and available tools (such as convergence diagnostics) that these existing software programs provide. An unbiased method for handling data with missing censoring indicators is crucial in studies where failure status is not instantaneously recorded, a setting which may lead to missing failure indicators. This study framework is common for diseases that are difficult or expensive to diagnose, such as TMD.

Our method requires that the missing data be, at minimum, MAR. Bair et al.[2013] showed that there were no significant differences between subjects who did attend their clinical examinations versus those who did not, with respect to both demographic variables and risk factors for TMD measured by the questionnaire. This suggests that the MAR assumption is appropriate for the OPPERA data.

In each simulation scenario under the MAR assumption, our method produced no significant bias, as well as the either the narrowest or very close to the narrowest credible interval. When using a fully Bayesian method, the average posterior variance of the estimated regression coefficient is slightly greater than or approximately equal to the mean squared error, showing that the proposed method does not underestimate the variance of the regression coefficient.

In the application of our fully Bayesian method to the OPPERA data, it is evident that some of the hazard ratios associated with some variables were noticeably different than the estimated hazard ratios after using both multiple imputation, and treating all missing censoring indicators as censored. Precise calculation of the hazard ratios associated with putative risk factors for TMD is important, as the results of OPPERA may standardize future orofacial pain literature and the standard of patient orofacial care. Therefore, we suggest that the best estimation of the regression parameters in a Cox model in the presence of missing failure indicators is done via a fully Bayesian model.

Table 3.1: Simulation Results for MAR

True β_j	Inference Method	Bias	Average SE	$\sqrt{\text{MSE}}$	Width	Coverage
-0.5	Full data	0.0010	0.0783	0.0717	0.3065	0.980
	Complete case	-0.0375	0.0958	0.1011	0.3752	0.952
	Treat missing as censored	0.1285	0.1397	0.1822	0.5163	0.871
	Treat missing as failures	-0.0018	0.1006	0.0992	0.3601	0.968
	Cook and Kosorok	-0.0025	0.0913	0.0937	0.3698	0.980
	Multiple Imputation	-0.0012	0.0911	0.1008	0.3559	0.972
	Fully Bayesian	-0.0010	0.0941	0.0856	0.3571	0.978
-1.5	Full data	0.0018	0.1346	0.1329	0.5271	0.976
	Complete case	-0.1779	0.1692	0.2521	0.6629	0.820
	Treat missing as censored	0.1628	0.2416	0.2931	0.9488	0.869
	Treat missing as failures	0.0424	0.1728	0.1931	0.7177	0.969
	Cook and Kosorok	-0.0115	0.1951	0.2181	0.7715	0.936
	Multiple Imputation	-0.0031	0.1804	0.1927	0.7079	0.960
	Fully Bayesian	-0.0023	0.1661	0.1604	0.6509	0.961
-3	Full data	-0.0185	0.2978	0.2857	1.1655	0.975
	Complete case	-0.3268	0.3746	0.4776	1.4644	0.745
	Treat missing as censored	-0.2648	0.4899	0.5387	2.1412	0.858
	Treat missing as failures	0.1211	0.4128	0.4439	1.7651	0.942
	Cook and Kosorok	-0.0343	0.5179	0.6387	2.2135	0.948
	Multiple Imputation	0.0641	0.4491	0.4890	1.9410	0.954
	Fully Bayesian	0.0239	0.3699	0.3545	1.4271	0.966

Bias is the empirical estimate of the sampling bias of the estimate computed from 250 data replicates; Average SE is the average standard error of the estimator $\hat{\beta}$ from 250 data replicates; $\sqrt{}$(MSE) is the empirical estimate of the square root of the mean squared error from 250 data replicates; Width is the average width of the 95% interval estimates based on 250 data replicates; Coverage is the empirical proportion of the 250 replicates of the interval estimates that contain the true parameter.

Table 3.2: Clinical Results from the Orofacial Pain: Prospective Evaluation and Risk Assessment.

	Treat Missing as Censored				Multiple Imputation				Fully Bayesian			
	HR	LCL	UCL	P	HR	LCL	UCL	P	HR	LCL	UCL	P
Clinical Variable												
Mouth	3.31	1.74	5.89	<0.001	2.35	1.39	3.96	0.002	3.22	1.76	4.26	0.002
Chronic Pain	3.09	2.29	4.37	<0.001	2.36	1.79	3.11	<0.001	3.24	2.38	4.32	<0.001
Respiratory Conditions	1.41	1.02	1.77	0.038	1.44	1.13	1.85	0.004	**1.08**	**0.83**	**1.47**	**0.072**
Smoking: current	1.28	0.89	1.77	0.221	1.48	1.13	1.85	0.005	**1.55**	**1.10**	**1.21**	**0.005**
Smoking: former	1.80	1.25	2.79	0.006	1.70	1.18	2.46	0.005	**1.49**	**0.94**	**2.31**	**0.043**
Right Temporalis	1.81	1.26	2.45	<0.001	1.54	1.18	2.02	0.002	1.75	1.29	2.18	<0.001
Left Temporalis	1.62	1.15	2.21	0.005	1.50	1.13	1.98	0.005	1.55	1.12	2.20	0.004
Right Masseter	1.88	1.26	2.57	<0.001	1.69	1.31	2.17	<0.001	1.66	1.22	1.96	<0.001
Left Masseter	1.68	1.20	2.39	0.001	1.50	1.15	1.97	0.003	1.60	1.18	2.12	0.003

HR: hazard ratio; LCL: lower limit of 95% interval estimate; UCL: upper limit of 95% interval estimate; P: p-value

Table 3.3: Psychosocial Results from the Orofacial Pain: Prospective Evaluation and Risk Assessment.

	Treat Missing as Censored				Multiple Imputation				Fully Bayesian			
	HR	LCL	UCL	P	HR	LCL	UCL	P	HR	LCL	UCL	P
Psychosocial variable												
PILL Global Score	1.49	1.38	1.82	<0.001	1.42	1.29	1.58	< 0.001	1.43	1.31	1.57	<0.001
EPQ-R Neuroticism	1.52	1.36	1.73	<0.001	1.25	1.11	1.42	0.003	1.37	1.18	1.53	<0.001
Trait Anxiety Inventory	1.40	1.19	1.62	<0.001	1.34	1.19	1.52	< 0.001	1.45	1.27	1.65	<0.001
Perceived Stress Scale	1.47	1.24	1.74	<0.001	1.29	1.15	1.44	< 0.001	1.37	1.16	1.62	<0.001
SCL 90R Somatization	1.43	1.33	1.64	<0.001	1.40	1.29	1.51	< 0.001	1.46	1.34	1.57	<0.001

HR: hazard ratio; LCL: lower limit of 95% interval estimate; UCL: upper limit of 95% interval estimate; P: p-value; PILL: Pennebaker Inventory of Limbic; Languidness; EPQ: Eysenck Personality Questionnaire;SCLR-90R: Symptom Checklist-90, Revised.

Table 3.4: Quantitative and Sensory Testing Results from the Orofacial Pain: Prospective Evaluation and Risk Assessment.

	Treat Missing as Censored				Multiple Imputation				Fully Bayesian			
	HR	LCL	UCL	P	HR	LCL	UCL	P	HR	LCL	UCL	P
QST variable												
Pressure: temporalis	1.25	1.06	1.59	0.007	1.14	1.00	1.31	0.047	**1.27**	**1.07**	**1.55**	< 0.001
Pressure: masseter	1.17	1.02	1.42	0.022	1.14	0.99	1.31	0.067	**1.25**	**1.07**	**1.53**	**0.008**
Pressure: TM joint	1.28	1.01	1.45	0.016	1.15	1.01	1.32	0.042	**1.26**	**1.07**	**1.46**	**0.003**
Mechanical pain, 15s:	1.19	1.10	1.42	<0.001	1.15	1.04	1.28	0.007	1.24	1.11	1.39	< 0.001
Mechanical pain, 30s:	1.14	1.06	1.33	0.021	1.12	1.02	1.24	0.024	1.21	1.09	1.33	< 0.001

HR: hazard ratio; LCL: lower limit of 95% interval estimate; UCL: upper limit of 95% interval estimate;P: p-value; QST: Quantitative sensory testing; TM: temporomandibular.

CHAPTER 4

GENERATING SIMULATED INTERVAL CENSORED DATA

4.1 Introduction

As discussed in previous sections, time to event data arise in many fields, particularly in medical or health studies that require periodic follow-up visits. In particular, interval censored data may occur if a patient misses one or more scheduled observation times, and returns to their next visit with a changed event status. We will refer to the last known observation time prior to diagnosis as A, and the first observation time that a changed event status is observed at as B. Even if all patients follow their exact examination schedule, investigators cannot determine the exact time that the event occurred. The only known information is that the event occurred sometime between times A and B. This ideal scenario results in grouped time to event data, wherein the observation for each patient is a subset of non-overlapping intervals. For an overview of methods for grouped time to event data, see Lawless [2015]. In this chapter, we focus on non-grouped interval censored time to event data, meaning that some patients will have overlapping censoring intervals.

When analyzing censored data, a common assumption is that that the censoring mechanism is independent of survival time T. For right censored data, this implies that the censoring time C and T are independent. For interval censored data, independent interval censoring means that the joint distribution of A and B contains no parameters that are involved in the survival function of T. This assumption can be written as

$$\mathrm{P}[T \in [t, t+dt) | A = a, B = b] = P[T \in [t, t+dt)] \tag{4.1}$$

In other words, the assumption of independent interval censoring means that some interval $[A, B)$ doesn't give any other information than that the survival time T is simply bracketed between two observed values [Self and Grossman, 1986, Zhang et al., 2005]. Under this assumption, one does not have to model the censoring function along with the survival function.

4.2 Review of Methods for Generating Interval Censored Data

Data simulation is an important aspect of research. Simulation studies are used to explore the behavior of various estimators, as well as compare the results of statistical procedures under different conditions. In this section, we review some methods for simulating interval censored data. We denote T as the failure time random variable following some specified distribution $F(t)$. We want to generate censoring intervals $[A, B)$ such that the censoring occurs non-informatively. Thus, the conditional distribution of A and B given T must satisfy (4.1).

The most naive way of simulating a censoring interval is to define the upper and lower bounds as $A_i = T_i - U_i^{(1)}$ and $B_i = T_i + U_i^{(2)}$, where U_i^1 and U_i^2 are independent, uniformly distributed random variables in the interval $(0, C)$. This method does not satisfy the condition of non-informative censoring.

Many interval censored data occur in the context of medical studies, where a subject is seen for periodic follow-up times until the event of interest has occurred. Given that the subjects are not continuously monitored, the only known information are the time of the last inspection at which the event had not yet occurred, and the time of the first inspection that the event was diagnosed at. Therefore, many simulation methods model inspection times as well as the failure times and the upper and lower boundaries of the censoring interval.

Calle and Gomez [2005] mimic a longitudinal study with periodic follow-up visits, while taking into account that a patient may miss some scheduled appointments. They assume that there are N potential inspection times $a_j, j = 0, 1, \ldots, N$. The probability that a patient completes each of these scheduled appointments is denoted as p. For each patient i, the censoring interval $[A_i, B_i)$ is constructed by defining $A_i = \max\{a_j : a_j < T_i; \delta_j^i = 1\}$ and $B_i = \min\{a_j : a_j \geq T_i; \delta_j^i = 1\}$, where δ_j^i is the indicator of whether the inspection at time a_j occurred ($\delta_j^i = 1$) or was missed ($\delta_j^i = 0$). Different values of inspection appearance probability p will lead to differing widths of the censoring intervals. If a subject did not experience the event of interest prior to the end of the study, $B_i = \infty$, meaning that patient was right-censored.

Another method of mimicking a longitudinal study is given by the model of Schick and Yu [2000]. In this simulation method, they give a set of examination times $N_{ai}, a = 1, \ldots \tau_i$, where N_{ai} are the sum of the independent follow-up times, $N_{ai} = \sum_{b=1}^{a-1} \zeta_{bi}$. For each subject, the number of examination times must satisfy the criteria that $\tau_i = \sup\{a \geq 1 : \sum_{b=1}^{a-1} \zeta_{bi} \leq \tau_i\}$, where τ is the

length of the study. The observed censoring intervals are defined as $A_i = \max\{N_{ai} : N_{ai} < T_i\}$ and $B_i = \min\{N_{ai} : N_{ai} \geq T_i\}$. The length of the observed censoring intervals are controlled by $E(\zeta_{bi})$, and the percentage of right-censored vs. interval-censored observations is controlled by τ_i.

Lawless and Babineau [2006] give an overview of the generation of interval-censored data. They break down the simulation of interval censored data into two prescriptions. In the more naive prescription, inspection times and failure times are independently generated, and the lower and upper observed censoring boundaries are determined from the simulated failure times. This method requires four steps:

1. Generating the failure times T_i from some specified distribution $F(t)$,
2. Generating the inspections N_i for each subject i,
3. Determining the observed lower bound A_i,
4. Determining the observed upper bound B_i

Both of the simulation methods described above by Calle and Gomez and Schick and Yu follow this general outline, in which it is necessary to simulate the individual inspection times for each subject.

To overcome the need for simulating individual inspection times for each subject, Lawless and Babineau use the concept of adjacency. Two potential inspection times N_a and N_b, given $N_a < N_b$ for a subject are said to be adjacent if that subject has an inspection at N_a and N_b, but not at any times in between. Then, for each subject i, the interval censored bounds $(A_i, B_i]$ are selected from the set $\{(N_a, N_b, 0 \leq a \leq b\}$, where N_a, N_b are chosen with probability g_{ab}

$$g_{ab} = \mathrm{P}(A = a, B = b | T \in (a, b]). \tag{4.2}$$

This method still requires the simulation of all possible inspection times for each subject.

In section 4.3, we propose a method for simulating interval censored data based on a Poisson process. Our method of data simulation requires only three distinct steps and does not require the generation of subject specific inspection times, which makes is computationally more efficient than currently available methods. We show that data simulated using our method will fulfill the necessary independence condition outlined in (4.1). We give the algorithmic steps for our simulation method,

and we give a brief simulation study showing the increase in computational efficiency compared a four step simulation method. In section 4.5, we show how our method for data simulation can easily be extended to a simulate data from a non-homogeneous Poisson process. We introduce a dataset that we propose fulfills the conditions of a non-homogeneous Poisson process, and analyze that data, followed by a discussion in section 4.6.

4.3 Simulating Interval-Censored Data

4.3.1 Poisson Processes

A Poisson process is a counting process in which events occur continuously and independently of one another. Let $N(t)$ denote the number of events observed to occur in time interval $(0, t]$. The counting process $\{N(t), t \geq 0\}$ is a homogeneous Poisson process with fixed event rate $\gamma > 0$ if the following conditions hold:

1. $N(0) = 0$
2. The number of events $N(t)$ in non-overlapping intervals are independent
3. The number of events in any time interval depend only on the length of the interval

We say that a homogeneous Poisson process has stationary increments, meaning that the distribution of the number of arrivals depends only on the width of the interval, and not on the bounds of the interval itself. Following this definition of a Poisson process, we know that the number of events $N(t)$ in any interval of length $\tau > 0$ follows a Poisson$(\gamma\tau)$ distribution. The time of the m^{th} inspection is a Gamma(m, γ) random variable, where γ is the fixed event rate. Lastly, the waiting times W will be independent Exp(γ) random variables. In Figure 4.1 below, we give a depiction of the process.

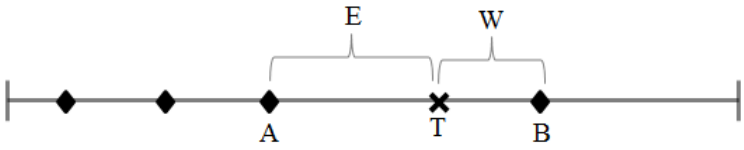

Figure 4.1: Depiction of elapsed and waiting time in relation to inspection times and failure time in interval censored data.

We let E = the elapsed time, or the time from the last observed inspection time A to the subject's failure time T, and W = waiting time, or the time from the subject's failure time T to the next observed inspection time B.

The lower boundary of the observed interval can be 0, meaning that no inspections have taken place before the patient fails at their first follow-up appointment. In this scenario, the elapsed time E will be equivalent to the failure time T. Otherwise, E will take on some value u between 0 and T.

$$E_i = \begin{cases} T & \text{with probability } \exp(-\gamma T) \\ u & \text{with density } \gamma \exp(-\gamma u). \end{cases} \tag{4.3}$$

The lower bound A of the observed censoring intervals can be generated as $A_i = T_i - Ei$. The waiting times W_i are exponentially distributed with rate γ. We generate the upper bound B of the observed censoring intervals as $B_i = T_i + W_i$.

Our method for simulating interval censored data requires only three steps:

1. Generating the failure times T_i from some specified distribution $F(t)$,
2. Generating the elapsed time E_i, which will give the observed lower bound as $A_i = T_i - E_i$
3. Generating the waiting time W_i, which will give the observed upper bound as $B_i = T_i + W_i$

Because this method does not require the simulation of inspections for all subjects at multiple time-points, it is less computationally intensive than currently available methods for the simulation of interval censored data. In addition, we can show that our method will produce independent interval censored data, which is an important assumption made when analyzing this type of data [Little and Rubin, 2002]. This proof is shown in Appendix A.3.

4.3.2 Simulation Study

We are interested in the increase in computational efficiency when simulating interval censored data via our proposed method that does not require simulating individual event times and outcomes for all patients. We will compare the efficiency of this method to a more traditional four step simulation method in which the number of inspection times prior to the lower boundary of the censoring interval is generated.

For this simulation study, we choose to use a more general form of the hazard function, as given by the Weibull distribution. This distribution is characterized by scale and shape parameters λ

and ν, respectively. We note that exponentially distributed survival times (such as the simulation study in Chapter 3) are a special case of the Weibull distribution where $\nu = 1$. The hazard function of a Cox model with the baseline hazard of a Weibull distribution is

$$h(t|Z) = \lambda \exp(\beta' Z)\nu t^{\nu-1} \tag{4.4}$$

The hazard function will increase over time t for $\nu > 1$, and will decrease monotonically for $0 < \nu < 1$. In this simulation study, survival times for 500 subjects were simulated under a proportional hazards model, such that the survival time for each subject is distributed according to (4.4), with baseline hazard $\lambda_0(t) = 1$ and fixed shape parameter $\nu = 0.5$. A single baseline covariate Z_i is generated from a normal distribution with mean 0 and unit variance. This gives the failure times T_i, conditioned on covariate Z_i, to be Weibull distributed with a hazard $\exp(\beta' Z)0.5t^{-0.5}$, where $\beta \in \{-0.5, -1.5, -3\}$.

We let $\gamma = 0.75$ be the rate of examinations per unit of study time. For example, if the unit of study time is months, the chosen rate of $\gamma = 0.75$ is equivalent to a subject being scheduled to be seen every 1.5 months.

For our proposed three-step simulation method, the data will be simulated as follows: First, generate the true failure time T_i for each subject from a Weibull distribution. Then, for each subject i, we let the lower bound of the observed interval $A = T_i - E_i$, where E_i is generated as per (4.3). Lastly, we simulate the upper bound B of the interval for each subject, such that $B_i = T_i + W_i$, where for each subject i, $W \sim \text{Expo}(\gamma)$.

We will compare the computational time from the above three-step method to a four-step simulation method also based on a Poisson process, where the data is simulated as follows: First, the true failure times T_i are generated from a Weibull distribution for each subject i. We simulate the number of inspection times N_i for each subject i, such that $N(T_i) \sim Poisson(\gamma T_i)$. To find the lower bound of the censoring interval A_i, we let $A_i|N(T) = m \sim \max(U_1, \ldots, U_m)$, where $U_j \sim Uniform(0, T_i)$. Lastly, the upper bound of the censoring interval B_i is simulated for each subject, such that $B_i = T_i + W_i$, where for each subject i, $W \sim \text{Expo}(\gamma)$.

Table 4.1: Interval Censored Data Simulation Results

True β_j	Simulation Method	Run Time (sec.)
-0.5	4-step	5.334
	3-step	3.843
-1.5	4-step	183.571
	3-step	5.652
-3	4-step	920.940
	3-step	8.509

As expected, the three step simulation method is considerably faster for all levels of β as compared to a more traditional four step simulation method. This computational efficiency is particularly important when performing large scale simulation studies that may be necessary in a medical setting. For clinical studies, it is often of interest to have simulated data that mirrors the actual study of interest, which may include thousands of patients. In order to achieve the goals of a simulation study, which may include examining the behavior of various estimators, comparing the results of statistical procedures under different conditions, etc., it may be necessary to run over a thousand replicates. With even larger sample size and even more replicates than shown in our study, the increase in computational efficiency when simulating interval censored data using our three-step method versus a traditional four step method will be even more apparent.

4.4 Extension to Non-Homogeneous Poisson Processes

As with a standard Poisson process, we wish to model the number of arrivals $N(t)$ during some time interval $(A, B]$. However, we do not want to make the assumption of stationary increments. This means that the arrival rate (γ) depends on the time-points A and B. Therefore, the time dependent arrival rate $\gamma(t)$ is a non-negative, integrable function. The counting process $\{N(t), t \geq 0\}$ can be called a Non-Homogeneous Poisson Process (NHPP) with rate $\gamma(t)$ if the following conditions are held:

1. $N(0) = 0$

2. $N(t)$ has independent increments

3. For each $0 \leq t_1 < t_2 < \ldots < t_m, N(t_1), N(t_2) - N(t_1), \ldots N(t_m) - N(t_{m-1})$ are independent random variables

In section 4.2, we outlined the steps to generate observed interval censoring bounds for a homogeneous Poisson process. This can can be extended to the more general case of a NHPP, wherein the assumption of a constant inspection rate is relaxed. We can still stimulate data for this scenario using the same three step procedure as for the homogeneous Poisson process simulated data. First, we denote the time specific inspection rate $\Gamma(t) = \int_0^t \gamma(t)dt$. The failure times T_i are again generated from some specified distribution $F(t)$. The elapsed time will be

$$E_i = \begin{cases} T & \text{with probability } \exp\{-\Gamma(T_i)\} \\ u & \text{with density } \gamma(T_i - u)\exp\{-(\Gamma(T_i) - \Gamma(T_i - u)\}. \end{cases}$$

The waiting time will be generated from the following distribution

$$W_i = \gamma(T_i + u) \times \exp(-\Gamma(T_i + u) - \Gamma(T_i))$$

4.4.1 Prostate Cancer Data

Prostate cancer studies commonly yield interval censored data. Following surgery to remover the cancerous prostate, the Prostate-Specific Antigen (PSA) is a widely accepted marker of prostate cancer recurrence [Freedland et al., 2005, Pound et al., Schroek et al., Kowalczyk et al.], as PSA level in the blood is undetectable in absence of metastatic disease. PSA recurrence is defined as the time following surgery that a patient's PSA level is greater than 0.2ng/ml. In a typical prostate cancer study, patients will have PSA blood tests at fixed intervals following surgery. The exact time of prostate cancer recurrence is interval censored, as the patient is not continuously monitored, and we only know whether a patient has PSA recurrence between two consecutive post surgery visits.

Our data is from a German prostate cancer study that follows 600 patients after they receive a minimally invasive procedure to remove the prostate. We are interested in the effect of baseline predictors on time to prostate cancer recurrence (time from surgery to PSA recurrence). The baseline characteristics of interest include age at surgery, number of positive cores, and type of surgery (0 = robotic DaVinci, 1 = RRP), surgical margin (0 = negative, 1 = positive), pT stage

(0 = has not spread, 1 = spread), and Gleason score (a measure of cancer aggressiveness based on how cancer cells are arranged in the prostate where 0 = less aggressive, 1 = more aggressive).

This data consists of a mixture of interval censored and right censored observations. Patients are right censored if they do not have a prostate cancer recurrence by the end of the study monitoring time. Approximately 20% of patients have interval censored survival times, leaving about 80% of patients that have not yet had a prostate cancer recurrence by the end of the study monitoring time.

We want to model the patient inspection times as a Poisson process. However, it is reasonable to assume that this process does not have stationary increments, or that the rate of the Poisson process changes over time. Thus, we assume that patient inspection times are a non-homogeneous Poisson process.

Table 4.2: Prostate Cancer Data Analysis

	HR	LCL	UCL	P
Age at Surgery	1.011	0.988	1.036	0.439
Number of Positive Cores	1.083	1.0258	1.140	0.022
Surgery Type	0.863	0.578	1.147	0.394
Surgical Margin	1.701	1.384	2.018	0.006
pT Stage	1.801	1.502	2.114	0.001
Gleason Score	2.084	1.314	2.854	0.117

HR: hazard ratio; LCL: lower limit of 95% interval estimate;UCL: upper limit of 95% interval estimate; P: p-value

4.5 Discussion

Simulation studies play an important part of data analysis in every field. These studies require the generation of simulated data that fulfills whatever assumptions researchers are making when analyzing the data. In particular, we are concerned with the simulated data fulfilling the commonly made assumption of independent censoring. In this assumption, interval boundaries give no more information about the survival time T than the fact that the survival time is contained within this interval.

We propose a new method for simulating observed boundaries for interval censored data using properties of Poisson processes. We show that our method fulfills the assumption of independent censoring, which is an assumption that is widely assumed, but rarely checked. In addition to fulfilling this important assumption, our method is more efficient and less complex, in that our simulation method does not require generating all inspection times and outcomes for all subjects. Therefore, our method requires only three steps as opposed to the usual four.

It is not always reasonable to make the assumption that the inspection rate is stationary. In many real world scenarios, it is more reasonable to assume that the inspection rate would change over time, meaning that the rate parameter in the Poisson process is time dependent. We give a solution for the simulation of data under the relaxed condition of the inspection rate changing over time, versus the more stringent assumption that the inspection process is stationary. This makes our methodology for simulating observed interval censoring boundaries even more applicable.

APPENDIX A

DERIVATION AND PROOFS FROM CHAPTER 3

A.1 Derivation of (3.5) in 3.2

The quantity that Brownstein et al. [2015] and Cook and Kosorok [2004] are estimating is the probability of a subject failing ($\Delta_i = 1$) given that it has a missing censoring indicator ($\zeta_i = 0$) at the last observation time t, i.e, $\mathrm{P}(\Delta_i = 1|V_i = t, \zeta_i = 0, Z_i, X_i)$. The failure probability can be written in terms of the failure time T and the censoring time C, as

$$\mathrm{P}(T_i = t, C_i > t|V_i = t, \zeta_i = 0, Z_i, X_i).$$

Under the assumption of non-informative censoring [Kalbfleisch and Prentice, 2002], we have

$$\mathrm{P}(T_i = t, C_i > t|V_i = t, Z_i, X_i) = \frac{\mathrm{P}(T_i = t, C_i > t, V_i = t|Z_i, X_i)}{\mathrm{P}(V_i = t|Z_i, X_i)} \tag{A.1}$$

Let $\tilde{S}(t|Z_i, X_i)$ be the survival function of the censoring time C_i, with corresponding hazard $\tilde{h}(t|Z_i, X_i)$. The hazard and survival function are $\lambda(t|Z_i)$ and $S(t|Z_i)$, as defined in Section 2. (A.1) can be written as

$$\begin{aligned}
&= \frac{\lambda(t|Z_i) \times (1 - \tilde{h}(t|Z_i, X_i))}{\lambda(t|Z_i) \times (1 - \tilde{h}(t|Z_i, X_i)) + \tilde{h}(t|Z_i, X_i) \times (1 - \lambda(t|Z_i))} \\
&= \frac{S(t|Z_i) \times \tilde{S}(t|Z_i, X_i) \times \lambda(t|Z_i)}{S(t|Z_i) \times \tilde{S}(t|Z_i, X_i) \times \left\{\lambda(t|Z_i) + \tilde{h}(t|Z_i, X_i)\right\}} \\
&= \frac{1}{1 + \frac{\tilde{h}(t|Z_i, X_i)}{\lambda(t|Z)}} \\
&= \frac{1}{1 + \tilde{h}(t|Z_i, X_i)\frac{\exp(-\beta' Z_i)}{\lambda_0(t)}}.
\end{aligned} \tag{A.2}$$

However, Brownstein et al. [2015] assume that

$$\mathrm{P}(\Delta_i = 1|V_i = t, \zeta_i = 0; Z_i, X_i) = \frac{\exp(\alpha X_i + \gamma Z_i + \eta_i t)}{1 + \exp(\alpha X_i + \gamma Z_i + \eta_i t)} = \frac{1}{1 + \exp(-\alpha X_i - \gamma Z_i - \eta V_i)} \tag{A.3}$$

To ensure (A.2) and (A.3) are equal, additional and highly restrictive assumptions must be placed on both the censoring hazard $\tilde{h}(t|Z_i, X_i)$ and the relationship between the baseline hazard λ_0 and the censoring hazard.

A.2 Proofs of Results 1 and 2 in 3.3

Result 1:

We assume that the censoring indicators are MAR [Rubin, 1976]. Thus,

$$\begin{aligned} &\mathrm{P}(V_i, \Delta_i, \zeta_i = 1|Z_i) \\ &= \mathrm{P}(V_i, \Delta_i|Z_i) \times \mathrm{P}(\zeta_i = 1|V_i, \Delta_i, Z_i) \\ &= \mathrm{P}(V_i, \Delta_i|Z_i) \times (1 - p_i), \\ &\text{where } p_i = \mathrm{P}(\zeta_i = 0|V_i, Z_i). \end{aligned} \tag{A.4}$$

Similarly, we can write

$$\begin{aligned} &\mathrm{P}(V_i, \zeta_i = 0|Z_i) = \mathrm{P}(V_i, \Delta_i = 0, \zeta_i = 0|Z_i) + \mathrm{P}(V_i, \Delta_i = 1, \zeta_i = 0|Z_i) \\ &= \mathrm{P}(V_i, \Delta_i = 0|Z_i) \times \mathrm{P}(\zeta_i = 0|V_i, Z_i) + \mathrm{P}(V_i, \Delta_i = 1|Z_i) \times \mathrm{P}(\zeta_i = 0|V_i, Z_i) \\ &= p_i\big\{\mathrm{P}(V_i, \Delta_i = 0|Z_i) + \mathrm{P}(V_i, \Delta_i = 1|Z_i)\big\}, \end{aligned} \tag{A.5}$$

where $p_i =$ is defined as in (A.4). We can express (A.5) in terms of the survival time C_i and the failure time T_i:

$$\mathrm{P}(V_i, \zeta_i = 0|Z_i) = p_i\big\{\mathrm{P}(T_i > V_i] \times \mathrm{P}[C_i = V_i) + \mathrm{P}(T_i = V_i] \times \mathrm{P}[C_i > V_i)\big\} \tag{A.6}$$

From (A.6), we can see that C_i can't be factored out, and thus the probability of a subject having a missing censoring indicator depends on the censoring distribution.

Result 2:

We want to estimate the probability p_{ij} that subject i with a missing censoring indicator is a failure in I_j, i.e, $\mathrm{P}(\Delta_i = 1|\zeta_i = 0, V_i \in I_j; Z_i, X_i)$.

$$\begin{aligned} &= \frac{\mathrm{P}(\Delta_i = 1, \zeta_i = 0, V_i \in I_j|Z_i, X_i)}{\mathrm{P}(\zeta_i = 0, V_i \in I_j|Z_i, X_i)} \\ &= \frac{\mathrm{P}(\Delta_i = 1, \zeta_i = 0, V_i \in I_j|Z_i, X_i)}{\mathrm{P}(\Delta_i = 1, \zeta_i = 0, V_i \in I_j|Z_i, X_i) + \mathrm{P}(\Delta_i = 0, \zeta_i = 0, V_i \in I_j|Z_i, X_i)} \end{aligned} \tag{A.7}$$

Let $\lambda(V_i|Z_i)$ be the baseline hazard, $S(V_i^-|Z_i)$ be the survival function and $\tilde{h}(V_i|Z_i, X_i)$ be the censoring hazard. Then, (A.7) can be written as

$$= \frac{\lambda(V_i|Z_i) \times S(V_i^-|Z_i)}{\lambda(V_i|Z_i) \times S(V_i^-|Z_i) + S(V_i^-|Z_i) \times (1 - \lambda(V_i|Z_i)) \times \tilde{h}(V_i|Z_i, X_i)}$$

$$= \frac{\lambda(V_i|Z_i)}{\lambda(V_i|Z_i) + [1 - \lambda(V_i|Z_i)]\tilde{h}(V_i|Z_i, X_i)} \tag{A.8}$$

Given that $\lambda(V_i|Z_i) \times \tilde{h}(V_i|Z_i, X_i)$ will be small, we can approximate (A.8) as

$$p_{ij} \approx \frac{\lambda(V_i|Z_i)}{\lambda(V_i|Z_i) + \tilde{h}(V_i|Z_i, X_i)}$$

APPENDIX B

PROOF FROM CHAPTER 4

B.1 Proof of Result in 4.3

To show that our method of simulating interval censored data fulfill the assumption of independent censoring, we need show that the following equation holds.

$$\mathrm{P}[T \in [t, t+dt) | A = a, B = b] = P[T \in [t, t+dt)] \tag{B.1}$$

$$\begin{aligned}
&\mathrm{P}[T \in [t, t+dt) | A = a, B = b] \\
&= \frac{\mathrm{P}[T \in [t \in t+dt), A = a, B = b]}{\mathrm{P}[A = a, B = b]} \\
&= \frac{\mathrm{P}[A = a, B = b | T \in [t, t+dt)] \times \mathrm{P}[T \in [t, t+dt)]}{\mathrm{P}[A = a, B = b]} \\
&= \frac{\int_a^b \lambda e^{-\lambda(t-a)} \lambda e^{-\lambda(b-t)} f(t|x) dt}{\mathrm{P}[A = a, B = b]} \\
&= \frac{\lambda^2 e^{-\lambda(b-a)} \int_a^b f(t|x) dt}{\mathrm{P}[A = a, B = b]}
\end{aligned} \tag{B.2}$$

The denominator in (B.2) can we written as

$$\mathrm{P}[A = a, B = b] = \sum_{m=0}^{\infty} \mathrm{P}[A = a, B = b, N(a^-) = m] \tag{B.3}$$

The probability of being in some interval $(a, b]$ is equal to the probability of being in some interval $(a, b]$ on the $(m+1)^{th}$ inspection, summed over all inspections m.

If there are m inspections prior to time a, then a is the $(m+1)^{th}$ inspection. The inspections follow a gamma distribution with shape and scale parameters $m+1$ and λ, respectively. Thus, denominator in (B.3) is equivalent to

$$\sum_{m=0}^{\infty} \frac{a^m \lambda^{m+1} e^{-\lambda a}}{\Gamma(m+1)} \times e^{-\lambda(b-a)}$$

$$e^{-\lambda(b-a)} e^{-\lambda a} \lambda^2 \sum_{m=0}^{\infty} \frac{(a\lambda)^m}{m!}$$

$$e^{-\lambda(b-a)} \lambda^2 e^{-\lambda a} e^{\lambda a} = \lambda^2 e^{-\lambda(b-a)}$$

Substituting this back into (B.1) gives

$$\frac{\lambda^2 e^{-\lambda(b-a)} \int_a^b f(t|x)dt}{\lambda^2 e^{-\lambda(b-a)}} = \int_a^b f(t|x)dt = \mathrm{P}[T \in [t, t+dt)]$$

Thus, our method of simulating the observed censoring intervals A and B will yield independent interval censored data.

APPENDIX C

IRB APPLICATION AND EXEMPTION FORMS

Human Subjects Application For Full IRB and Expedited Exempt Review

Go back

1. Project Title and Identification

1.1 Project Title

PhD Thesis in Biostatistical Applications

Project is: Dissertation

1.2 Principal Investigator (PI)

Name(Last name, First name MI): sinha, debajyoti	Highest Earned Degree: Doctorate
Mailing Address: [redacted]	Phone Number: [redacted] Fax:
University Department: ARTS AND SCIENCES, DEANS OFFICE	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records: NIH	Occupational Position: Faculty

1.3 Co-Investigators/Research Staff

Name(Last name, First name MI): [redacted]	Highest Earned Degree: Master's Degree
Mailing Address: 4330	Phone Number: [redacted] Fax:
University Department: STATISTICS DEPARTMENT	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records: None	Occupational Position: Student
Name(Last name, First name MI): [redacted]	Highest Earned Degree: Master's Degree
Mailing Address: [redacted]	Phone Number: [redacted] Fax:
University Department: [redacted]	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records:	Occupational Position: Student
Name(Last name, First name MI): [redacted]	Highest Earned Degree: Master's Degree
Mailing Address: [redacted]	Phone Number: [redacted] Fax:
University Department: STATISTICS DEPARTMENT	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records:	Occupational Position: Student
Name(Last name, First name MI): Bunn, Veronica Florida State Un; Co-Investigator	Highest Earned Degree: Master's Degree
Mailing Address: [redacted]	Phone Number: [redacted] Fax:
University Department: STATISTICS DEPARTMENT	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records:	Occupational Position: Student

1.4 Faculty Advisor/Department Chair/Dean Information

Name(Last name, First name MI): niu, xufeng ; Chair	Highest Earned Degree:
Mailing Address: [redacted]	Phone Number: [redacted] Fax:
University Department: STATISTICS DEPARTMENT	Email: [redacted]
The training and education completed in the protection of human subjects or human subjects records:	Occupational Position:

Office of the Vice President for Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 · FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 02/16/2018

To: debajyoti sinha <sinhad@stat.fsu.edu>

Address: Statistics Dept., Room 106F OSB, 105 N Woodword Ave, PO Box 3064330, Tallahassee, FL 32306-4330

Dept.: ARTS AND SCIENCES, DEANS OFFICE

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
PhD Thesis in Biostatistical Applications

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Exempt per 45 CFR § 46.101(b)4 and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 02/15/2019 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the chairman of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: veronica bunn <v.bunn@stat.fsu.edu>, Chair
HSC No. 2018.23156

REFERENCES

Odd O. Aalen and Hakon K. Gjessing. Understanding the shape of the hazard rate: A process point of view. *Statistical Science*, 16(1):1–22, 2001.

Eric Bair, Naomi C. Brownstein, Richard Ohrbach, Joel D. Greenspan, Ronald Dubner, Roger B. Fillingim, et al. Study protocol, sample characteristics, and loss to follow-up: The oppera prospective cohort study. *The Journal of Pain*, 14(12):T2–19, 2013.

Ralf Bender, Thomas Augustin, and Maria Blettner. Generating survival times to simulate cox proportional hazards models. *Statistics in Medicine*, 24(11):1713–1723, 2005.

N.C. Brownstein, Jianwen Cai, Gary Slade, and Eric Bair. Parameter estimation in cox models with missing failure indicators and the oppera study. *Statistics in Medicine*, 34(30):3984—-3996, Apr 2015.

J. Burridge. Empirical bayes analysis of survival time data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 45:65–75, 1981.

M. Luz. Calle and Guadalupe Gomez. A semiparametric hierarchical method for a regression model with an interval-censored covariate. *Australian and New Zealand Journal of Statistics*, 47: 351–364, 2005.

Angelo Canty and Brian Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2016. R package version 1.3-18.

David Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall, Boca Raton, 2015.

Thomas D Cook and Michael R Kosorok. Analysis of time-to-event data with incomplete event adjudication. *Journal of the American Statistical Association*, 99(468):1140–1152, 2004.

D. R. Cox. Partial likelihood. *Biometrika*, 62:543–558, 1975.

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

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

Gregg E. Dinse. Nonparametric estimation for partially-complete time and type of failure data. *Biometrics*, 38:417–431, 1982.

R. L. Dykstra and Purushottam Laud. A bayesian nonparametric approach to reliability. *The Annals of Statistics*, 9:356–367, 1981.

Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

Polly Feigl and Martin Zelen. Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21(4):826–838, 1965.

Thomas S. Ferguson and Eswar G. Phadia. Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 7:163–186, 1979.

Stephen J. Freedland, Elizabeth B. Humphreys, Leslie A. Mangold, M. Eisenberger, F.J. Dorey, P.C. Walsh, and A.W. Partin. Risk of prostate cancer-specific mortality following biochemical recurrence after radical prostatectomy. *Journal of the American Medical Association*, 294(4): 433–439, 2005.

Michael Friedman. Piecewise exponential models with survival data with covariates. *The Annals of Statistics*, 10:101–113, 1982.

Andrew Gelman and Jennifer Hill. Opening windows to the black box. *Journal of Statistical Software*, 40, 2011. R package version 9.15.

Amy H. Herring and Joseph G. Ibrahim. Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302, 2001.

Nils Lid Hjort. Nonparametric bayes estimators on beta processes in models for life history data. *the Annals of Statistics*, 18:1259–1294, 1990.

Joseph G. Ibrahim, Ming-Hui Chen, and Steven MacEachern. Bayesian variable selection for proportional hazards. *Canadian Journal of Statistics*, 27:557–565, 1999.

Joseph G. Ibrahim, Ming-Hui Chen, and Debajyoti Sinha. *Bayesian Survival Analysis*. Springer, New York, 2001.

Joseph G. Ibrahim, Ming-Hui Chen, Stuart R. Lipsitz, and Amy H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–347, 2005.

John D Kalbfleisch. Non-parametric bayesian analysis of survival time data. *Journal of the Royal Statistical Society*, 40:214–221, 1978.

John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 2002.

E.L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association)*, 53(282):457–481, 1958.

Keith J. Kowalczyk, Aaron C. Weinburg, Xiangmei Gu, Hua-Yin Yu, Stuart R. Lipsitz, Stephen B. Williams, and Jim C. Hu. Comparison of outpatient narcotic prescribing patterns after minimally invasive versus retropubic and perineal radical prostatectomy. *The Journal of Urology*, 186), year=2011, pages=1843-1848.

Jerald F. Lawless. *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons, Hoboken, 2015.

Jerald F. Lawless and Denise Babineau. Models for interval censoring and simulation-based inference for lifetime distributions. *Biometrika*, 93:671–686, 2006.

Yan Lin, Stuart R. Lipsitz, Debajyoti Sinha, Garrett Fitzmaurice, and Steven Lipshultz. Exact bayesian p-values for a test of independence in a 2 x 2 contingency table with missing data. *Statistical Methods in Medical Research*, 2017.

Roderick J.A. Little and Donald B. Rubin. *The Statistical Analysis of Interval-censored Failure Time Data*. Wiley Series in Probability and Mathematical Statistics, 2002.

William Maixner, Luda Diatchenko, Ronald Dubner, Roger B. Fillingim, Joel D. Greenspan, Charles Knott, et al. Orofacial pain prospective evaluation and risk assessment study – the oppera study. *The Journal of Pain*, 12(11):T4–11, 2011.

M. Mezzetti and Joseph G. Ibrahim. Bayesian inference for the cox model using correlated gamma process priors. Technical report, Department of Biostatistics, Harvard School of Public Health, 2000.

Charles R. Pound, Alan W. Partin, Mario A. Eisenberger, et al. Natural history of progression after psa elevation following radical prostatectomy. *Journal of the American Medical Association*, 281), year=1999, pages=1591-1597.

German Rodriguez. Parametric survival models. Technical report, Princeton University, 2010.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B. Rubin. Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489, 1996.

Donald B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons, New York, 2004.

Anton Schick and Qiqing Yu. Consistency of the gmle with mixed case interval-censored data. *Scandinavian Journal of Statistics*, 27:45–55, 2000.

Florian R. Schroek, Leon Sun, Stephen J. Freedland, David M. Albala, Vladimir Mouraviev, Thomas J. Polascik, and Judd W. Moul. Comparison of prostate-specific antigen recurrence-free survival in a contemporary cohort of patients undergoing either radical retropubic or robot-assisted laparoscopic radical prostatectomy. *British Journal of Urology International*, 102), year=2008, pages=28-32.

Steven G. Self and Elizabeth A. Grossman. Linear rank tests for interval-censored data with applications to pcb levels in adipose tissue of transform repair workers. *Biometrics*, 42:521–530, 1986.

Gary D Slade, Eric Bair, Kunthel By, Flora Mulkey, Cristina Baraian, Rebecca Rothwell, et al. Study methods, recruitment, sociodemographic findings, and demographic representativeness in the oppera study. *The Journal of Pain*, 12(11):T12–T26, 2011.

Sibylle Sturtz, Uwe Ligges, and Andrew Gelman. R2winbugs: A package for running winbugs from r. *Journal of Statistical Software*, 12(3):1–16, 2005.

Jianguo Sun. *Statistical Analysis with Missing Data*. Springer, New York, 2006.

V. Susarla and J. Van Ryzin. Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71:897–902, 1976.

T Therneau. *A Package for Survival Analysis in S*, 2015. R package version 2.38.

Zhigang Zhang, Jianguo Sun, and Liuquan Sun. Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, 24:1399–1407, 2005.

Calvin Zippin and Peter Armitage. Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter. *Biometrics*, 22:665–672, 1966.

BIOGRAPHICAL SKETCH

Veronica Bunn was born in Lexington, Kentucky in 1993. She completed Bachelor of Science degrees in Mathematics and Economics at University of Kentucky in 2013. In 2015, she received an Master of Science in Applied Statistics from Florida State University and will presumably receive a Ph.D. in Biostatistics from Florida State University in the summer of 2018. While pursuing her graduate work, she completed internships at Pacific Northwest National Laboratory, Sanofi Genzyme, and Takeda Pharmaceuticals. After finishing her Ph.D., Veronica will join Takeda Pharmaceuticals as a Senior Statistician in Cambridge, MA.