τ-Inflated Beta Regression Model for Censored Time-to-Event and Recurrent Event Data

by

Yizhuo Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2023

Doctoral Committee:

Professor Susan Murray, Chair
Professor Jeremy M G Taylor
Assistant Professor Nabihah Tayob
Professor Alexander Tsodikov
Assistant Professor Zhenke Wu

Yizhuo Wang

yizhuow@umich.edu

ORCID iD: 0000-0003-3656-1267

ACKNOWLEDGMENTS

I would like to express my sincerest gratitude to my PhD advisor, Dr. Susan Murray, for her invaluable guidance, support, and encouragement throughout my doctoral journey. She has consistently demonstrated exceptional patience and provided insightful solutions that have helped me navigate the challenges I encountered along the way. The past couple of years have been especially challenging due to the COVID-19 pandemic that has impacted our lives in numerous ways. Despite the pandemic-induced disruptions and uncertainties, my advisor has been a constant source of motivation and support, going above and beyond her role to ensure my well-being and success. I really appreciate my advisor's sense of humor and joy, especially when things get tough. Even though I don't always get her jokes, she always manages to make me laugh and brighten up my day. I feel incredibly fortunate to have had such an unwaveringly supportive and dedicated mentor, committed to my academic and personal growth.

Graduate school has been a journey filled with challenges and successes. I am deeply grateful to my family for their unwavering support during the tough times and for sharing in my accomplishments. My parents, Lin and Zhimin, have been a constant source of encouragement and pride for their daughter. I would also like to thank my loving husband, Zhanghao, for his support, love, and encouragement throughout my doctoral journey. He has been my rock and biggest cheerleader, always there to listen and provide the emotional support I needed during the most challenging moments of my journey. His wonderful sense of humor that has helped me maintain a positive attitude, even during the most stressful times. I could not have achieved this milestone without the love, support, and encouragement of my family. I am incredibly blessed to have them in my life and will forever cherish their unwavering support and love.

I am deeply grateful to all our collaborators in the pulmonary division at the University of Michigan Hospital, who have played an instrumental role in my research journey. I extend my sincere thanks to Dr. Jeffrey L. Curtis and Dr. David O'Dwyer for their patience and guidance during my early days as a GSRA. I am also thankful to Dr. Meilan Han for her exceptional mentorship and support throughout my graduate studies. Collaborating with Dr. Shijing Jia has been a great pleasure, and I am proud of the research questions we have tackled together. I would like to express my sincere appreciation to my dissertation committee members, Drs. Jeremy M G Taylor, Nabihah Tayob, Alexander Tsodikov, and Zhenke Wu, for their unwavering support and

responsiveness to my requests. I am grateful to all the faculty and staff members in the Biostatistics Department for their guidance, assistance, and support. Without their wisdom and expertise, my research journey would not have been possible. Lastly, I want to give a special shout-out to all of my cutest friends, Xiayi Li, Kexin Jiang, Yaoxuan Xia, Yujing Zhang, Yuan Jin, Jiongming Wang, Yuming Sun and Chang Xu, for the unforgettable memories we shared in Ann Arbor. I am forever grateful to my beloved University of Michigan for its rich academic and cultural environment, which has left an indelible mark on my life. Go BLUE!

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

ABSTRACT

τ-Restricted Mean Survival Time (τ-RMST) models are popular for modeling censored time-to-event data. One data feature that has not received adequate attention in τ-RMST literature is the point mass of events at τ for the τ-restricted event time, min(τ,T). Individuals who remain event-free during the restricted follow-up time are certainly of interest when evaluating impacts of factors. This dissertation introduces a novel model framework that takes advantage of the point mass to improve the precision of the estimation of RMST and understanding of the association of the predictors and τ-restricted times-to-event. We explore three different settings within this dissertation and demonstrate that our proposed model provides statistical advantages in each of these settings after effectively handling censoring as part of model fitting process.

In Chapter 2, we leverage mixture distribution ideas from cure rate model literature, viewing the study cohort as a mixture of patients who experience the event versus do not experience the event during the restricted follow-up time. We propose a τ-inflated beta regression (τ-IBR) model using joint logistic and beta regression to explore associations between predictors and a potentially censored time-to-event in these two sub-populations and improve the precision of RMST estimation. To deal with censored nature of the data and fit our proposed models we develop both expectation-maximization (EM) and multiple imputation (MI) approaches. Simulations indicate excellent performance of the τ-IBR model(s), and higher precision of corresponding τ-RMST estimates compared to the traditional τ-RMST model, in independent and dependent censoring setting.

In Chapter 3, we generalize the τ-IBR model framework proposed in Chapter 2 to the setting with potentially censored recurrent event times. We first restructure recurrent-event data into a censored longitudinal data structure of τ-restricted-times-to-first-event observed in τ-duration follow-up windows initiated at regularly-spaced intervals. Models used to analyze single restricted event times in Chapter 2 are then applied to the censored longitudinal dataset of times-to-first-event; a generalized estimating equation (GEE) approach is used to address the correlated nature of the τ-restricted times-to-first-event across the follow-up windows. Multiple imputation (MI) and expectation-solution (ES) approaches appropriate for censored data are developed as part of the model fitting process. Simulations indicate good statistical performance of the proposed τ-IBR approach to modeling censored recurrent event data.

In Chapter 4 we extend the τ-inflated beta model to the setting with dependently censored data. This chapter is motivated by lung allocation waitlist survival data, where waitlist deaths are dependently censored as more urgent patients are selected for transplantation. An inverse probability of censoring algorithm is incorporated into the multiple imputation of the censored waitlist death times. Ideas from Chapters 2 and 3 are used to develop an appropriate τ-restricted inflated beta regression model that allow for improved 1-year RMST estimation, which is an essential component of the current lung allocation score that measures waitlist urgency.

CHAPTER 1

Introduction

Restricted mean survival time methodology for single and recurrent censored event data are popular data analysis techniques. In this dissertation work, we introduce τ-inflated beta regression (τ-IBR) model techniques for these settings. These models consider the study cohort as a mixture of patients who either experience the event or do not experience the event during the restricted follow-up time, and it models these two components separately. We leverage mixture distribution ideas from zero/one inflated beta model literature [37, 73] and cure rate model literature [8, 7, 11, 17, 74, 56, 63]. However, these methods have never been applied to estimate and perform inference on restricted mean survival times (RMST).

For a single time-to-event, the outcome min(τ, T) is a restricted event time, where T and τ are the time-to-event and follow-up duration of interest, respectively. The follow-up period may be restricted to an interesting time period, for example 1 or 5 years of follow-up, or follow-up may be restricted due to funding. When evaluating τ-restricted times-to-event, it is common to observe many individuals that remain event-free throughout the restricted follow-up period, resulting in frequent occurrences of min(τ, T) = τ. Risk profiles (covariates) associated with observing min(τ, T) = τ may be quite different from risk profiles associated with min(τ, T) given $\min(\tau, T) < \tau$, which is often overlooked in models of these restricted event times. The cure rate model literature handles such point masses effectively by utilizing a logistic regression approach to model the proportion of cured patients in the data, along with a proportional hazards or accelerated failure time model for event times. However, we were motivated to pursue this particular model for use with pulmonary research data, in which the concept of cure is not possible. Our τ-inflated nomenclature makes the goals of our analysis more transparent and very explicitly tied to a follow-up window duration that may be varied, if desired. Moreover, our modeling framework is more closely tied to restricted mean regression models, which have not been used for cure rate modeling to date. This nomenclature also aligns with one-inflated beta regression models proposed in settings without censoring [37, 73], another inspiration for our approach. The beta distribution, rescaled to cover the range from zero to τ, is particularly flexible and suitable for modeling events constrained to a follow-up window of fixed length. By incorporating beta regression to model event times for "unsusceptible" patients instead of relying on proportional hazards or accelerated failure time models, we can easily estimate and perform inference on restricted mean survival times, which served as our initial starting point when developing our proposed model.

In order to develop methodologies for censored single or recurrent event data, it is necessary to address the censored nature of the data. In this dissertation, we employ three different approaches where appropriate: the expectation-maximization (EM) algorithm, the expectation-solution (ES) algorithm, and the multiple imputation (MI) approach. The EM and ES algorithms are iterative methods that rely on likelihood and estimating equations, respectively, to estimate the parameters in statistical models that depend on unobserved latent variables. The ES algorithm is a variation of the EM algorithm for censored longitudinal data. Different from the EM and ES algorithms, the MI approach replaces censored outcomes with imputes to obtain multiple complete datasets that can be analyzed using standard methods. One advantage of the MI approach over the EM and ES algorithms is that it allows for additional analyses using uncensored data methods based on the multiply imputed datasets. In Chapters 2 and 3, where censoring is assumed to be independent to the restricted event times, the only parametric element of the MI procedure is defining a risk set of similar individuals to the censored individual being imputed, where risk set selection differs in each of the chapters in this dissertation. Once the risk set is defined, we use an inverse transform (IT) method of imputing from the risk set that is entirely nonparametric, a technique that has been developed and modified by many authors [30, 57, 20, 21, 70, 60]. The general idea of the IT method is to obtain imputes by sampling from the non-parametric Kaplan-Meier survival estimate based on patients in the risk set. The IT method slightly differs in Chapter 4 compared to Chapters 2 and 3, where we use the inverse probability of censoring (IPCW) adjusted Kaplan-Meier to obtain a consistent estimator of the survival function in the presence of dependent censoring.

In Chapters 3 and 4, we repurpose the data into a censored longitudinal data structure for the analysis of censored recurrent event and single time-to-event data. Through conducting this preprocessing step, we extend our proposed τ-IBR model framework to recurrent events setting in Chapter 3, while in Chapter 4, we enhance the estimation of τ-RMST by incorporating information beyond the initial τ-length follow-up window. This alternative censored longitudinal data structure was first proposed by Tayob and Murray [58], and further developed in other literature [59, 60, 67, 69, 68]. For each individual, the recurrent event data structure is transformed to a series of τ-length follow-up windows initiated at regularly spaced follow-up times $t \in t_1, ..., t_b$. The newly formatted longitudinal outcomes of interest for each individual are the τ-restricted time to the first event following each pre-specified time t. By converting traditional recurrent event data into a censored longitudinal data structure, we are able to take advantage of standard longitudinal data analysis methods such as GEE to analyze the data and account for correlations within individuals. In addition, pre-specified follow-up windows also avoid dependent censoring issues commonly observed in the analysis of recurrent events data, particularly in correlated gap-time derived data structures[3, 29, 54]. By converting single times-to-event into a censored longitudinal data structure, we attempt to combine information across the multiple follow-up windows to improve modeling efficiency. In settings where a considerable proportion of patients experience events beyond the initial τ-restricted period with updated risk factors, focusing only on information from that period results in an unfortunate waste of statistical information. Using follow-up windows that span a longer duration improves the efficiency of estimating τ-RMST. The choice of the length of follow-up windows τ depends on clinical interest and the spacing between adjacent windows is determined by study design and modeling efficacy, which are further explored in the literature [67, 58], as well as in Chapters 3 and 4 of this dissertation.

In Chapter 2, we establish the fundamental model of our proposed τ-inflated beta regression (τ-IBR) approach. Our proposed τ-IBR model explicitly models the point mass and continuous components of min(τ, T) by decomposing min(τ, T) into $\tau[I(T \geq \tau) + (T/\tau)I(T < \tau)]$ and modeling the mean of this latter expression using joint logistic and beta regression models. In addition to parameter estimation of τ-IBR model, we also provide the formula for estimating the τ-RMST and its variance estimate. In simulation studies, we evaluate finite sample properties of our proposed model's parameter estimates and compare the τ-RMST estimated by our proposed model to the τ-RMST estimated by the standard τ-RMST model. Attractive τ-IBR model results are found regardless of estimation approach (EM, MI) or censoring mechanism (none, independent, dependent). We also demonstrate how to apply the proposed τ-IBR model to data from a randomized clinical trial assessing the efficacy of azithromycin treatment in reducing exacerbations in patients with chronic obstructive pulmonary disease (COPD).

In Chapter 3, we take a fresh look at the τ-IBR modeling framework applied to recurrent events data. We are motivated by the idea that many individuals in the Azithromycin in COPD study are at an earlier disease stage that makes them less susceptible to recurrent exacerbations during follow-up, so that τ-RMST = τ is often observed. These are the same types of individuals with a tendency to have zero recurrent event counts, which inspired the use of zero-inflated count models. Our modeling framework for the analysis of censored recurrent event data that is particularly useful when there is a mixture of (a) individuals who are generally less susceptible to recurrent events and (b) heterogeneity in duration of event-free periods amongst those who experience events. The τ-IBR model is applied to a restructured version of the recurrent event data that consists of τ-restricted times-to-first event for each individual in follow-up windows initiated at regularly spaced follow-up times. Censoring of recurrent event times is handled through expectation-solution (ES) or multiple imputation (MI) approaches, with generalized estimating equation (GEE) methods then used to analyze the resulting longitudinal data structure. In simulation studies, we assess the performance of our proposed τ-IBR model with ES and MI methods in the case of no censoring and 20% censoring and compare them to the results of the model proposed by Xia, Murray and Tayob (2020) [70]. An example is given based on the Azithromycin for Prevention of COPD Exacerbations Trial.

In Chapter 4 we extend our multivariate τ-IBR modeling approach to the analysis of complex dependently censored time-to-event data. Our approach is motivated by United States lung waitlist candidate urgency estimation, which is based on estimates of τ-restricted mean survival time (τ-RMST) for τ =1 year. One statistical challenge of analyzing this data is that lung waitlist candidates with high LAS values are typically offered lung transplants. Hence their one-year-restricted waitlist survival times are dependently censored at the time of transplant. We use an inverse-transform multiple imputation (MI) approach that incorporates inverse-probability-of-censoring weights (IPCW) to multiply impute time-to-event outcomes. To make best use of our τ-IBR modeling framework, event-time data is restructured into a longitudinal dataset that includes multiple τ-length follow-up windows per individual with updated risk factors at the start of each window. This chapter makes the case that the τ-RMST estimation is improved with additional attention given to modeling the point mass of individuals who achieve a τ-RMST of τ via simulation studies and highlights advantages of our approach with an analysis of lung candidate data.

CHAPTER 2

τ-Inflated Beta Regression Model for Analysis of Restricted Mean Survival Subject to Censoring

2.1 Introduction

τ-Restricted Mean Survival Time (RMST) models [22, 77, 4, 6, 30, 71, 72, 62, 60] are popular for modeling censored time-to-event data. This modeling framework assesses the impact of predictors on a time-to-event during a τ-length follow-up period. Finite follow-up periods are ubiquitous in clinical trial and observational study settings. For example in the Azithromycin for Prevention of Chronic Obstructive Pulmonary Disease (COPD) Exacerbations Trial [2], patients were randomized to azithromycin or placebo and followed for the primary endpoint of time-to-first acute exacerbation during the subsequent year. In this setting, a 1-year-RMST model estimates acute exacerbation-free days over the follow-up year by treatment group and other patient characteristics.

More formally, for time-to-event, T, the τ-restricted event-time is min(τ, T). In this paper, and in most τ-restricted event-time literature, τ is treated as a non-random, pre-specified quantity within the range of support for the data. For the azithromycin study, the $\tau = 1$ year follow-up period will be our focus. Alternatively, Tian et al [61] frame τ as a tuning parameter to be estimated from the observed data.

For a particular covariate profile, Z, the τ-RMST, E[min$(\tau, T)|Z$], can be estimated using the model described by Andersen et al. [4] and implemented in the R pseudo package [23]; we will hereafter call this model the standard or traditional τ-RMST model. This method first converts each τ-restricted, (potentially censored) time-to-event to a pseudo-observation (PO). Linear regression can then be applied to the PO outcomes to estimate τ-RMST values.

One data feature that has not received adequate attention in τ-RMST literature is the point mass of events at τ for the τ-restricted event time, min(τ, T). Individuals who remain event-free during follow-up are certainly of interest when evaluating treatment benefit. In the azithromycin study, treatment benefit may manifest as (1) a decrease in the chance that an exacerbation occurs during the year and/or (2) a longer exacerbation-free period amongst those who have an event during that year. Different patient risk profiles may lend themselves to treatment benefit of the form (1) or (2), but standard τ-RMST models do not differentiate between these two types of treatment effects, which is a limitation of these models.

Cure rate model literature considers such point masses [8, 7, 11, 17, 74, 56, 63, 27], viewing the study cohort as a mixture of cured or uncured patients depending on whether they experience the event during follow-up or not. Typically the cure probability, $\text{Pr}(T > \tau|Z)$, is modeled via logistic regression and the censored times-to-event are simultaneously modeled via a Cox proportional hazards model or accelerated failure time (AFT) model. In the 1-year azithromycin trial, exacerbation-free patients are not truly cured. However, we may borrow mixture distribution ideas for those experiencing treatment benefit of the form (1) and/or (2) when estimating $\text{E}[\text{min}(\tau, T)|Z]$ via the relationship,

$$\text{E}[\text{min}(\tau, T)|Z] = \text{E}(T|T < \tau, Z)\text{Pr}(T < \tau|Z) + \tau\text{Pr}(T \geq \tau|Z).$$

τ-RMST estimation based on the right-hand side of this expression is hypothesized to gain precision over standard estimates due to better modeling of its individual elements. This intuition has been borne out time and again when estimating survival quantities using appropriate weighted expressions in the presence of censored data [13, 34, 35, 32]. We have not seen mixture distributions that address the point mass of events at τ applied to τ-RMST estimation in censored time-to-event literature. In addition to improving efficiency of τ-RMST estimation, we will later demonstrate how separate models for those who experience versus do not experience events during the τ-duration follow-up period enhance our understanding of treatment effect in the azithromycin trial.

In this manuscript, we develop τ-inflated beta regression (τ-IBR) models for censored times-to-event that (1) provide a more enriched understanding of the association between predictors and τ-restricted times-to-event and (2) allows for more efficient estimation of τ-RMST values. In developing these models, we assume a one-inflated beta distribution for $\tau^{-1}\text{min}(\tau, T)$ and, without loss of generality, rescale to the desired sample space of zero to one; parameter estimates maintain interpretability on either scale (zero to one or zero to τ). Although we have seen one-inflated beta regression models developed for uncensored outcomes [37, 38, 73], we have not seen these ideas adapted and rescaled for use with censored τ-restricted event times.

To fit our proposed models we develop both expectation-maximization (EM) and multiple imputation (MI) algorithms. Multiple imputation of censored survival outcomes has grown in popularity [30, 12, 57, 20, 70], since it enables a variety of complete case analysis methods to be applied to the multiply imputed datasets once the imputations are done. We develop an MI algorithm that falls within the class of inverse transform (IT) methods where, using the inverse transform theorem, imputes are sampled from Kaplan-Meier survival function estimates within an appropriate risk set of individuals comparable to the censored individual. Our risk set definition is based on the τ-IBR model, but otherwise MI procedure is entirely nonparametric, conveying some robustness over fully parametric MI methods.

The remainder of the manuscript is organized as follows. In section 2.2, we introduce notation and the τ-IBR modeling approach with corresponding τ-RMST estimates in the special case with no censored data. Section 2.3 describes EM and MI model fitting procedures for censored data along with details for τ-RMST inference based on the τ-IBR model. Finite sample properties of our methods are given via simulation in Section 2.4. We highlight attractive properties of our modeling approach compared to traditional τ-RMST models through an analysis of the azithromycin study in Section 2.5, followed by a discussion in Section 2.6.

2.2 Notation and Model Specification

For patient i, let T_i and C_i be the latent time-to-event and censoring time, respectively, where C_i is independent of T_i, $i = 1, 2, \ldots, n$. The observable data becomes $\{X_i, \Delta_i, Z_i\}$, where $X_i = \min(T_i, C_i)$ with censoring indicator $\Delta_i = I(T_i \leq C_i)$, and Z_i is a vector of covariates, $i = 1, \ldots, n$. For the remainder of this section, we assume the special case with no censoring, so that patient i's τ-restricted event time, $\min(\tau, T_i)$, is uncensored, $i = 1, 2, \ldots, n$. We will return to the censored setting in section 2.3.

Define $B_i = I(T_i \geq \tau)$, $Y_i = \tau^{-1} T_i \,|\, T_i < \tau$, $\pi_i = \mathrm{E}(B_i|Z_i)$ and $\mu_i = \mathrm{E}(Y_i|Z_i)$, where $0 < \pi_i < 1, 0 < \mu_i < 1$, $i = 1, \ldots, n$. Since $\min(\tau, T_i) = T_i(1 - B_i) + \tau B_i$, $\mathrm{E}\left[\min\left(\tau, T_i\right)|Z_i\right]$, becomes

$$\mathrm{E}\left(T_i \,|\, T_i < \tau, Z_i\right) \Pr(T_i < \tau|Z_i) + \tau \Pr(T_i \geq \tau|Z_i) = \tau\left[\mu_i(1 - \pi_i) + \pi_i\right]. \tag{2.1}$$

Hence, we may estimate τ-RMST values for individuals with different covariate profiles via joint models of π_i and μ_i. One advantage of this approach over the standard τ-RMST model is the flexibility to model different relationships between covariates and μ_i as opposed to π_i. Let $Z_{\pi i}$ and $Z_{\mu i}$ be potentially reduced subsets of Z_i such that $\pi_i = \mathrm{E}(B_i|Z_i) = \mathrm{E}(B_i|Z_{\pi i})$ and $\mu_i = \mathrm{E}\left(Y_i \,|\, Z_i\right) = \mathrm{E}\left(Y_i \,|\, Z_{\mu i}\right)$.

We assume a logistic regression model for π_i that is applied to outcomes, $B_i, i = 1, \ldots, n$,

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1^T Z_{\pi i}, \tag{2.2}$$

where β_1 is a vector of parameters associated with $Z_{\pi i}$. Hence, for a one unit increase in the k^{th} element of $Z_{\pi i}$, $Z_{\pi ik}$, the corresponding odds ratio of being event-free through time τ is $\exp(\beta_{1k})$ when adjusted for other factors in the model, where β_{1k} is the k^{th} element of β_1. Later it will be convenient to express π_i in terms of $\beta = (\beta_0, \beta_1^T)^T$, that is, $\pi_i = 1/\big(1 + e^{-\beta_0 - \beta_1^T Z_{\pi i}}\big)$. When emphasizing the dependence of π_i on β, we will use the notation $\pi_i(\beta)$.

To model μ_i, we assume that $Y_i \mid Z_{\mu i}$, follows a beta$[\mu_i\nu, (1-\mu_i)\nu]$ distribution with probability density function, $f_{Y_i}(y_i; \mu_i, \nu) = \frac{\Gamma(\nu)}{\Gamma(\mu_i\nu)\Gamma[(1-\mu_i)\nu]} {y_i}^{\mu_i\nu - 1}(1-y_i)^{(1-\mu_i)\nu - 1}$, $i = 1, \dots, n$. We then assume a beta regression model for μ_i that is applied to outcomes, $Y_i, i = 1, \dots, n$,

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \alpha_0 + \alpha_1^T Z_{\mu i}, \tag{2.3}$$

where α_1 is a vector of the parameters associated with $Z_{\mu i}$. When emphasizing the dependence of $\mu_i = 1/\big(1+e^{-\alpha_0 - \alpha_1^T Z_{\mu i}}\big)$ on $\alpha = (\alpha_0, \alpha_1^T)^T$, we will use the notation $\mu_i(\alpha)$. To interpret this model, consider the k^{th} element of $Z_{\mu i}$, $Z_{\mu ik}$, with corresponding parameter α_{1k}. For a one unit increase of $Z_{\mu ik}$ from z to $z+1$, the fold change for μ_i becomes $e^{\alpha_{1k}}(1 + e^{\alpha_0 + z\alpha_{1k}})/(1 + e^{\alpha_0 + \alpha_{1k} + z\alpha_{1k}})$ when all other predictors in the model are zero. In subject area manuscripts, we center continuous predictors and use zero values for reference groups of categorical predictors to aid in fold-change interpretations.

Together, models (2.2) and (2.3) will hereafter be called the τ-inflated beta regression model, or the τ-IBR model. To estimate parameters in models (2.2) and (2.3) based on completely observed data $\{(Y_i, B_i) = (y_i, b_i), i = 1, \dots, n\}$, we use the log-likelihood function:

$$l(\theta) = \sum_{i=1}^{n} \big\{b_i \log[\pi_i(\beta)] + (1-b_i)\log[1-\pi_i(\beta)] + (1-b_i)\log[f_{Y_i}(y_i; \mu_i(\alpha), \nu)]\big\}. \tag{2.4}$$

where $\theta = (\alpha^T, \beta^T, \nu)^T$ denotes the entire set of parameters. Differentiating (2.4) with respect to each component of θ yields the score vector:

$$U_\theta = (U_\beta^T, U_\alpha^T, U_\nu)^T,$$

where

$$U_\beta = \frac{\partial l(\theta)}{\partial \beta} = \sum_{i=1}^{n} [b_i - \pi_i(\beta)] Z_i^\pi,$$

$$U_\alpha = \frac{\partial l(\theta)}{\partial \alpha} = \sum_{i=1}^{n} (1-b_i)\nu(y_i^* - \psi_i^*)\mu_i(\alpha)[1-\mu_i(\alpha)] Z_i^\mu,$$

$$U_\nu = \frac{\partial l(\theta)}{\partial \nu} = \sum_{i=1}^{n}(1-b_i)\big[\mu_i(\alpha)(y_i^* - \psi_i^*) + \phi_i^* + \log\big(1-y_i\big)\big].$$

Here, $Z_i^\pi = (1, Z_{\pi i}{}^T)^T$, $Z_i^\mu = (1, Z_{\mu i}{}^T)^T$, $y_i^* = \log[y_i/(1-y_i)]$, $\psi_i^* = \psi[\mu_i(\alpha)\nu] - \psi\{[1-\mu_i(\alpha)]\nu\}$ and $\phi_i^* = \psi(\nu) - \psi\{[1-\mu_i(\alpha)]\nu\}$, where $\psi(x) = d\log\Gamma(x)/dx$. Later it will also be convenient to define $\psi'(x) = [1/\Gamma(x)][d\Gamma(x)/dx]$.

The observed Fisher information matrix becomes:

$$J_\theta = \begin{pmatrix} J_\beta & 0 & 0 \\ 0 & J_\alpha & J_{\alpha\nu} \\ 0 & J_{\nu\alpha} & J_\nu \end{pmatrix} = \begin{pmatrix} -\frac{\partial^2 l(\theta)}{\partial\beta\partial\beta^T} & 0 & 0 \\ 0 & -\frac{\partial^2 l(\theta)}{\partial\alpha\partial\alpha^T} & -\frac{\partial^2 l(\theta)}{\partial\alpha\partial\nu} \\ 0 & -\frac{\partial^2 l(\theta)}{\partial\alpha\partial\nu} & -\frac{\partial^2 l(\theta)}{\partial\nu\partial\nu^T} \end{pmatrix}$$

where

$$\mathcal{J}_\beta = \sum_{i=1}^{n} Z_i^\pi Z_i^{\pi^T} \pi_i(\beta)[1-\pi_i(\beta)],$$

$$\mathcal{J}_\alpha = -\sum_{i=1}^{n}(1-b_i)[\nu(y_i^* - \phi_i^*)\mu(\alpha)[1-\mu(\alpha)][1-2\mu(\alpha)] - \nu^2\mu_i(\alpha)^2[1-\mu(\alpha)]^2\psi_i^\dagger]Z_i^\mu Z_i^{\mu^T},$$

$$\mathcal{J}_\nu = -\sum_{i=1}^{n}(1-b_i)\{-\mu_i^2(\alpha)\psi_i^\dagger + 2\mu_i(\alpha)\psi_i'\{[1-\mu_i(\alpha)]\nu\} + \phi^\dagger\},$$

$$\mathcal{J}_{\alpha\nu} = -\sum_{i=1}^{n}(1-b_i)\{y_i^* - \psi_i^* - \nu\mu_i(\alpha)\psi_i^\dagger + \nu\psi_i'[\nu - \mu_i(\alpha)\nu]\}\mu_i(\alpha)[1-\mu_i(\alpha)]Z_i^\mu$$

with $\psi_i^\dagger = \psi'[\mu_i(\alpha)\nu] + \psi'\{[1-\mu_i(\alpha)]\nu\}$ and $\phi_i^\dagger = \psi'(\nu) - \psi'[(1-\mu_i)\nu]$.

The maximum likelihood estimator, $\hat\theta = (\hat\alpha^T, \hat\beta^T, \hat\nu)^T$, is the solution to $U_\theta = 0$. Various software programs can solve for $\hat\theta$, which includes quantities of interest, $\hat\beta$ and $\hat\alpha$. We used the nlminb function from the stats package in RStudio version 1.0.153 to obtain the maximum likelihood estimator $\hat\theta$, which is a quasi-Newton method optimizer [14]. We then estimated $\widehat{\text{Var}}(\hat\theta)$, which includes corresponding covariance matrices, $\hat V_\beta$ and $\hat V_\alpha$, related to models (2.2) and (2.3), respectively, using the inverse of the information matrix J_θ at $\hat\theta$.

In addition to parameter estimation, we may estimate the τ-RMST for each individual $i = 1, \ldots, n$. Returning to equation (2.1), and defining $\hat\mu_i(\hat\alpha) = 1/(1 + e^{-\hat\alpha_0 - \hat\alpha_1^T Z_{\mu i}})$ and $\hat\pi_i(\hat\beta) = 1/(1 + e^{-\hat\beta_0 - \hat\beta_1^T Z_{\pi i}}), i = 1, \ldots, n$, the estimated τ-RMST for subject i becomes

$$\hat{\text{E}}[\min(\tau, T_i)|Z_i] = \tau\hat\mu_i(\hat\alpha)[1-\hat\pi_i(\hat\beta)] + \tau\hat\pi_i(\hat\beta). \tag{2.5}$$

After some algebraic manipulation relegated to Appendix A.1 ,

$$\begin{aligned}\widehat{\text{Var}}\left\{\hat{\text{E}}\left[\min\left(\tau, T_i\right)|Z_i\right]\right\} &= \tau^2\left(1-\frac{1}{1+e^{-\hat{\alpha}^T Z_i^\mu}}\right)^2 Z_i^{\pi T}\hat{V}_\beta Z_i^\pi \frac{(e^{-\hat{\beta}^T Z_i^{\pi T}})^2}{(1+e^{-\hat{\beta}^T Z_i^\pi})^4} \\ &+ \tau^2\left(1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^\pi}}\right)^2 Z_i^{\mu T}\hat{V}_\alpha Z_i^\mu \frac{\left(e^{-\hat{\alpha}^T Z_i^\mu}\right)^2}{\left(1+e^{-\hat{\alpha}^T Z_i^\mu}\right)^4},\end{aligned} \tag{2.6}$$

where $\hat{V}_\beta$ and $\hat{V}_\alpha$ are estimated coefficient covariance matrices from models (2.2) and (2.3).

2.3 Extension to Censored Time-to-event Outcomes

In this section, we extend methods from section 2.2 to the setting with censored time-to-event data. In sections 2.3.1 and 2.3.2, we describe EM and MI algorithms, respectively, for fitting the τ-IBR model and performing inference.

2.3.1 EM Algorithm

We first describe contributions to the complete data log-likelihood from individuals $i = 1, \ldots, n$, where, without loss of generality, we assume that individuals $i = 1, \ldots, n_1$ have observed $B_i = b_i$ as well as observed $Y_i = y_i$ for those with $B_i = 0$. For the remaining $i = n_1 + 1, \ldots, n$ individuals with censored τ-restricted times-to-event, we observe $Y_i \geq y_i$, but do not observe B_i. Contributions to the complete data log-likelihood from individuals $i = 1, \ldots, n_1$, follow equation (2.4), that is,

$$l_1(\theta) = \sum_{i=1}^{n_1}\left\{b_i \text{log}[\pi_i(\beta)] + (1-b_i)\text{log}[1-\pi_i(\beta)] + (1-b_i)\text{log}[f_{Y_i}(y_i; \mu_i(\alpha), \nu)]\right\}. \tag{2.7}$$

For individuals $i = n_1 + 1, \ldots, n$, the contribution to the complete data log-likelihood depends on the unobserved B_i values, that is,

$$l_2(\theta) = \sum_{i=n_1+1}^{n}\left\{B_i \text{log}[\pi_i(\beta)] + (1-B_i)\text{log}[1-\pi_i(\beta)] + (1-B_i)\text{log}[1-F_{Y_i}(y_i; \mu_i(\alpha), \nu)]\right\} \tag{2.8}$$

where $F_{Y_i}(y_i; \mu_i(\alpha), \nu)$ is the cumulative distribution for Y_i evaluated at y_i. Hence, the complete data log-likelihood function across individuals, $i = 1, \ldots, n$, is $l_c(\theta) = l_1(\theta) + l_2(\theta)$.

The EM algorithm is an iterative procedure with an expectation step (E-step) and a maximation step (M-step) that are repeated until convergence of model parameters according to predefined criteria. Let $\hat{\theta}^{(r)} = \{\hat{\alpha}^{(r)}, \hat{\beta}^{(r)}, \hat{\nu}^{(r)}\}, r = 1, \ldots$ be the vector of τ-IBR model parameter estimates obtained at the r^{th} iteration of the M-step of the EM algorithm. Initial parameter estimates, $\hat{\theta}^{(0)}$, are based on fitting the τ-IBR model given in section 2.2 to a dataset completed using the completely nonparametric IT imputation method for censored values given by Hsu, Taylor and Murray [20].

The r^{th} iteration of the E-step in the EM algorithm is to compute the expectation of the log-likelihood function with respect to B_i, conditional on the observed data and current parameter estimates, $\hat{\theta}^{(r-1)}$. In equation (2.7), $B_i = b_i$ is already known, so we only need to calculate the conditional expectation for equation (2.8), which becomes

$$\mathrm{E}\big[l_2(\theta)|\hat{\theta}^{(r-1)}, Y_i \geq y_i, i = n_1+1,\ldots,n\big] = \sum_{i=n_1+1}^{n} \big\{w_i^{(r-1)}\log[\pi_i(\beta)] + \big[1-w_i^{(r-1)}\big] \log[1-\pi_i(\beta)] + \big[1-w_i^{(r-1)}\big][1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]\big\}, \tag{2.9}$$

$$\begin{aligned}\text{where } w_i^{(r-1)} &= \mathrm{E}(B_i|Y_i \geq y_i, i = n_1+1,\ldots,n, \hat{\theta}^{(r-1)})\\ &= \frac{\pi_i(\hat{\beta}^{(r-1)})}{\pi_i(\hat{\beta}^{(r-1)}) + [1-\pi_i(\hat{\beta}^{(r-1)})][1-F_{Y_i}(y_i;\mu_i(\hat{\alpha}^{(r-1)}),\hat{\nu}^{(r-1)})]}.\end{aligned}$$

The M-step in the EM algorithm maximizes $\mathcal{Q}(\theta;\hat{\theta}^{(r-1)}) = \mathrm{E}\big[l_c(\theta)|\hat{\theta}^{(r-1)}, Y_i \geq y_i, i = n_1+1,\ldots,n\big] = l_1(\theta) + \mathrm{E}\big[l_2(\theta)|\hat{\theta}^{(r-1)}, Y_i \geq y_i, i = n_1+1,\ldots,n\big]$ with respect to the τ-IBR model parameters, resulting in updated estimates, $\hat{\theta}^{(r)}$. For implementing the M-step we use a quasi-Newton method optimizer available through the nlminb function from the R stats package [14]. The EM algorithm iterates between the E- and M-steps until $|\hat{\theta}^{(r)} - \hat{\theta}^{(r-1)}| < \epsilon$ for some $\epsilon > 0$; for results given in this manuscript, we used $\epsilon = 10^{-4}$. Hereafter, $\hat{\theta}^{EM}$ denotes τ-IBR parameter estimates obtained via this EM algorithm approach.

The estimated asymptotic variance-covariance matrix of $\hat{\theta}^{EM}$ is determined using the method of Louis [31, 55], which in this case becomes

$$\widehat{\mathrm{Var}}(\hat{\theta}^{EM}) = \left[-\frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\theta\partial\theta^T} - \mathrm{Var}\left(\frac{\partial l_c(\theta)}{\partial\theta}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}^{EM}},$$

where

$$\frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\beta\partial\beta^T} & 0 & 0 \\ 0 & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\alpha^T} & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\nu} \\ 0 & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\nu} & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\nu^2} \end{bmatrix}$$

and

$$\text{Var}\left(\frac{\partial l_c(\theta)}{\partial \theta}\right) = \begin{bmatrix} \text{Var}(U_\beta) & 0 & 0 \\ 0 & \text{Var}(U_\alpha) & \text{Cov}(U_\alpha, U_\nu) \\ 0 & \text{Cov}(U_\alpha, U_\nu) & \text{Var}(U_\nu) \end{bmatrix}.$$

For details on the determination of $\widehat{\text{Var}}(\hat{\theta}^{EM})$ via the Louis approach, see Appendix A.2. EM-based τ-RMST and corresponding variance estimates are constructed using equations (2.5) and (2.6), substituting elements of $\hat{\theta}^{EM}$ and $\widehat{\text{Var}}(\hat{\theta}^{EM})$ for $\hat{\theta}$ and $\widehat{\text{Var}}(\hat{\theta})$, as appropriate.

2.3.2 MI Algorithm

The overall goal of MI methods is to produce M uncensored datasets, each of which can be analyzed using appropriate uncensored data analysis methods. Quantities may be combined across the M analyses using an approach laid out by Rubin [45, 46]; $M = 10$ MI datasets are often recommended and were used in this manuscript. Our proposed MI approach is a variation of inverse transform (IT) imputation. For each censored individual, IT imputation methods form a risk set of individuals similar to the censored individual (parametric component) and then use the inverse-transform theorem applied to the estimated Kaplan-Meier curve in this risk set to obtain imputes (non-parametric component). Details of our MI algorithm for estimating $\theta = (\alpha^T, \beta^T, \nu)^T$ and Var(θ) are given below, with MI estimates denoted $\hat{\theta}^{MI}$ and $\widehat{\text{Var}}(\hat{\theta}^{MI})$, respectively. When event times are subject to censoring, $\hat{\theta}^{MI}$ and $\widehat{\text{Var}}(\hat{\theta}^{MI})$ estimates obtained using the algorithm below replace corresponding estimates given in section 2.2 for the uncensored case.

In our setting a censored patient, j, with $C_j < \min(\tau, T_j)$ has incomplete data for the model (2.2) outcome, B_j, as well as the model (2.3) outcome, Y_j, which is relevant if $B_j = 0$. Completed outcomes are obtained by IT imputation of $\min(\tau, T_j)$ for $j = n_1+1, ..., n$. Step 1, below, describes risk set construction based on parameter estimates from models (2.2) and (2.3), fit using the EM algorithm given in section 2.3.1. Step 2 describes the imputation procedure based on these risk sets. Step 3 gives the final M imputed data sets and formulas for combining the M analyses.

Step 1: (Risk set definition step) For individual j with $C_j < \min(\tau, T_j)$, we define a risk set, $\mathcal{R}_j$, of similar individuals. The $l = 1, 2, ..., N_j$ individuals included in $\mathcal{R}_j$ must satisfy: (a) $\max(|\hat{\mu}_l - \hat{\mu}_j|, |\hat{\pi}_l - \hat{\pi}_j|) < \epsilon$, where $\hat{\mu} \in (0, 1)$ and $\hat{\pi} \in (0, 1)$ are taken from $\hat{\theta}^{EM}$, and (b) $X_l > X_j$, . Condition (b) requires individuals in $\mathcal{R}_j$ to be at risk at individual j's censoring time. Condition (a) ensures that individuals in $\mathcal{R}_j$ have similar predicted outcomes to individual j. We typically set ϵ in condition (a) to 0.01.

For any particular dataset, we have found it useful to be flexible in defining variations of condition (a). For instance, one may stipulate that if N_j is small, ϵ is increased by 0.001 until N_j reaches a predetermined size. In our simulations and example, we incrementally increased ϵ by 0.001 until either $N_j \geq 15$ or $\epsilon > 0.5$. If the largest value in the risk set is a censored value $< \tau$, we incrementally increased ϵ by 0.001 until the largest value in the risk set was uncensored. Condition (a) may also stipulate that individuals in $\mathcal{R}_j$ have an exact match to censored individual j on a particularly important predictor, such as treatment group.

$Step$ 2: (IT imputation step) For each censored individual, $j = n_1 + 1, ..., n$, we impute a value of T_j from which imputed values for B_j and Y_j, as appropriate, can be calculated. Define $\hat{S}_{T_j}(v|\mathcal{R}_j), 0 \leq v \leq \tau$, as the nonparametric Kaplan-Meier survival estimate for individuals in $\mathcal{R}_j$ based on the data $\{(X_l, \Delta_l), l = 1, \ldots, N_j\}$. For each of the $j = n_1+1, ..., n$ censored individuals, the IT imputation algorithm first generates a uniform(0,1) random variable, u_j, and then finds the smallest observed event time v^* where $\hat{S}_{T_j}(v^*|\mathcal{R}_j) \leq u_j$. If $v^* \geq \tau$, then impute 1 for B_j and no further imputation for Y_j is needed, otherwise we impute 0 for B_j and v^* for Y_j. Completing this step results in a fully imputed dataset.

$Step$ 3: (Multiple imputation step) Repeat step 2 M times. For each imputed data set $m = 1, ..., M$, fit models (2.2) and (2.3) and obtain parameter estimates $\hat{\theta}_m^{MI}$ with corresponding estimated covariance matrix $\widehat{\text{Var}}(\hat{\theta}_m^{MI})$. The final MI estimate of θ becomes $\hat{\theta}^{MI} = \sum_{m=1}^{M} \hat{\theta}_m^{MI}$ with corresponding estimated covariance matrix:

$$\widehat{\text{Var}}(\hat{\theta}^{MI}) = \frac{1}{M}\sum_{m=1}^{M} \widehat{\text{Var}}(\hat{\theta}_m^{MI}) + (1 + M^{-1})\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m^{MI} - \hat{\theta}^{MI})(\hat{\theta}_m^{MI} - \hat{\theta}^{MI})^T.$$

The terms, $\hat{V}_\alpha^{MI}$ and $\hat{V}_\beta^{MI}$, can be extracted from $\widehat{\text{Var}}(\hat{\theta}^{MI})$, as appropriate. MI-based τ-RMST and corresponding variance estimates are constructed using equations (2.5) and (2.6), substituting elements of $\hat{\theta}^{MI}$ and $\widehat{\text{Var}}(\hat{\theta}^{MI})$ for $\hat{\theta}$ and $\widehat{\text{Var}}(\hat{\theta})$, as appropriate.

The final M imputed data sets are not restricted for use with the τ-IBR model. In our example section we produce a heat map based on multiply imputed $\tau-$restricted event times, an application that we have not previously seen with censored survival data. Multiple imputation of τ-restricted outcomes makes such graphics simple to produce.

2.4 Simulation Study

In this section we show the finite sample behavior of our proposed τ-IBR model, summarizing the quality of (i) EM and MI parameter estimates for models (2.2) and (2.3) and (ii) EM and MI τ-RMST estimates for individuals, $i = 1, \ldots, n$. Section 2.4.1 summarizes data generation details for each of the 1000 iterations of the simulation studies and section 2.4.2 gives results.

Metrics used to assess the quality of parameter estimates across simulations based on 500 subjects include the average bias (Bias), the empirical standard deviation (ESD), the average model-based standard error (ASE), and the average coverage probability of model-based 95% confidence intervals (CP). The quality of τ-RMST estimates across simulations based on n=100, 500, 1000 and 1500 subjects was evaluated by comparing the bias (Bias), the empirical mean squared error (EMSE), the ASE, and the 95% CP to those obtained from the standard τ-RMST Model. Since true τ-RMST values are specific to each of the n individuals in each of the 1000 simulations, summary statistics for the performance of τ-RMST estimates are defined as Bias $= \sum_{j=1}^{1000}\sum_{i=1}^{n}\frac{\hat{\mathrm{E}}[\min(T_{ij},\tau)] - \mathrm{E}[\min(T_{ij},\tau)]}{n \times 1000}$, EMSE $= \sum_{j=1}^{1000}\sum_{i=1}^{n}\frac{\{\hat{\mathrm{E}}[\min(T_{ij},\tau)] - \mathrm{E}[\min(T_{ij},\tau)]\}^2}{n \times 1000}$, ASE $= \sum_{j=1}^{1000}\sum_{i=1}^{n}\frac{\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[\min(\tau, T_{ij})]\}}{n \times 1000}$ and CP $= \sum_{j=1}^{1000}\sum_{i=1}^{n}\frac{I(\mathrm{Lower}_{ij} < \hat{\mathrm{E}}[\min(\tau, T_{ij})] < \mathrm{Upper}_{ij})}{n \times 1000}$, where $\mathrm{Lower}_{ij} = \hat{\mathrm{E}}[\min(\tau, T_{ij})] - 1.96 \times \sqrt{\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[\min(\tau, T_{ij})]\}}$ and $\mathrm{Upper}_{ij} = \hat{\mathrm{E}}[\min(\tau, T_{ij})] + 1.96 \times \sqrt{\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[\min(\tau, T_{ij})]\}}$.

2.4.1 Data Generation

Two independent uniform(0,1) covariates, Z_{1i}, Z_{3i} and one independent Bernoulli(0.7) covariate, Z_{2i}, were generated for each individual, $i = 1, \ldots, n$. Models (2.2) and (2.3) were taken to be $\log[\pi_i/(1-\pi_i)] = -1.0 + 1.0Z_{1i} + 2.0Z_{2i} - 1.5Z_{3i}$, and $\log[\mu_i/(1-\mu_i)] = -2.0 + 1.2Z_{1i} + 2.0Z_{2i}$, respectively. From these models, each individual's calculated π_i and μ_i was used to independently generate a Bernoulli(π_i) random variable, B_i, and a Beta$[\mu_i\nu, (1-\mu_i)\nu]$ random variable, Y_i. The corresponding τ-restricted survival time for the i^{th} individual was $\min(\tau, T_i) = \tau B_i + \tau Y_i(1 - B_i), i = 1, \ldots, n$, where $\tau = 30$.

We considered three different censoring mechanisms: (1) no censoring, (2) independent censoring and (3) dependent censoring, where in these latter two cases, approximately 30% of censored events occurred prior to τ. For the independent censoring case, we independently generated a Bernoulli(0.56) random variable, B_i^*, and a $\mathrm{Uniform}(0, \tau)$ random variable, U_i and defined the censoring random variable to be $C_i = U_i(1 - B_i^*) + \tau B^*, i = 1, \ldots, n$. For the dependent censoring case, subjects with $Z_{2i} = 1$ had a of 36% chance of being censored according to a Uniform(0,τ) distribution; otherwise outcomes were uncensored.

2.4.2 Simulation Results

Table 2.1 displays finite sample properties of parameter estimates from models (2.2) and (2.3) assuming no censoring, 30% independent censoring and 30% dependent censoring. In each of these settings, τ-IBR parameter estimates are approximately unbiased with estimated 95% confidence

Table 2.1: Finite sample properties of model (2.2) and (2.3) parameter estimates based on 1000 iterations with n=500 subjects assuming no censoring, independent censoring* and dependent censoring† mechanisms.

Coef.	**%Cens.**	**Method**	**Bias**	**ASE**	**ESD**	**CP**
α_0 = -2.0	0	τ-IBR	-0.008	0.160	0.162	0.949
	30*	EM	-0.011	0.173	0.178	0.949
	30*	MI	-0.006	0.17	0.18	0.935
	30†	EM	-0.009	0.170	0.172	0.948
	30†	MI	-0.005	0.167	0.176	0.941
α_1 = 1.2	0	τ-IBR	0.006	0.235	0.236	0.944
	30*	EM	0.006	0.258	0.261	0.939
	30*	MI	-0.010	0.253	0.265	0.941
	30†	EM	0.004	0.257	0.260	0.952
	30†	MI	-0.007	0.252	0.271	0.930
α_2 = 2.0	0	τ-IBR	0.005	0.146	0.147	0.947
	30*	EM	0.005	0.158	0.163	0.942
	30*	MI	0.007	0.156	0.165	0.943
	30†	EM	0.007	0.155	0.159	0.942
	30†	MI	0.008	0.152	0.161	0.935
β_0 = -1.0	0	τ-IBR	-0.014	0.314	0.306	0.955
	30*	EM	-0.014	0.357	0.350	0.954
	30*	MI	-0.038	0.348	0.36	0.936
	30†	EM	-0.019	0.347	0.335	0.968
	30†	MI	-0.034	0.338	0.353	0.949
β_1 = 1.0	0	τ-IBR	0.012	0.356	0.355	0.956
	30*	EM	0.015	0.418	0.419	0.952
	30*	MI	0.030	0.401	0.425	0.936
	30†	EM	0.020	0.417	0.409	0.960
	30†	MI	0.026	0.399	0.432	0.932
β_2 = 2.0	0	τ-IBR	0.026	0.237	0.233	0.956
	30*	EM	0.036	0.268	0.266	0.956
	30*	MI	0.039	0.261	0.274	0.948
	30†	EM	0.030	0.255	0.252	0.949
	30†	MI	0.032	0.249	0.253	0.943
β_3 = -1.5	0	τ-IBR	-0.015	0.361	0.356	0.956
	30*	EM	-0.035	0.425	0.425	0.954
	30*	MI	-0.005	0.408	0.435	0.935
	30†	EM	-0.016	0.422	0.416	0.960
	30†	MI	0.006	0.406	0.443	0.931

*Independent censoring, † Dependent censoring; ASE: Average Standard Error Estimates; ESD: Empirical Standard Deviation; CP: Coverage of 95% Confidence Interval.

Table 2.2: Comparison of estimated τ-RMST using (1) τ-IBR model and (2) τ-RMST model based on 1000 iterates.

	No censoring		**30% Independent censoring**			**30% Dependent censoring**		
	τ-IBR	**τ-RMST**	**τ-IBR (MI)**	**τ-IBR (EM)**	**τ-RMST (PO)**	**τ-IBR (MI)**	**τ-IBR (EM)**	**τ-RMST (PO)**
1500 Subjects								
Bias	0.000	-0.001	-0.003	-0.002	-0.001	-0.002	-0.002	-0.248
EMSE	0.189	0.641	0.226	0.220	0.671	0.227	0.215	0.792
ASE	0.384	0.417	0.425	0.439	0.453	0.416	0.431	0.447
CP	0.945	0.746	0.943	0.957	0.770	0.938	0.954	0.754
1000 Subjects								
Bias	-0.008	-0.01	-0.004	-0.006	-0.008	-0.008	-0.008	-0.256
EMSE	0.277	0.737	0.343	0.330	0.787	0.342	0.325	0.920
ASE	0.470	0.510	0.520	0.536	0.555	0.510	0.527	0.547
CP	0.946	0.804	0.941	0.955	0.823	0.933	0.951	0.803
500 Subjects								
Bias	0.004	-0.001	-0.012	-0.007	-0.01	0.002	0.004	-0.244
EMSE	0.540	1.020	0.680	0.659	1.128	0.653	0.619	1.233
ASE	0.663	0.722	0.733	0.757	0.784	0.718	0.743	0.774
CP	0.943	0.873	0.935	0.947	0.883	0.934	0.950	0.871
100 Subjects								
Bias	0.036	0.046	0.013	0.033	0.05	0.017	0.033	-0.196
EMSE	2.860	3.454	3.596	3.531	3.980	3.374	3.306	3.954
ASE	1.451	1.601	1.604	1.655	1.744	1.578	1.627	1.719
CP	0.908	0.928	0.898	0.910	0.932	0.897	0.912	0.929

Bias: Average difference between the true and predicted τ-RMST values across all subjects and simulations; EMSE: Empirical mean squared error of τ-RMST values across all subjects and simulations; ASE: Average of the model-based standard error estimates corresponding to the τ-RMST estimates across all subjects and simulations; CP: Empirical coverage probability of the true τ-RMST value by the model-based 95% confidence interval across all subjects and simulations.

intervals showing appropriate coverage of the true parameter values. As one might expect, the variability of parameter estimates from the model with dichotomous outcomes, i.e., model (2.2), is somewhat larger than the variability of parameter estimates from the model with continuous outcomes, i.e., model (2.3).

Table 2.2 shows finite sample properties of τ-RMST estimates using both our proposed τ-IBR methodology and the standard τ-RMST model. We initially generated results for the case with n=500 subjects as in Table 2.1, noting that τ-IBR-based τ-RMST estimates were approximately unbiased regardless of estimation approach (EM, MI) or censoring mechanism (none, independent, dependent) and 95% confidence intervals showed appropriate coverage. However, when viewing results from the standard τ-RMST model, two interesting findings emerged. First, as has been reported by others [30, 71, 72, 60], we noted biased τ-RMST estimates when using the standard τ-RMST model in the dependent censoring case. Even more interesting was that in all scenarios, τ-RMST estimates based on the standard τ-RMST model had both larger ASE values and poorer coverage probabilities than estimates based on the τ-IBR model.

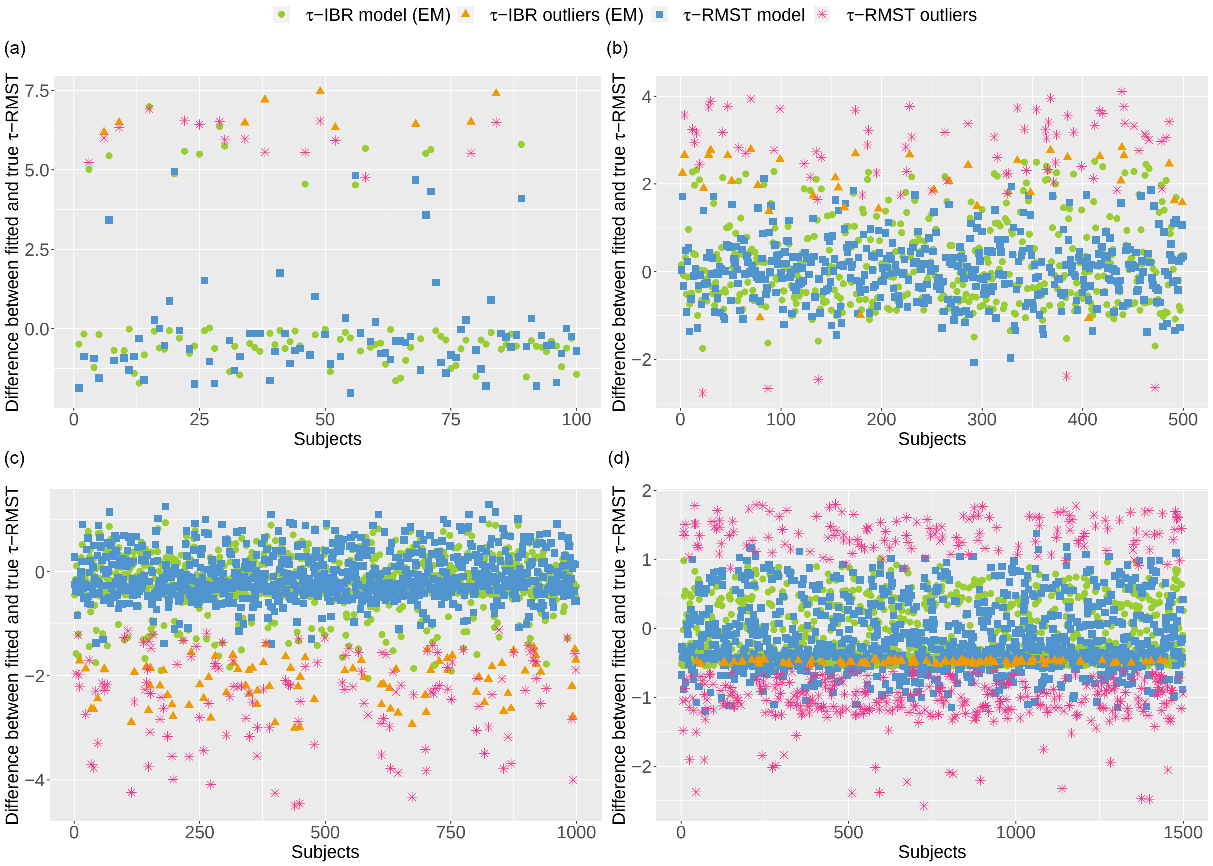

Figure 2.1: Difference between EM-fitted τ-RMST and actual τ-RMST based on (a) 100 subjects (b) 500 subjects (c) 1000 subjects and (d) 1500 subjects.

To further investigate the poor coverage probabilities of the standard τ-RMST model seen in our simulations, we first looked at a range of sample sizes to see if coverage using the standard method improved with increased sample size. Rather than correcting the problem, increased sample sizes resulted in increasingly worse coverage probabilities using the standard τ-RMST model. Our next strategy to gain intuition about the poor coverage probabilities using the standard τ-RMST model in our simulations was to plot differences between estimated and true τ-RMST values for representative datasets with n=100, 500, 1000 and 1500 subjects. Figure 2.1 displays these differences for the independent censoring setting, highlighting subjects whose true τ-RMST values were not covered by each method's estimated 95% confidence interval. τ-IBR estimates in Figure 2.1 are based on the EM algorithm described in section 2.3.1. Additional figures of this nature for the dependent censoring setting and using τ-IBR estimates based on the MI algorithm given in section 2.3.2 are located in Supplemental Figure A.1 in Appendix A.3.

Although the average bias is near zero for the standard τ-RMST model estimates in Figure 2.1, high absolute bias is seen in many cases due to not taking into account the behavior of the point mass of min(τ, T_i) at τ. This also is reflected in the higher EMSE results seen for the standard τ-RMST model in Table 2.2. Hence, in very plausible settings where the point mass at τ contains important statistical information about the τ-restricted mean, we see significant gains in precision using the more suitable τ-IBR modeling approach.

2.5 Azithromycin for Prevention of COPD Exacerbations Trial

In this section, we use the proposed τ-IBR model to analyze data from the Azithromycin for Prevention of COPD Exacerbations Trial. In this trial, COPD patients were randomized to daily azithromycin or placebo for 1 year in addition to their usual care. Our analysis focuses on 1112 participants who were available for multivariable modeling with adjustments for age [in decades and centered at 65 years], sex [male=1, female=0], forced expiratory volume in one second (FEV_1) [in percentage of predicted units by tens and centered at 40$\%$ of predicted], smoking status [current=1, former=0] and study site.

Multiply imputed datasets based on our proposed τ-IBR model were used to create the heatmap in Figure 2.2. Heatmap entries are individual 1-year RMST values averaged across 10 imputed datasets, with higher values in the yellow color range and lower values in the purple color range. Additional individual characteristics (treatment, age, gender, $\text{FEV}_1\%$ of predicted and smoking status) are color coded on the top of the heatmap. Descriptively, the cluster with the highest 1-year RMST values is more likely to be taking azithromycin, is predominantly male, and tends to have $\text{FEV}_1\%$ $>$50$\%$ of predicted. The cluster with the lowest 1-year RMST values is more likely to be taking placebo, is predominantly female and tends to have $\text{FEV}_1\%$ $<$30$\%$ of predicted.

Table 2.3 presents parameter estimates from the proposed multivariable 1-year-IBR models and the traditional 1-year-RMST model; parameter estimates for study site are submerged in this table. For those who experienced an exacerbation during the 1 year of follow-up, the beta regression component of the 1-year-IBR model [Model (2.3)] estimates the ratio of estimated exacerbation-free time during the year when comparing those with versus without a one unit increase in each predictor (Fold Change), assuming other predictors in the model are zero; zero values for the continuous variables, age and $\text{FEV}_1\%$ of predicted are assumed to be at their centered values of 65 and 40, respectively, when interpreting fold changes. The logistic regression component of the 1-year-IBR model [Model (2.2)] gives estimated odds ratios for remaining exacerbation-free at one year comparing those with versus without a one unit increase in the predictor, adjusted for other covariates in the model. The 1-year RMST model was fit with an identity link, so that each regression coefficient divided by τ, Coef/τ, is the percentage increase in 1-year-RMST for each unit increase of the predictor, adjusted for other covariates in the model. Results in Table 2.3 include corresponding 95$\%$ confidence intervals (CIs) and p-values.

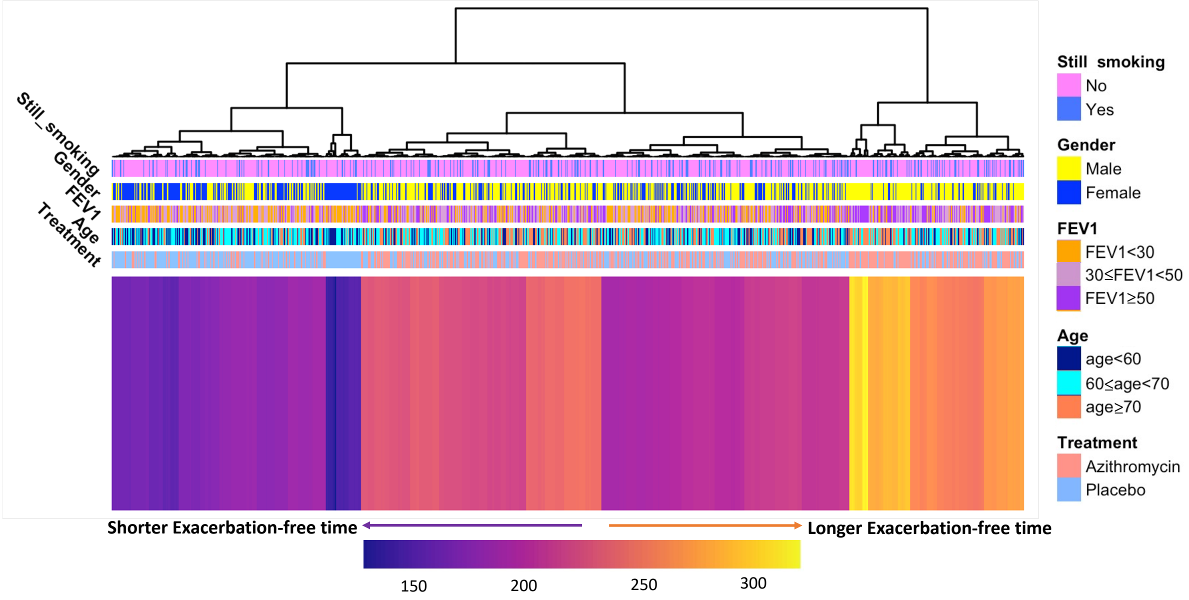

Figure 2.2: One-year RMST values for participants in the COPD Exacerbations Trial. Heatmap entries are individual 1-year RMST values averaged across 10 multiply imputed datasets using the τ-IBR method of imputation.

In evaluating the effect of azithromycin, the 1-year-IBR model estimates an odds ratio of 1.766 based on the MI method (95% CI: 1.358-2.296, p<0.001) and an odds ratio of 1.748 based on the EM method (95% CI: 1.345-2.271, p<0.001) for remaining exacerbation-free at one year when taking azithromycin versus placebo, adjusted for other factors in the model. The effect of azithromycin on exacerbation-free time during the year, amongst those who experienced an exacerbation during the year and adjusted for other factors in the model, is not statistically significant (MI: fold change of 1.025 comparing azithromycin to placebo, 95% CI: 0.926-1.123, p = 0.617; EM: fold change of 1.026 comparing azithromycin to placebo, 95% CI: 0.927-1.125, p = 0.606). Hence, taking azithromycin seems to decrease the overall odds of experiencing an exacerbation during the year, as opposed to increasing exacerbation-free time for those subject to experiencing an exacerbation during the year. The traditional 1-year-RMST model estimates a 9.1% increase in 1-year-RMST for those taking azithromycin versus placebo (95% CI: 4.7%-13.6%, p<0.001), which correctly interprets the impact of treatment on the estimated 1-year-RMST, but does not characterize impact in terms of patients being more or less susceptible to exacerbations during the year as the 1-year-IBR model does. Supplemental Table A.1 in Appendix A.4 displays the estimated effect of azithromycin for varying $\tau = 3, 6, 9$ and 12 months in τ-IBR and τ-RMST models, adjusted for same predictors. A similar pattern is seen in each case when interpreting results from the τ-IBR models, that is, that azithromycin is associated with being less susceptible to exacerbations during the various follow-up periods of interest. This pattern is discernible even for the

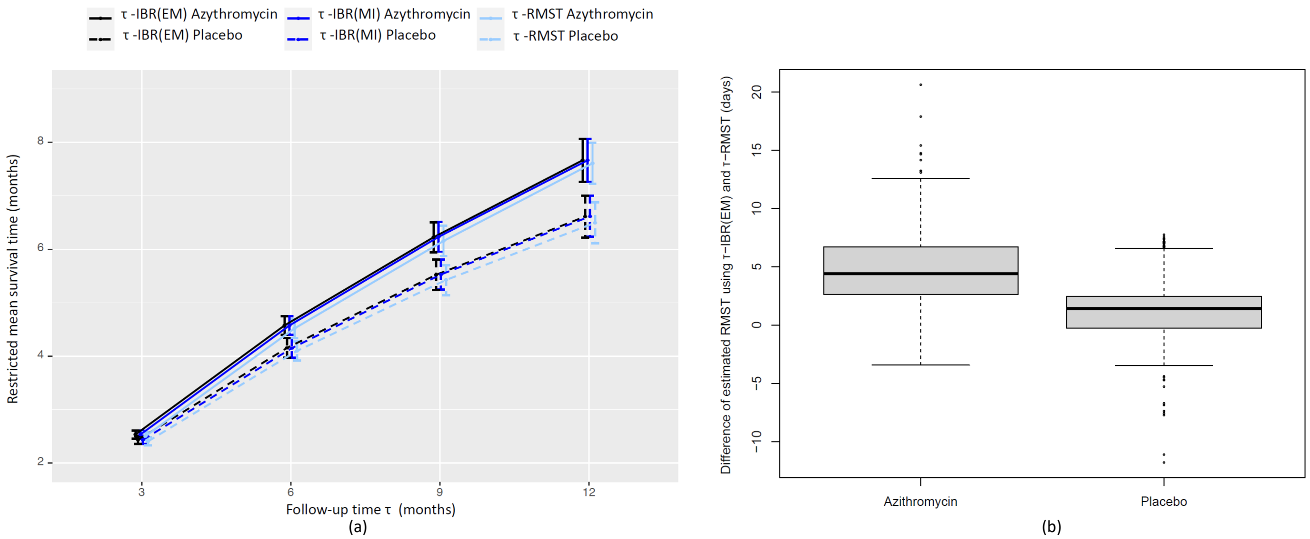

Figure 2.3: (a) Estimated RMST by treatment group for varying τ in τ-IBR and τ-RMST multivariable models.
(b) Individual level differences between 1-year-RMST estimates using the 1-year-IBR EM-fitted model versus estimates using the traditional 1-year-RMST model by treatment group.

$\tau = 3$ month follow-up window when using the 3-month-IBR models (p=0.003). The traditional 3-month-RMST model does not achieve statistical significance for the azithromycin effect in this case (p=0.084).

One-year-RMST estimates of interest are shown in Figures 2.3(a) and 2.3(b). Figure 2.3(a) shows estimated τ-RMST values and corresponding 95% CIs by treatment group for $\tau = 3$, 6, 9 and 12 month follow-up periods based on the τ-IBR and τ-RMST models given in Table 2.3, assuming average values for the overall study cohort for predictors other than treatment group. When adjusted for other predictors in the model(s), all τ-RMST estimates are close to one another, with τ-RMST estimates being slightly higher using the τ-IBR models versus the standard τ-RMST model. Figure 2.3(b) shows individual level differences between 1-year-RMST estimates using the 1-year-IBR EM-fitted model versus estimates using the traditional 1-year-RMST model by treatment group. At the individual level, RMST estimates are typically close regardless of estimation method used. Differences ranged from 20 days (5%) higher to 11 days (3%) lower when using the 1-year-IBR versus the standard model. Recall that the standard τ-RMST model was seen to have much higher MSE when compared to the τ-IBR method in simulation (see Figure 1) due to higher bias on the absolute value scale. In this context, the τ-IBR model RMST estimates seen in Figure 2.3 are preferred.

Interestingly, the only predictor that showed statistical significance in the beta component of the 1-year-IBR model given in Table 2.3 was age. That is, among those experiencing an exacerbation during the 1 year of follow-up, the estimated exacerbation free time increased by 7% for every

Table 2.3: Azithromycin for Prevention of COPD Exacerbations Trial: Estimated 1-year-IBR and 1-year-RMST multivariable model parameters with 95% confidence intervals and p-values. All models are additionally adjusted for study site (data not shown).

	Azithromycin	**Age**	**Male**	**FEV$_1$**	**Current Smoker**
	(vs. Placebo)	**(per 10 Years)**	**(vs. Female)**	**(per 10% Predicted)**	**(vs. Ex)**
τ-IBR model (EM) (Beta Regression)					
Fold Change*	1.026	1.070	1.018	1.021	1.004
95% CI	(0.927, 1.125)	(1.007, 1.132)	(0.919, 1.118)	(0.987, 1.054)	(0.876, 1.132)
P-value	0.606	0.027	0.714	0.224	0.952
τ-IBR model (EM) (Logistic Regression)					
Odds Ratio†	1.748	1.138	1.686	1.123	1.076
95% CI	(1.345, 2.271)	(0.965, 1.341)	(1.282, 2.218)	(1.030, 1.224)	(0.766, 1.512)
P-value	<0.001	0.124	<0.001	0.008	0.672
τ-IBR model (MI) (Beta Regression)					
Fold Change*	1.025	1.070	1.021	1.019	1.100
95% CI	(0.926, 1.123)	(1.009, 1.132)	(0.919, 1.123)	(0.986, 1.052)	(0.869, 1.131)
P-value	0.617	0.024	0.689	0.247	0.998
τ-IBR model (MI) (Logistic Regression)					
Odds Ratio†	1.766	1.134	1.670	1.124	1.108
95% CI	(1.358, 2.296)	(0.964, 1.333)	(1.267, 2.201)	(1.030, 1.227)	(0.790, 1.554)
P-value	<0.001	0.128	<0.001	0.009	0.552
τ-RMST Model					
Coef/τ⋆	0.091	0.035	0.080	0.023	0.014
95% CI	(0.047, 0.136)	(0.007, 0.063)	(0.034, 0.126)	(0.008, 0.038)	(-0.045, 0.072)
P-value	<0.001	0.014	<0.001	0.002	0.641

*Among those experiencing an exacerbation during the 1 year of follow-up, fold change is the ratio of estimated exacerbation-free time during the year when comparing those with versus without a one unit increase in the predictor, assuming all other predictors are zero. Age is centered at 65 years and percent of predicted FEV$_1$ is centered at 40% to aid in interpreting fold changes.
†Odds ratio for remaining exacerbation-free at one year comparing those with versus without a one unit increase in the predictor shown, adjusted for other covariates in the model including treatment group, age, gender, percent of predicted FEV$_1$, smoking status and study site.
⋆Percentage increase in 1-year-RMST for each unit increase of the predictor, adjusted for other covariates in the model.

additional 10-year increase in age (MI: fold change of 1.070, 95% CI 1.009-1.132, p=0.024; EM: fold change of 1.070, 95% CI 1.007-1.132, p=0.027), adjusted for other predictors in Table 2.3. Upon further investigation, a significant interaction between treatment group and age was seen in the logistic component (MI: p=0.013; EM: p=0.014), but not the beta component (MI: p=0.754; EM: p=0.644), of the τ-IBR model, indicating an increasing odds of remaining exacerbation-free for older patients taking azithromycin versus placebo during the 1 year of follow-up. This interaction was also detected by the τ-RMST model (p=0.041) and was noted by the investigators based on a Cox model.

Supplemental Table A.2 in Appendix A.4 summarizes results of interaction tests between treatment group and age, gender, percent of predicted FEV$_1$ and current smoking status. Current smokers were observed to benefit less from azithromycin that former smokers with marginal significance in the logistic component of the τ-IBR model (MI: p=0.089; EM: p=0.083) and significance in the τ-RMST model (p=0.030). The τ-RMST model also showed a marginally significant interaction between treatment and percent of predicted FEV_1 (p=0.073), suggesting a stronger benefit of azithromycin for those with higher FEV_1. When tested, the τ-IBR model did not give any indication of an interaction between treatment and FEV_1 (beta component MI: p=0.125, EM: p=0.161; logistic component MI: p=0.400, EM: p=0.379). Instead, based on results from Table 2.3, the statistical signal was relegated to the logistic regression component of the τ-IBR model, with the odds of remaining exacerbation-free during the year increasing by 12% for every 10 unit increase in percent of predicted FEV_1. Otherwise, based on the beta regression component of the τ-IBR model, no significant effect of percent of predicted FEV_1 was seen to impact exacerbation-free time for those experiencing exacerbations during the year of follow-up (MI: p=0.247, EM: p=0.224).

2.6 Discussion

To our knowledge, a τ-inflated beta regression model has never been proposed as a way to model time-to-event data, censored or otherwise. The key advantages of this method are (1) a better understanding of predictors associated with no events in the τ-restricted period of interest, as opposed to predictors associated with shorter expected event-free time amongst those who experienced the event and (2) more efficient estimation of restricted means due to properly modeling the point mass of min(τ, T) events at τ. Weighted estimation expressions, such as that given by equation (2.5), have often been seen to improve efficiency of estimation in censored data settings [13, 34, 35, 32]. In the case of the weighted Kaplan-Meier survival estimate proposed by [34], closed-form asymptotic variance calculations confirmed gains in efficiency over the traditional Kaplan-Meier estimator when survival differences in subpopulations were taken into account and recombined using appropriate weighting methods. Efficiency gains of the τ-IBR method RMST estimates over traditionally estimated RMST estimates from τ-RMST models seem to bear out this intuition as well.

We developed both an EM algorithm and a semi-parametric MI algorithm for fitting and reporting results for the τ-IBR model. In this paper, we form risk sets for the IT imputation method based on the EM algorithm fitted parameter estimates. In simulation, both EM and MI estimation methods performed well in terms of efficiency and bias in settings with both independent and dependent censoring mechanisms. One could certainly argue for use of EM algorithm approach since the MI approach requires more computational time to fit. An advantage of the MI algorithm approach over the EM algorithm approach is the ability to easily perform additional analyses using uncensored data methods based on the multiply imputed datasets, for instance, in creating data views like Figure 2.2. The only parametric element to our MI algorithm is defining a risk set of similar individuals to the censored individual being imputed; the IT method of imputing from the risk set is thereafter entirely nonparametric.

A common issue in multiple imputation algorithms is whether or not to include a bootstrap step that accounts for population variability more appropriately; a bootstrap step is often recommended for producing a 'proper' imputation algorithm in Bayesian literature parlance. We took a very pragmatic approach of looking at our simulated coverage rates without a bootstrap step first and, after some experimentation with including a bootstrap step, elected to skip this step after noting little improvement in coverage rates. Again, this is a matter of taste, with those of a more pure Bayesian mindset likely to include a bootstrap step as a matter of principle, but we found the gain in obtaining results more quickly quite satisfying and continue to recommend skipping the bootstrap step. This is, of course, an easily incorporated change for those who wish to do so. We also looked to see if bootstrapped variance estimates of EM algorithm parameters would improve coverage rates in simulation and again found that coverage rates were similar to those reported without bootstrapping.

When viewed within the context of restricted mean regression models, which was our initial starting point when developing our proposed model, some clear advantages emerge from our simulation results and from our analysis of the COPD data. Our simulation results confirm better precision of τ-RMST estimates and better corresponding confidence interval coverage rates when the point mass at τ is more appropriately modeled. As seen in the COPD example, the relative importance of risk factors in models (2.2) and (2.3) shifted between the two models, with most statistical signal appearing in model (2.2). In particular, the treatment effect manifested significantly in model (2.2) and not model (2.3). The traditional τ-RMST model identifies a significant treatment effect as well, but is not able to distinguish the nature of the treatment effect as clearly. That is, the τ-RMST model identifies those on azithromycin as having longer estimated τ-RMST values but does not capture the intuition that fewer patients are susceptible to exacerbations during the follow-up period when taking this treatment. The restricted mean modeling approach is becoming more popular as an alternative to the proportional hazards model, and our proposed method fills a gap in restricted mean model literature.

In this work, the τ value used in modeling corresponded to the follow-up period of the Azithromycin study that was approximately 1 year long. Any analysis method applied to this data is restricted to this follow-up period, whether the methodology explicitly indicates this restricted period of time or not. Note that any method can consider different follow-up times of interest, for example what event times restricted to 6-months might look like. The restricted survival time framework makes this choice more explicit than other methods that, for instance, could censor event times at 6-months to evaluate short term outcomes in this study. We view the nature of specifying the follow-up period of interest using the τ-IBR and τ-RMST methods as a strength, rather than a weakness, of using the τ notation when reporting results.

CHAPTER 3

τ-Inflated Beta Regression Model for Censored Recurrent Events

3.1 Introduction

In clinical and observational studies of chronic disease, patients often experience recurrent events that are central to how the disease manifests. For example, in Chronic Obstructive Pulmonary Disease (COPD) patients, acute respiratory exacerbations are a frequent concern and much attention has been given to how to prevent or delay them. Regression models that identify risk factors, biomarkers and effective treatments for preventing exacerbations are key to understanding how to clinically manage COPD patients.

Modern regression methods for recurrent events address possible dependence between recurrent event times in an individual to obtain valid inference. Models for (a) the number of recurrent events per unit time and models for (b) the times between recurrent events are both commonly used.

Models of type (a) include Poisson and negative binomial count models that employ a dispersion parameter to better reflect count distributions that deviate from model assumptions [25, 16, 66]. Generalized estimating equations (GEE) have also been proposed to estimate parameters and corresponding variance terms in models of dependent counts over time, which allow covariates to be updated at more frequent intervals than their predecessor models [48, 1]. Zero-inflated versions of these count models are also available [24, 49], which address subpopulations of patients who may not be subject to experiencing recurrent events at their stage of chronic disease, i.e., patients who inflate the number of zero recurrent event counts in the data.

For type (b) models that focus on recurrent event time random variables, several methods have emerged since Andersen and Gill's [3] extension of the Cox model to analyze times between events, or gap times, as independent random variables subject to censoring. This independence assumption for gap time random variables taken from the same individual was relaxed by Pepe and Cai (1993) [40], Lawless and Nadeau (1995) [26] and Lin et al (2000) [28]. Model parameters are typically related to recurrent event intensity rate ratios, although if only time-independent covariates are used in these models, an estimate for the multiplicative effect on the mean number of events per unit time is available so that model (a) type interpretations of the data can be made.

A third modeling paradigm for recurrent events has recently been introduced by Xia, Murray and Tayob (2020) [70] that incorporates ideas from restricted event time [77, 4, 30, 71] and landmark analyses [5, 64, 36, 39, 53] for single times-to-event. The recurrent events data structure is first transformed to a censored longitudinal data structure of τ-restricted times-to-first event for each individual i in follow-up windows initiated at regularly spaced follow-up times $t \in \{t_1, ..., t_b\}$. More formally, the newly formatted longitudinal outcomes for individual i are written as $T_i(t)$, which is the τ-restricted time to the first event following each pre-specified time, t. Censoring of recurrent event times is handled through pseudo observation (PO) or multiple imputation (MI) approaches, with generalized estimating equation (GEE) methods then used to analyze the resulting data structure. The overall approach naturally accommodates dependence between correlated event times measured from the same individual, while avoiding dependent censoring issues that appear, for instance, in correlated gap-time derived data structures [29, 54]. Parameter estimates from the resulting models for $\mathrm{E}[T_i(t)]$ can be used to estimate and interpret longitudinal trajectories of mean τ-restricted event times across the follow-up windows according to different patient covariate profiles.

The current manuscript fits within this third modeling paradigm that restructures recurrent event data into a censored longitudinal data structure of τ-restricted times-to-first-event, $T_i(t)$, in predetermined, regularly spaced follow-up windows starting at $t \in \{t_1, ..., t_b\}$. In particular, we are motivated by the idea that many individuals in the Azithromycin in COPD study are at an earlier disease stage that makes them less susceptible to recurrent exacerbations during follow-up, so that $T_i(t) = \tau$ is often observed. These are the same types of individuals with a tendency to have zero recurrent event counts, which inspired the use of zero-inflated count models that were mentioned earlier. As in the case with zero-inflated count models, risk profiles (covariates) associated with observing $T_i(t) = \tau$ may be quite different from risk profiles associated with $T_i(t)$ given $T_i(t) < \tau$, perhaps requiring different sets of predictors to model the data appropriately. To understand trends in longitudinal restricted event times, then, it seems sensible to take into account the mixture of individuals that achieve an active recurrent event endpoint, $T_i(t) < \tau$, during a follow-up window starting at t as opposed to individuals who are event-free during a follow-up window.

In this manuscript we develop a modeling framework for recurrent event data that more explicitly characterizes the point mass of event-free individuals with $T_i(t) = \tau$ in the follow-up window starting at t versus those with active events in this follow-up window with $T_i(t), T_i(t) < \tau$. As in Xia, Murray and Tayob (2020), our model will allow for estimation of $\mathrm{E}[T_i(t)]$. However, we approach this estimation problem through the useful decomposition:

$$\mathrm{E}[T_i(t)] = \mathrm{E}[T_i(t)|T_i(t) < \tau]\mathrm{Pr}[T_i(t) < \tau] + \tau\mathrm{Pr}[T_i(t) \geq \tau].$$

By decomposing $\mathrm{E}[T_i(t)]$ in this manner and modeling its separate components, we will also be able to evaluate patient profiles that lend themselves to (1) a lower overall chance of experiencing an exacerbation during a restricted period and (2) a longer event-free time amongst patients who experience exacerbation during a restricted period. In the context of evaluating treatment effects during a clinical trial, this modeling framework will naturally allow for more nuanced evaluation of treatment effects, not only allowing for treatment effects of the type (1) versus (2), but also supporting a description of these effects over time through modeling the imposed censored longitudinal data structure. In terms of modeling efficiency, taking into account potentially different associations between predictors and outcomes in components of the decomposition above may increase precision of statistical inferences regarding times to the next recurrent event. This intuition has been verified many times when estimating survival quantities using appropriate weighted methods in the presence of censored data [13, 34, 35, 32].

The remainder of this manuscript is organized as follows. In Section 3.2, we define notation and describe how to transform the censored recurrent event data into the more regularly spaced censored longitudinal data structure for these events. In section 3.3, we develop our model in the special case with no censoring, with censoring later addressed using expectation-solution (ES) and multiple imputation (MI) approaches given in Section 3.4. Section 3.5 describes finite sample properties of our methodology via simulation. An analysis of the Azithromycin in COPD clinial trial highlighting advantages of our approach is given in Section 3.6, followed by a discussion in Section 3.7.

3.2 Notation and Construction of Censored Longitudinal Data

In this section, we define notation, review how to convert traditional recurrent event data into a censored longitudinal data structure as described in refs [58][67][69][70] and develop additional longitudinal data notation to suit the purposes of this research.

For each patient $i, i = 1, \ldots, n$, let T_{ij} be the time from the beginning of follow-up to the j^{th} recurrent event, $j = 1, \ldots, J_i$. Let C_i be the censoring time for patient i, with C_i assumed to be independent of T_{ij} for $j = 1, \ldots, J_i$. Data from different individuals indexed by $i = 1, \ldots, n$ are assumed to be independent of one another. However, correlation between recurrent event times, $T_{i1}, \ldots, T_{iJ_i}$, contributed by the same individual i is allowed. The censored nature of the data only allows us to observe $X_{ij} = \min[T_{ij}, C_i]$ with its corresponding censoring indicator $\Delta_{ij} =$

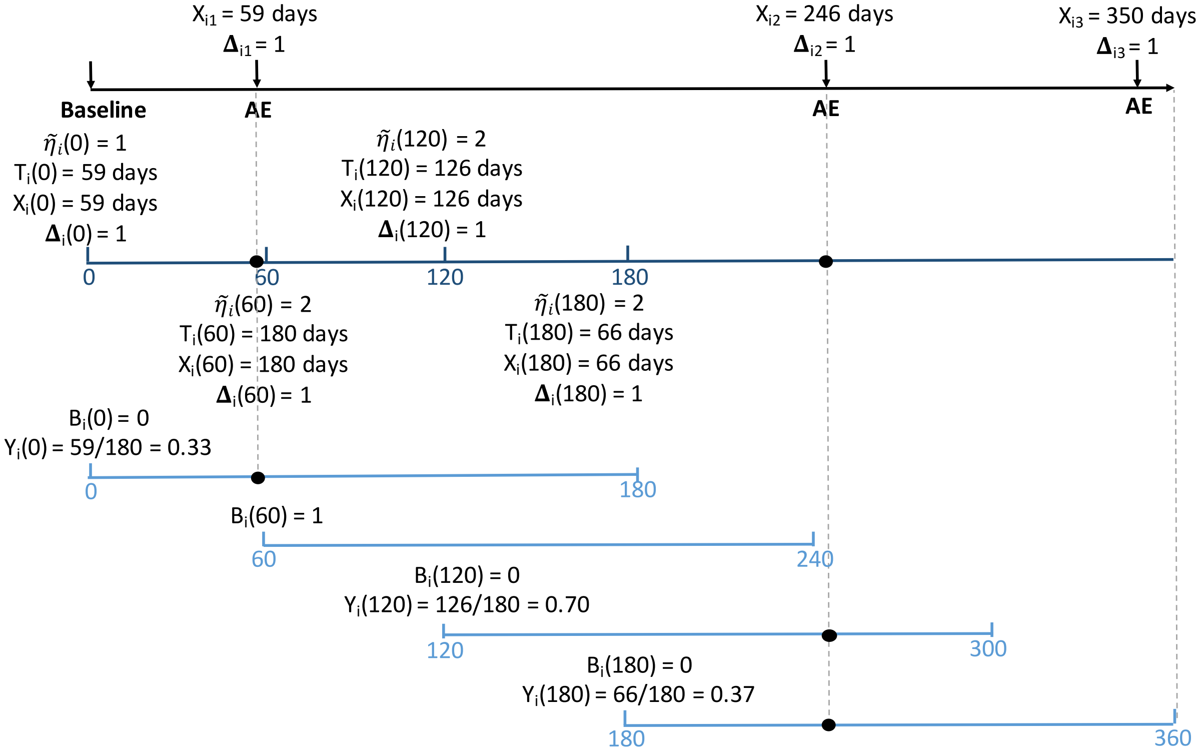

Figure 3.1: An individual from the Azithromycin in COPD Trial with 360 follow-up days shown using both the traditional and proposed longitudinal notation. (AE: Acute Exacerbation)

$I[T_{ij} < C_i]$, $i = 1, \ldots, n, j = 1, \ldots, \tilde{J}_i$ with $\tilde{J}_i \leq J$. Our definition of a recurrent event time random variable is broad, allowing for composite endpoints that might include both recurring and terminal events. In the azithromycin study, the recurrent event times were based on composite endpoints, $T_{ij} =$ min(time to j^{th} acute exacerbation (AE) for patient i, time to death for patient i), $i = 1, \ldots, n, j = 1, \ldots, J_i$.

An alternative censored longitudinal data structure for the analysis of recurrent event data was first proposed by Tayob and Murray [58]. For each individual, the recurrent event data structure is transformed to a series of regularly-spaced follow-up windows with censored times-to-first-event recorded for each window. Figure 3.1 shows data for 360 follow-up days from an example participant in the Azithromycin Study using both traditional and the proposed longitudinal notation that will be helpful as we review notation from Tayob and Murray and introduce additional longitudinal outcome notation used in this manuscript. This individual experienced three AEs at 59, 246 and 350 days before being censored at 360 days of follow-up. Hence, using traditional recurrent event notation, the data becomes $\{(X_{i1} = 59 \text{ days}, \Delta_{i1} = 1), (X_{i2} = 246 \text{ days}, \Delta_{i2} = 1), (X_{i3} = 350 \text{ days}, \Delta_{i3} = 1), (X_{i4} = 360 \text{ days}, \Delta_{i4} = 0)\}$, with $\tilde{J} = 4$. We will continue to refer back to this example patient when defining notation throughout this section.

The proposed longitudinal data structure is based on (potentially censored) times-to-first-event in follow-up windows of length τ measured from follow-up times $t \in \{t_1, ..., t_b\}$, with $t_1 = 0$ and $t_k = t_{k-1} + a, k = 2, \dots, b$. For a study with s follow-up days, t_b is at most $s - \tau$. The follow-up window length, τ, corresponds to a clinically meaningful duration in the context of the current study and patient population. For the example patient shown in Figure 3.1, s = 360 days, τ = 180 days and a, b and τ are 60 days and 4, respectively, giving $\{t_1, ..., t_4\} = \{0, 60, 120, 180\}$ days. Recommendations for the choice of a, b and τ are discussed in Tayob and Murray (2015)[58], Xia and Murray (2019)[67], and Xia, Murray and Tayob (2020)[70] in terms of computational burden, the probability of capturing each recurrent event in at least one follow-up window once the data conversion to the longitudinal structure is complete and statistical efficiency. For exponentially distributed recurrent event times, using a equal to one-third of the expected recurrent event time tends to capture 90% of the recurrent events in at least one follow-up window. It is theoretically possible to set $a = 1$ day, i.e., new follow-up windows of length τ starting every day until $t_b = s - \tau$, although this increases the computational burden of an analysis.

With the structure of the follow-up windows defined in terms of a, b and τ, we now define notation for τ-restricted times-to-first-event in each of the follow-up windows mapped from the original recurrent event time random variables. For each $t \in \{t_1, \dots, t_b\}$, we define:

$$\eta_i(t) = \text{min}\{j = 1, \dots, J_i : T_{ij} > t\} \text{ and}$$
$$T_i(t) = \text{ min}[T_{i\eta_i(t)} - t, \tau],$$

where $\eta_i(t)$ indexes the first of the original recurrent events to appear after follow-up-time t and $T_i(t)$ is the corresponding τ-restricted time-to-first-event measured from t. In the presence of censoring, individual i's observed data for the follow-up window starting at t becomes:

$$\tilde{\eta}_i(t) = \text{min}\{j = 1, \dots, \tilde{J}_i : X_{ij} > t\},$$
$$X_i(t) = \text{ min}[X_{i\tilde{\eta}_i(t)} - t, \tau] \text{ and}$$
$$\Delta_i(t) = I\left\{\text{min}[X_{i\tilde{\eta}_i(t)} - t, \tau] < C_i - t\right\},$$

where $\tilde{\eta}_i(t)$ indexes the first of the original observed recurrent event times to occur after t, with $X_i(t)$ and $\Delta_i(t)$ being the corresponding τ-restricted time-to-first-observed-event and censoring indicator, respectively, measured from the start of that follow-up window. Any individual i who is not at risk at the beginning of a follow-up window starting at t is assumed to have $X_i(t) = \Delta_i(t) = 0$.

Returning to Figure 3.1, where follow-up for patient i through day $s = 360$ is complete, the longitudinal data for patient i contributed from follow-up windows starting at $\{t_1, ..., t_4\} =$

$\{0, 60, 120, 180\}$ days becomes:

$$\{T_i(0) = X_i(0) = 59 \text{ days}, \Delta_i(0) = 1, \eta_i(0) = \tilde{\eta}_i(0) = 1\},$$
$$\{T_i(60) = X_i(60) = 180 \text{ days}, \Delta_i(60) = 1, \eta_i(60) = \tilde{\eta}_i(60) = 2\},$$
$$\{T_i(120) = X_i(120) = 126 \text{ days}, \Delta_i(120) = 1, \eta_i(120) = \tilde{\eta}_i(120) = 2\} \text{ and}$$
$$\{T_i(180) = X_i(180) = 66 \text{ days}, \Delta_i(180) = 1, \eta_i(180) = \tilde{\eta}_i(180) = 2\}.$$

Note that $T_i(60) = 180$ days reflects a $\tau = 180$ day follow-up window with no recurrent events. In the azithromycin study, approximately 48% of the 180-day length follow-up windows constructed in this way had no recurrent events, resulting in a point mass for $T_i(t)$ at 180 days. To emphasize that $T_i(t)$ is a mixture distribution with both continuous and point mass components, we may rewrite it as

$$T_i(t) = \tau B_i(t) + T_i(t)[1 - B_i(t)] = \tau\big\{B_i(t) + Y_i(t)[1 - B_i(t)]\big\}, \tag{3.1}$$

where $B_i(t) = I[T_i(t) = \tau]$ is a Bernoulli random variable with mean $\pi_i(t) = \text{Pr}[T_i(t) = \tau]$ and $Y_i(t) = \tau^{-1}T_i(t)$ defined conditionally for $T_i(t) < \tau$ (i.e., $B_i(t) = 0$), is a continuous random variable on the sample space between zero and one. Since $Y_i(t)$ is a conditional random variable that is only defined when $B_i(t) = 0$, $Y_i(t)$ is independent of $B_i(t)$ in equation (3.1). Hence, we may model these quantities separately in obtaining inferences on $T_i(t)$. Returning to Figure 3.1, individual i contributes $\{[B_i(0) = 0, Y_i(0) = 0.33], [B_i(60) = 1], [B_i(120) = 0, Y_i(120) = 0.70], [B_i(180) = 0, Y_i(180) = 0.37]\}$. Values for $Y_i(t)$ and $B_i(t)$ are only partially observed in follow-up windows starting at t, where the patient is at risk at t and censored prior to $t + \tau$. In the presence of censoring, it will be convenient to define the observed value, $\tilde{Y}_i(t) = \tau^{-1}X_i(t)$, defined conditionally for $X_i(t) < \tau$. For instance, imputation of $Y_i(t)$ for individuals with $\tilde{Y}_i(t) = \tilde{y}_i(t)$ will be discussed in section 3.4.

The following two sections of the manuscript develop methodology first in the uncensored data setting (section 3.3) and then in the censored data setting (section 3.4), where both a multiple imputation and an ES algorithm are introduced for fitting the model.

3.3 Model Specification for the Uncensored Data Setting

Using the (potentially censored) longitudinal data structure for recurrent events data described in section 3.2, we would like to understand the association between predictors and $T_i(t)$. In this section, we describe estimation and inference for our proposed model(s) in the special case with no censoring. We will address the potentially censored nature of the data in section 3.4, where a multiple imputation approach for generating uncensored datasets for analysis is developed and analyses combining inferences across these uncensored datasets are described.

To date, the only regression model applied to this proposed recurrent event data structure is the multivariate τ-Restricted Mean Survival Time (τ-RMST) model described by Xia, Murray and Tayob [70]:

$$\mathrm{E}[T_i(t)|Z_i(t)] = \tilde{\beta}^T Z_i(t), \tag{3.2}$$

$i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, where $\tilde{\beta}$ is a vector of parameters corresponding to a vector of predictors, $Z_i(t)$. In the special case with no censoring, a generalized estimating equation (GEE) approach may be used to obtain parameter estimates and corresponding variance terms. Xia, Murray and Tayob introduced the overall modeling strategy for fitting and performing inference with this model using both a multiple imputation and a pseudo-observation approach to handle the censored nature of the data and advice for structuring GEE correlation matrices that allow correlation between $T_i(t_{k_1})$ and $T_i(t_{k_2})$ for $t_{k_1} \neq t_{k_2}$. The assumed model-based variance function, $\mathrm{Var}[T_i(t)|Z_i(t)] = \sigma(t)$, is based on the Normal distribution. Although they develop their model for $T_i(t)$ on the log scale, we slightly modify their approach to maintain $T_i(t)$ on the original scale in model (3.2) to allow for more direct comparison with our method when estimating restricted mean times-to-first-event in each follow-up window. Otherwise we follow their approach when reporting results for the τ-RMST method for modeling recurrent events in this manuscript.

The key feature not addressed in model (3.2) is the point-mass of $T_i(t)$ at τ seen in the azithromycin study data. Instead of modeling $\mathrm{E}\{T_i(t)\}$ as in model (3.2), we propose modeling the mean of the expression for $T_i(t)$ seen on the right hand side of equation (3.1), namely,

$$\mathrm{E}[T_i(t)] = \mathrm{E}\big(\tau\big\{B_i(t) + Y_i(t)[1 - B_i(t)]\big\}\big) = \tau\big\{\pi_i(t) + \mu_i(t)[(1 - \pi_i(t)]\big\}, \tag{3.3}$$

$i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$ and $\mu_i(t) = \mathrm{E}[Y_i(t)]$. In the special case with no censored outcomes, both $\pi_i(t)$ and $\mu_i(t)$ in equation (3.3) can be estimated using generalized estimating equations (GEE) applied to $B_i(t), i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$ and $Y_i(t), i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, respectively. This framework allows exploration of different associations between covariates and $\pi_i(t)$ versus associations between covariates and $\mu_i(t)$. Not only does this added flexibility over model (3.2) potentially enhance our understanding of relationships between predictors and $T_i(t)$, but improved modeling of the point mass of $T_i(t)$ at τ has potential to increase efficiency when estimating $\mathrm{E}[T_i(t)]$.

Each patient $i = 1, \ldots, n$ contributes data $\{Y_i(t), B_i(t), Z_i(t)\}$ for $t \in \{t_1, ..., t_b\}$, where $Z_i(t)$ is a vector of covariates. Some of the predictors included in $Z_i(t)$ maybe more relevant in modeling the mean of $B_i(t)$ as opposed to the mean of $Y_i(t)$ and, accordingly, we define $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$

to reflect the potentially different subsets of covariates from $Z_i(t)$ relevant to modeling these two different outcomes, respectively.

Inspired by equation (3.3), the underlying assumption of the τ-IBR approach is that patient i's τ-RMST for the follow-up window starting at time t satisfies

$$\mathrm{E}[T_i(t)|Z_i(t)] = \tau\big\{\mathrm{E}[B_i(t)|Z_{\pi i}(t)] + E[Y_i(t)|Z_{\mu i}(t)]\{(1 - \mathrm{E}[B_i(t)|Z_{\pi i}(t)]\}\big\}. \tag{3.4}$$

For simplicity, we will continue to use the notation $\pi_i(t)$ for $\mathrm{E}[B_i(t)|Z_{\pi i}(t)]$ and the notation $\mu_i(t)$ for $\mathrm{E}[Y_i(t)|Z_{\mu i}(t)]$. As in standard longitudinal analysis, $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$ can include window start times t, time-dependent covariates that change at the window start times, and interactions between t and other covariates.

Models for $\pi_i(t)$ and $\mu_i(t)$ require specification of the mean and variance functions as well as the correlation structure for outcomes taken from the same individual. For the model applied to $\{B_i(t), Z_{\pi i}(t)\}$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, we specify the mean structure, $\pi_i(t)$, via

$$g[\pi_i(t)] = \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = \beta_0 + \beta_1^T Z_{\pi i}(t), \tag{3.5}$$

where β_1 is a vector of the parameters corresponding to $Z_{\pi i}(t)$. A Bernoulli variance function, $\mathrm{Var}[B_i(t)|Z_{\pi i}(t)] = \pi_i(t)[1 - \pi_i(t)]$, is assumed. Later it will be convenient to express $\pi_i(t)$ in terms of $\beta = (\beta_0, \beta_1^T)^T$ and $Z_i^\pi(t) = [1, Z_{\pi i}^T(t)]^T$, that is, $\pi_i(t) = 1/\big[1 + e^{-\beta^T Z_i^\pi(t)}\big]$.

For the model applied to $\{Y_i(t), Z_{\mu i}(t)\}$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, we specify the mean structure, $\mu_i(t)$, via

$$g[\mu_i(t)] = \log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = \alpha_0 + \alpha_1^T Z_{\mu i}(t), \tag{3.6}$$

where α_1 is the vector of the parameters associated with $Z_{\mu i}(t)$. The assumed variance function, $\mathrm{Var}[Y_i(t)|Z_{\mu i}(t)] = (\nu+1)^{-1}\mu_i(t)[1-\mu_i(t)]$, is based on $Y_i(t)$ following a beta$\big(\mu_i(t)\nu, [1-\mu_i(t)]\nu\big)$ distribution with probability density function,

$$f[y_i(t); \mu_i(t), \nu] = \frac{\Gamma(\nu)}{\Gamma[\mu_i(t)\nu]\Gamma\{[1-\mu_i(t)]\nu\}} {y_i}^{\mu_i(t)\nu - 1}(1-y_i)^{[1-\mu_i(t)]\nu - 1}$$

with corresponding cumulative density function $F[y_i(t); \mu_i(t), \nu]$. Later we will express $\mu_i(t)$ in terms of $\alpha = (\alpha_0, \alpha_1^T)^T$ and $Z_i^\mu(t) = [1, Z_{\mu i}^T(t)]^T$, that is, $\mu_i(t) = 1/\big[1 + e^{-\alpha^T Z_i^\mu(t)}\big]$. Although we've selected very robust distributions when specifying model-based variance functions corresponding to (3.5) and (3.6), GEE methodology allows for an additional layer of robustness based on sandwich variance estimation methods. Since models (3.5) and (3.6) are structured to model means $\pi_i(t)$ and $\mu_i(t)$, respectively, there is no proportionality assumption imposed on incidence rates for the recurrent event times as is assumed in proportional incidence and proportional means models for recurrent events.

Using Liang and Zeger's GEE methodology [76], the estimate of parameters β in model (3.5) may be obtained by solving the following estimating equation:

$$\sum_{i=1}^{n} \frac{\partial \pi_i}{\partial \beta}^T V_{\pi_i}^{-1}(B_i - \pi_i) = 0. \tag{3.7}$$

Here $B_i = [B_i(t_1), \ldots, B_i(t_b)]^T$, $\pi_i = [\pi_i(t_1), \ldots, \pi_i(t_b)]^T$ and $\pi_i(t) = 1/\big[1 + e^{-\beta^T Z_i^{\pi}(t)}\big]$, with $(\partial \pi_i(t)/\partial \beta)^T = Z_i^{\pi}(t) e^{\beta^T Z_i^{\pi}(t)}/\big[1 + e^{\beta^T Z_i^{\pi}(t)}\big]^2$; $V_{\pi_i} = A_{\pi i}^{\frac{1}{2}} R_{\pi i} A_{\pi i}^{\frac{1}{2}}$ is the variance matrix of B_i, where $A_{\pi i} = \text{Diag}\{\pi_i(t_1)[1-\pi_i(t_1)], \ldots, \pi_i(t_b)[1-\pi_i(t_b)]\}$ with the b^{th} element being the variance of $B_i(t_b)$, and $R_{\pi i}$ is the working correlation matrix for B_i. Similarly, the estimating equation for α in model (3.6) can be written as:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \alpha}^T V_{\mu_i}^{-1} U_{\mu_i}(Y_i - \mu_i) = 0. \tag{3.8}$$

Here $Y_i = [Y_i(t_1), \ldots, Y_i(t_b)]^T$, $\mu_i = [\mu_i(t_1), \ldots, \mu_i(t_b)]^T$ and $\mu_i(t) = 1/\big[1 + e^{-\alpha^T Z_i^{\mu}(t)}\big]$ with $(\partial \mu_i(t)/\partial \alpha)^T = Z_i^{\mu}(t) e^{\alpha^T Z_i^{\mu}(t)}/\big[1 + e^{\alpha^T Z_i^{\mu}(t)}\big]^2$; $V_{\mu_i} = A_{\mu i}^{\frac{1}{2}} R_{\mu i} A_{\mu i}^{\frac{1}{2}}$ is the variance matrix of Y_i, where $A_{\mu i}$ is a diagonal matrix with $\text{Var}[Y_i(t)] = (\nu + 1)^{-1}\mu_i(t)[1 - \mu_i(t)]$,$t = t_1, \ldots, t_b$, along the diagonal, $U_{\mu_i} = \text{diag}\{(1 - B_i(t_1), ..., (1 - B_i(t_b))\}$ and $R_{\mu i}$ is the working correlation matrix for Y_i. A variety of user specified working correlation structures for fitting these models are available. In addition to an unstructured working correlation matrix, users may consider a Toeplitz correlation structure similar to that described by Xia and Murray [70] or, if follow-up windows used to construct the censored longitudinal data do not overlap, an exchangeable working correlation matrix may be appropriate. The score equations (3.7) and (3.8) have no closed form solution; therefore, a two-stage iterative algorithm is required to estimate parameters $\hat{\beta}$ and $\hat{\alpha}$ and corresponding sandwich-based parameter covariance matrices, $\hat{V}_\beta$ and $\hat{V}_\alpha$ for models (3.5) and (3.6), respectively. Once convergence is achieved, $\hat{\alpha}$ and $\hat{\beta}$ are consistent estimators of α and β with an asymptotic multivariate normal distribution [75, 76].

In addition to parameter estimation, we may estimate the $\text{E}[T_i(t)]$ for each individual $i = 1, \ldots, n$ and each window $t = t_0, \ldots, t_b$ based on equation (3.3). Defining $\hat{\mu}_i(t) = 1/\big[1 + e^{-\hat{\alpha}^T Z_i^{\mu}(t)}\big]$ and $\hat{\pi}_i(t) = 1/\big[1 + e^{-\hat{\beta}^T Z_i^{\pi}(t)}\big]$, the estimated $\text{E}[T_i(t)]$ for subject i and window t becomes

$$\hat{\mathrm{E}}[T_i(t)] = \tau\hat{\mu}_i(t)[1-\hat{\pi}_i(t)] + \tau\hat{\pi}_i(t). \tag{3.9}$$

After some algebraic manipulation relegated to Appendix B.1,

$$\begin{aligned}\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[T_i(t)]\} &= \tau^2\Big[1-\frac{1}{1+e^{-\hat{\alpha}^T Z_i^{\mu}(t)}}\Big]^2 Z_i^{\pi}(t)^T\hat{V}_{\beta}Z_i^{\pi}(t)\frac{\big[e^{-\hat{\beta}^T Z_i^{\pi}(t)}\big]^2}{\big[1+e^{-\hat{\beta}^T Z_i^{\pi}(t)}\big]^4} \\ &+ \tau^2\Big[1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}(t)}}\Big]^2 Z_i^{\mu}(t)^T\hat{V}_{\alpha}Z_i^{\mu}(t)\frac{\big[e^{-\hat{\alpha}^T Z_i^{\mu}(t)}\big]^2}{\big[1+e^{-\hat{\alpha}^T Z_i^{\mu}(t)}\big]^4}.\end{aligned} \tag{3.10}$$

3.4 Algorithms for Censored Recurrent Event Times

In section 3.3, we introduced methodology for the the special case with uncensored data. In this section, we extend methods to the censored case using two approaches for fitting the recurrent event τ-inflated beta regression model: an ES algorithm-based approach described in section 3.4.1 and a multiple imputation approach described in section 3.4.2.

3.4.1 ES Algorithm

Due to the censoring process, $B_i(t)$ and $Y_i(t)$, $i = 1, ..., n, t = \{t_1, \ldots, t_b\}$ are potentially only partially observed for each individual, i. Hence, a likelihood-based approach to obtaining parameter estimates for the τ-IBR model will need to address the incomplete nature of the data.

The ES algorithm is a popular variant of the expectation-maximization (EM) algorithm [44, 18, 24] that may be used with censored longitudinal data such as ours, and correctly handles dependence between outcomes taken from the same individual over time. As with the EM algorithm, the ES algorithm is an iterative procedure with an expectation step (E-step) and a solution step (S-step) for obtaining updated parameter estimates. These steps are iterated until convergence of estimated model parameters according to predefined criteria.

We adopt an ES algorithm approach similar to that laid out by Rosen et al. [44], which leverages generalized estimating equation (GEE) methodology and can be briefly described as follows. Estimating equations, or solving equations, for the S-step are developed in three stages: (1) First a complete data log-likelihood that assumes working independence between outcomes taken from the same individual over time is derived. Then, (2) estimating equations based on this complete data log-likelihood are developed. Next, (3) the independence working correlation matrices that appear in the estimating equations in (2) are replaced with more general working correlation matrices, as desired. The E-step allows updates to the partially observed components of the solving equations prior to the next iteration of the S-step. The ES algorithm then iterates between the solution equations found in stage (3) and the E-step until convergence of model parameter estimates.

For the recurrent-event τ-IBR model, the ES algorithm approach takes the following form.

Stage 1 of developing the S-step: Let $\theta = (\alpha^T, \beta^T, \nu)^T$. Assuming an independent working correlation structure for outcomes taken from the same individual, the complete data log-likelihood becomes:

$$\begin{aligned} l(\theta) = \sum_{i=1}^{n} \sum_{t=t_1}^{t_b} \big\{ & B_i(t) \text{log}[\pi_i(t)(\beta)] + [1 - B_i(t)] \text{log}[1 - \pi_i(t)(\beta)] \\ & + [1 - B_i(t)] \text{log}[f_{Y_i(t)}[Y_i(t); \mu_i(t)(\alpha), \nu)] \big\}, \end{aligned} \tag{3.11}$$

where for individuals, i, with completely observed data, $B_i(t) = b_i(t)$ and $Y_i(t) = y_i(t)$, we substitute these observed values into the above expression.

Stage 2 of developing the S-step: Let I be a $b \times b$ identity matrix. Maximizing (3.11) in terms of β results in the solving equations:

$$\sum_{i=1}^{n} \frac{\partial \pi_i}{\partial \beta}^T (A_{\pi i}^{\frac{1}{2}} I A_{\pi i}^{\frac{1}{2}})^{-1} (B_i - \pi_i) = 0, \tag{3.12}$$

and maximizing (3.11) in terms of α results in the solving equations:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \alpha}^T (A_{\mu i}^{\frac{1}{2}} I A_{\mu i}^{\frac{1}{2}})^{-1} U_{\mu i} (Y_i - \mu_i) = 0, \tag{3.13}$$

where $U_{\mu_i} = \text{diag}\{(1 - B_i(t_1), ..., (1 - B_i(t_b))\}$.

Stage 3 of developing the S-step: Note that both equations (3.12) and (3.13) take the form of GEE (or weighted GEE) with an independence working correlation matrix, I. As in Rosen et al.[44], more general working correlation matrices can be substituted for I in equations (3.12) and (3.13), so that the final S-step solution equations become:

$$\sum_{i=1}^{n} \frac{\partial \pi_i}{\partial \beta}^T (A_{\pi i}^{\frac{1}{2}} R_{\pi i} A_{\pi i}^{\frac{1}{2}})^{-1} (B_i - \pi_i) = 0 \tag{3.14}$$

and

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \alpha}^T (A_{\mu i}^{\frac{1}{2}} R_{\mu i} A_{\mu i}^{\frac{1}{2}})^{-1} U_{\mu i} (Y_i - \mu_i) = 0, \tag{3.15}$$

where $R_{\pi i}$ and $R_{\mu i}$ are the working correlation matrices for B_i and Y_i respectively. Initial parameter estimates, $\hat{\theta}^{(0)}$, are based on solving equations (3.14) and (3.15) using the complete-case subset of the dataset.

The r^{th} iteration of the E-step updates $B_i(t)$ and $Y_i(t)$, $i = 1, ..., n$, $t = \{t_1, \ldots, t_b\}$ with their expectations given the observed data, $\{X_i(t) = x_i(t), \Delta_i(t) = \delta_i(t)\}, i = 1, ..., n, t = t_1, ...t_b)$ and the most recently updated parameter estimates, $\hat{\theta}^{(r-1)}$. For $B_i(t)$ this expectation takes the form:

$$\begin{aligned} w_i(t)^{(r-1)} &= \mathrm{E}[B_i(t)|\hat{\theta}^{(r-1)}, \{X_i(t) = x_i(t), \Delta_i(t) = \delta_i(t)\}, i = 1, ..., n, t = t_1, ...t_b)] \\ &= b_i(t)\delta_i(t) + [1-\delta_i(t)]\mathrm{E}[B_i(t)|\hat{\theta}^{(r-1)}, Y_i(t) \geq \tilde{y}_i(t)] \\ &= b_i(t)\delta_i(t) + [1-\delta_i(t)]\frac{\pi_i(t)\hat{\beta}^{(r-1)}}{\pi_i(t)\hat{\beta}^{(r-1)} + [1-\pi_i(t)\hat{\beta}^{(r-1)}][1-F_{Y_i(t)}(\tilde{y}_i(t); \mu_i(\hat{\alpha}^{(r-1)}), \hat{\nu}^{(r-1)})]}, \end{aligned}$$

and for $Y_i(t)$ this expectation takes the form:

$$\begin{aligned} \zeta_i(t)^{(r-1)} &= \mathrm{E}[Y_i(t)|\hat{\theta}^{(r-1)}, \{X_i(t) = x_i(t), \Delta_i(t) = \delta_i(t)\}, i = 1, ..., n, t = t_1, ...t_b)] \\ &= y_i(t)\delta_i(t) + [1-\delta_i(t)]\mathrm{E}[Y_i(t)|\hat{\theta}^{(r-1)}, Y_i(t) \geq \tilde{y}_i(t)] \\ &= y_i(t)\delta_i(t) + [1-\delta_i(t)]\frac{\int_{\tilde{y}_i(t)}^{\infty} y_i(t)\mathrm{d}F_{Y_i(t)}[y_i(t); \mu_i(t)(\hat{\alpha}^{(r-1)}), \hat{\nu}^{(r-1)}]}{\int_{\tilde{y}_i(t)}^{\infty} \mathrm{d}F_{Y_i(t)}[y_i(t); \mu_i(t)(\hat{\alpha}^{(r-1)}), \hat{\nu}^{(r-1)}]}. \end{aligned}$$

The following (r^{th}) iteration of the S-step replaces B_i with $w_i^{(r-1)} = [w_i(t_1)^{(r-1)}, \ldots, w_i(t_b)^{(r-1)}]^T$ in solving equation (3.12) and replaces Y_i and $U_{\mu i}$ with $\zeta_i^{(r-1)} = [\zeta_i(t_1)^{(r-1)}, \ldots, \zeta_i(t_b)^{(r-1)}]^T$ and $U_{\mu i}^{(r-1)} = \mathrm{diag}[1 - w_i(t_1)^{(r-1)}, \ldots, 1 - w_i(t_b)^{(r-1)}]$, respectively, in solving equation (3.13).

We use the geem2 function from the R mmmgee package [33] to conduct this S-step, with outcomes $w_i^{(r-1)}$ for estimating $\hat{\beta}^{(r)}$ and correlation parameters in $\hat{R}_{\pi i}^{(r)}$ and with outcomes $\zeta_i^{(r-1)}$ and weights $U_{\mu i}^{(r-1)}$ for estimating $\hat{\alpha}^{(r)}$, $\hat{\nu}^{(r)}$ and correlation parameters in $\hat{R}_{\mu i}^{(r)}$. The E step and S step iterate until $|\hat{\theta}^{(r)} - \hat{\theta}^{(r-1)}| < \epsilon$ for some $\epsilon > 0$; we used $\epsilon = 10^{-4}$ for results given in this manuscript. Hereafter, $\hat{\eta}^{ES} = (\hat{\alpha}^{ES}, \hat{\beta}^{ES}, \hat{\nu}^{ES}, \hat{R}_{\mu i}^{ES}, \hat{R}_{\pi i}^{ES})$ denotes a set of vectors including all final model parameter estimates and correlation parameters obtained via this ES algorithm approach.

We follow the Kong et al. [24] approach for obtaining the estimated variance-covariance matrix of the parameters $\hat{\alpha}^{ES}$ and $\hat{\beta}^{ES}$, which appropriately takes into account the variability related to the partially missing (censored) outcomes. In particular, let $\psi = (\alpha^T, \beta^T)^T$ be the combined parameter vector of interest with corresponding ES estimates $\hat{\psi}^{ES}$. Then $\widehat{\mathrm{Var}}(\hat{\psi}^{ES})$ can be consistently estimated by $\hat{H}^{-1}\hat{Q}\hat{H}^{-1}$, where

$$\hat{Q} = \sum_{i=1}^{n} S_i(\hat{\eta}^{ES})S_i(\hat{\eta}^{ES})^T, \hat{H} = \sum_{i=1}^{n} \frac{\partial S_i(\eta)}{\partial \psi}\Big|_{\hat{\eta}^{ES}} \text{ and } S_i(\eta) = \begin{pmatrix} \frac{\partial \pi_i}{\partial \beta}^T V_{\pi_i}^{-1}(w_i - \pi_i) \\ \frac{\partial \mu_i}{\partial \alpha}^T V_{\mu_i}^{-1}\mathrm{Diag}(1 - w_i)(\zeta_i - \mu_i) \end{pmatrix}.$$

Based on the formula provided by Satten and Datta (2000)[50], $\partial S_i(\eta)/\partial\psi$ can be written as:

$$
\begin{aligned}
\frac{\partial S_i(\eta)}{\partial\psi} &= \begin{pmatrix} -\frac{\partial\pi_i}{\partial\beta}^T V_{\pi_i}^{-1}\frac{\partial\pi_i}{\partial\beta} & 0 \\ 0 & -\frac{\partial\mu_i}{\partial\alpha}^T V_{\mu_i}^{-1}\text{Diag}(1-w_i)\frac{\partial\mu_i}{\partial\alpha} \end{pmatrix} \\
&+ \begin{pmatrix} \frac{\partial\pi_i}{\partial\beta}^T V_{\pi_i}^{-1} \\ -\frac{\partial\mu_i}{\partial\alpha}^T V_{\mu_i}^{-1}\text{Diag}(\zeta_i-\mu_i) \end{pmatrix} \text{Var}(B_i|\zeta_i) \begin{pmatrix} \frac{\partial\pi_i}{\partial\beta}^T V_{\pi_i}^{-1} \\ -\frac{\partial\mu_i}{\partial\alpha}^T V_{\mu_i}^{-1}\text{Diag}(\zeta_i-\mu_i) \end{pmatrix}^T
\end{aligned}
$$

where $\text{Var}(B_i|\zeta_i) = \text{Diag}\{w_i(t)[1-w_i(t)]\}^{1/2}R_{\pi i}\text{Diag}\{w_i(t)[1-w_i(t)]\}^{1/2}$. ES-based $\text{E}[T_i(t)]$ and corresponding variance estimates are constructed using equations (3.9) and (3.10), substituting elements of $\hat{\psi}^{ES}$ and $\widehat{\text{Var}}(\hat{\psi}^{ES})$ for $\hat{\alpha}$, $\hat{\beta}$, $\hat{V}_\alpha$ and $\hat{V}_\beta$, as appropriate.

3.4.2 Multiple Imputation Algorithm For Censored Recurrent Event Times

In section 3.3, we introduced a new modeling framework for analyzing recurrent event data in the special case of no censoring. In this section, we describe a multiple imputation (MI) algorithm for generating uncensored data for each individual i that can be used in fitting models 3.5 and 3.6. The overall goal of MI is to generate M different imputed datasets, where imputed values for missing data are sampled from appropriate conditional distributions based on the observed data and are subject to the same variability as the fully observed data. Each of the M imputed datasets can be analyzed using methods for uncensored data described in section 3.3. In this manuscript, we set $M = 10$, which is usually sufficient for MI algorithms to produce results with good operating characteristics. Results from these M analyses are then combined using Rubin's method [45, 46], with further details to follow later in this section. Steps for implementing our MI algorithm to estimate $\theta = (\alpha^T, \beta^T, \nu)^T$ and $\text{Var}(\theta)$ are given below, with MI-based estimates denoted $\hat{\theta}^{MI}$ and $\widehat{\text{Var}}(\hat{\theta}^{MI})$, respectively. Once we obtain $\hat{\theta}^{MI}$ and $\widehat{\text{Var}}(\hat{\theta}^{MI})$, they can replace corresponding terms given in section 3.3 for estimating $\hat{\text{E}}[T_i(t)]$ and $\widehat{\text{Var}}\{\hat{\text{E}}[T_i(t)]\}$, respectively.

In our setting a censored patient, i, requires imputation of times-to-first-event in the set, S_i, of τ-length follow-up windows starting at $\{t \in \{t_1, \ldots, t_b\} : 0 < X_i(t) < \tau, \Delta_i(t) = 0\}$. If the imputation set S_i consists of more than one follow-up window, we need only impute the time-to-first-event for the last window in S_i, which is denoted as the b_i^{*th} follow-up window in the data set. For better understanding, we set t_i^{sup} as the start time of the b_i^{*th} window and $\tilde{T}_i(t_i^{sup})$ as the imputed τ-restricted event time for the window starting at t_i^{sup}. Then for other windows in S_i with start time $t < t_i^{sup}$, the imputed τ-restricted event times become $\tilde{T}_i(t) = \min[\tilde{T}_i(t_i^{sup}) + t_i^{sup} - t, \tau]$. According to the imputed $\tilde{T}_i(t)$ for each window in S_i, the outcomes of model (3.5) and model (3.6) can be imputed as $\tilde{B}_i(t) = \text{I}[\tilde{T}_i(t) = \tau]$ and $\tilde{Y}_i(t) = \tilde{T}_i(t)$ if $\tilde{B}_i(t) = 0$.

The general idea of the imputation procedure is to generate $\tilde{T}_i(t^*)$ by inverse probability transform imputation (IPTI) from a risk set of individuals with similar covariates to patient i at the window starting at t_i^{sup}. The only parametric component of this MI algorithm is the construction of risk set, which depends on the model being fitted. Step 1 describes how we define the risk sets and step 2 describes the imputation procedure based on these risk sets. Step 3 produces M imputed data sets for analysis. Details of these steps are given below.

$Step$ 1: (Risk set definition step) For each individual i, $i = 1, \dots, N_c$, requiring imputation for censored τ-restricted event time $\tilde{T}_i(t_i^{sup})$ in the b_i^{*th} window, the sup window, we define a risk set $\mathcal{R}_i$ of individuals that are similar to individual i. We consider two cases, $b_i^* = 1$ and $b_i^* > 1$. For the case $b_i^* = 1$, the $j = 1, 2, \dots, N_i$ individuals included in the risk set $\mathcal{R}_i$ need to satisfy two constraints: (a) max$[|\hat{\mu}_i(t_i^{sup}) - \hat{\mu}_j(t_i^{sup})|, |\hat{\pi}_i(t_i^{sup}) - \hat{\pi}_j(t_i^{sup})|] < \epsilon$, where $\hat{\mu}_i(t_i^{sup})$ and $\hat{\pi}_i(t_i^{sup})$ are taken from $\hat{\theta}^{ES}$ and (b) $X_j(t_i^{sup}) > X_i(t_i^{sup})$. Condition (b) in this step ensures that all subjects in $\mathcal{R}_i$ are at risk when individual i is censored. Condition (a) requires that individuals in $\mathcal{R}_i$ have similar predicted outcomes to the individual i in the current iteration of fitting model (3.5) and model (3.6). We set $\epsilon = 0.05$ in condition (a) and increase ϵ by 0.005 until either $N_i \geq 15$ or $\epsilon > 0.5$. If the last observed event time in $\mathcal{R}_i$ is a censored value $< \tau$, then we continue increasing ϵ by 0.001 until this is no longer the case. For the case $b_i^* > 1$, it is possible to include additional constraints related to a potentially quite sophisticated history for that individual, subject to available sample size. For instance in the azithromycin study, it may be attractive to incorporate information on previous exacerbations in creating the risk set. Suppose the random variable pairs $\{Y_i(t); B_i(t)\}$ for $t = 0, \dots, t_{b'_i}$ are known. One may, for instance, create a binary history variable $\bar{B}_i(t)$, $i = 1, \dots, n$, that indicates whether the individual has ever met $B_i(t) = 1$ for $t = 0, \dots, t_{b'_i}$, and required individuals included in the risk set to have $\bar{B}_j(t) = \bar{B}_i(t)$.

$Step$ 2: (IPTI step) In this step, for each censored individual i that requires imputation, we impute a value of $T_i(t_i^{sup})$ for the last window in S_i. This determines imputed $T_i(t)$ values for remaining windows in S_i as well as corresponding $\{B_i(t), Y_i(t)\}$ for each window in S_i. First, from individuals in $\mathcal{R}_i$ with data $\{X_j(t_i^{sup}), \Delta_j(t_i^{sup})\}$, we use the nonparametric Kaplan-Meier approach to obtain the estimated survival function $\hat{S}_{T_i(t_i^{sup})}(v|\mathcal{R}_i)$. Second, generate a uniform(0,1) random variable, u_i, and find the smallest observed event time v^* where $\hat{S}_{T_i(t_i^{sup})}(v^*|\mathcal{R}_i) \leq u_i$. If $v^* = \tau$, then impute 1 for $B_i(t_i^{sup})$; no further imputation for $Y_i(t_i^{sup})$ is required. Otherwise, impute 0 for $B_i(t_i^{sup})$ and v^* for $Y_i(t_i^{sup})$. Completing this step results in a fully imputed dataset.

$Step$ 3: (Multiple imputation step) Repeat step 2 M times, resulting in M imputed data sets to be analyzed. By fitting model (3.5) and model (3.6) using each imputed data set, we obtain M parameter estimates $\hat{\theta}_m^{MI}$ with corresponding estimated covariance matrix $\widehat{\text{Var}}(\theta_m^{\hat{M}I})$, $m = 1, \dots, M$. We combine the results from $M = 10$ imputed data sets using the approach proposed bu Rubin [45, 46]. The final estimate of θ based on M parameter estimates becomes $\hat{\theta}^{MI} = \sum_{m=1}^{M} \hat{\theta}_m^{MI}$ with corresponding estimated covariance matrix

$$\widehat{\text{Var}}(\hat{\theta}^{MI}) = \bar{U} + (1 + M^{-1})\bar{B},$$

where

$$\bar{U} = \tfrac{1}{M} \textstyle\sum_{m=1}^{M} \widehat{\text{Var}}(\hat{\theta}_m^{MI})$$

is the estimated within imputation variance and

$$\bar{B} = \tfrac{1}{M-1} \textstyle\sum_{m=1}^{M} (\hat{\theta}_m^{MI} - \hat{\theta}^{MI})(\hat{\theta}_m^{MI} - \hat{\theta}^{MI})^T$$

is the estimated between imputation variance. The terms, $\hat{V}_{\alpha}^{MI}$ and $\hat{V}_{\beta}^{MI}$, can be extracted from $\widehat{\text{Var}}(\hat{\theta}^{MI})$, as appropriate.

3.5 Simulation Study

In this section we evaluate the finite sample performance of our proposed τ-IBR Model in terms of (i) the quality of parameter estimates from models (3.5) and (3.6) and (ii) the quality of longitudinal τ-RMST estimates for individuals i, $i = 1, \ldots, n$ and follow-up windows starting at $t = \{t_1 \ldots, t_b\}$. Section 3.5.1 describes our simulation framework and gives data generation details for each of the 1000 iterations simulated per scenario. Section 3.5.2 summarizes simulation results.

The quality of parameter estimates for models (3.5) and (3.6) are evaluated in terms of bias, the average model-based standard error (ASE), the empirical standard deviation of the parameter estimates (ESD) and the average coverage probability of model-based 95% confidence intervals (CP). The quality of longitudinal τ-RMST estimates across simulations based on n=100, 500, 1000 and 1500 subjects was evaluated by comparing bias, the empirical mean squared error (EMSE), ASE, and CP using the proposed τ-IBR model to those based on the Xia, Murray and Tayob model (XMT Model) for recurrent event data. Since true τ-RMST values are specific to each individual i, $i = 1, \ldots, n$ for each window t, $t = t_1, ..., t_b$, summary statistics for the performance of longitudinal τ-RMST estimates are defined as

$$\text{Bias} = \sum_{j=1}^{1000} \sum_{i=1}^{n} \sum_{t=t_1}^{t_b} \frac{\hat{\text{E}}[T_{ij}(t)] - \text{E}[T_{ij}(t)]}{n \times b \times 1000},$$

$$\text{EMSE} = \sum_{j=1}^{1000} \sum_{i=1}^{n} \sum_{t=t_1}^{t_b} \frac{\{\hat{\text{E}}[T_{ij}(t)] - \text{E}[T_{ij}(t)]\}^2}{n \times b \times 1000},$$

$$\text{ASE} = \sum_{j=1}^{1000} \sum_{i=1}^{n} \sum_{t=t_1}^{t_b} \frac{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}{n \times b \times 1000} \text{ and}$$

$$\text{CP} = \sum_{j=1}^{1000}\sum_{i=1}^{n}\sum_{t=t_1}^{t_b} \frac{I(\text{Lower}_{ij} < \hat{\text{E}}[T_{ij}(t)] < \text{Upper}_{ij})}{n \times b \times 1000},$$

where $\text{Lower}_{ij} = \hat{\text{E}}[T_{ij}(t)] - 1.96 \times \sqrt{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}$ and $\text{Upper}_{ij} = \hat{\text{E}}[T_{ij}(t)] + 1.96 \times \sqrt{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}$.

3.5.1 Data Generation

In the following, we assume that the censored longitudinal data structure is based on $\tau = 30$ and $b = 4$ follow-up windows with $t \in \{t_1, t_2, t_3, t_4\} = \{0, 30, 60, 90\}$. For each individual, $i = 1, ..., n$, we generate three uniform (0,1) covariates, Z_{1i}, Z_{2i} and Z_{3i}. We consider two settings for the true τ-IBR model:

$$\text{Setting 1: } \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = -1.832 + 0.839Z_{1i} + 1.980Z_{2i} + 1.012Z_{3i},$$
$$\log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = -1.023 + 0.654Z_{1i} + 1.678Z_{2i};$$

$$\text{Setting 2: } \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = 0.5,\ \log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = -1.023 + 0.654Z_{1i} + 1.678Z_{2i}.$$

The model for $\mu_i(t)$ is the same in each of these two settings, with $Z_{\mu_i} = \{Z_{1i}, Z_{2i}\}$ being the key covariates. The model for $\pi_i(t)$ varies in the two settings, with $Z_{\pi_i} = \{Z_{1i}, Z_{2i}, Z_{3i}\}$ being important covariates for Setting 1 and no important covariates in Setting 2. We would not expect advantages of our method compared to the Xia, Murray and Tayob model in Setting 2, since there are no interesting covariate features related to the point mass of τ-restricted event times at τ. However, advantages of our approach are expected to emerge from Setting 1.

Simulated outcomes for subject i are based on a beta distribution for $Y_i(t)$ with mean $\mu_i(t)$ and a Bernoulli distribution for $B_i(t)$ with mean $\pi_i(t)$. Correlated outcomes $\{B_i(0), B_i(30), B_i(60), B_i(90)\}$ and $\{Y_i(0), Y_i(30), Y_i(60), Y_i(90)\}$ for each individual i proceed by first generating correlated multivariate normal random variables and then coercing them into the desired distributions. For correlated $\{Y_i(0), Y_i(30), Y_i(60), Y_i(90)\}$, we use a Gaussian copula approach. First, multivariate standard normal random variables $\{G_i(0), G_i(30), G_i(60), G_i(90)\}$ with exchangeable correlation structure (correlation coefficient=0.3) were generated for each individual i. We then transform the multivariate normal random variables $G_i(t)$ to multivariate uniform(0,1) variables $U_i(t)$ via $U_i(t) = \Phi(G_i(t))$, where $t = 0, 30, 60, 90$ and $\Phi()$ is the cumulative density function of the standard normal distribution. The last step is to use the inverse transform theorem to obtain multivariate beta random variables $Y_i(t) = F^{-1}(U_i(t))$, where $F^{-1}()$ is the inverse of cumulative density function of Beta$[\mu_i(t)\nu, (1-\mu_i(t))\nu]$ distribution, where $\nu = 3$ in all simulations.

To generate correlated binary random variables, $\{B_i(0), B_i(30), B_i(60), B_i(90)\}$, we adopt an algorithm proposed by Emrich and Piedmonte [10] that is implemented using the function rmvbin from the R package bindata. The general idea of this approach is to first generate multivariate standard normal random variables $\{N_i(0), N_i(30), N_i(60), N_i(90)\}$ that are transformed into multivariate binary random variables by setting $B_i(t) = I[N_i(t) > a(t)]$, where $t = 0, 30, 60, 90$, $a(t) = \Phi^{-1}[1 - \pi_i(t)]$ and $\Phi^{-1}()$ is the inverse cumulative density function of the standard normal distribution. The resulting binary random variables, $\{B_i(0), B_i(30), B_i(60), B_i(90)\}$, will have the desired marginal Bernoulli distributions with means $\{\pi_i(0), \pi_i(30), \pi_i(60), \pi_i(90)\}$, respectively. In our simulations, we assumed exchangeable correlation structure $R_{\pi i}$ with a correlation of 0.3 between dependent $B_i(t_1), B_i(t_2), t_1 \neq t_2$, in our simulations. Details of how the user-specified correlation is calibrated to the correlation of the multivariate normal random variables used in this algorithm is provided in Appendix B.2.

For the models assumed in Settings 1 and 2, we evaluated model performance with 0% and 20% of subjects censored prior to $t_b + \tau$. For the 20% censoring case, we independently generated a Bernoulli(0.7) random variable, B_i^*, and a Uniform$(0, b\tau)$ random variable, U_i^*. Then the censoring random variable was defined as $C_i = U_i^*(1 - B_i^*) + b\tau B^*, i = 1, \ldots, n$.

3.5.2 Simulation Results

Table 3.1 displays finite sample properties of τ-IBR model parameter estimates for both the uncensored case and the case with 20% censoring with $n = 500$ simulated individuals. Both MI and ES parameter estimation procedures perform well in simulation. Coverage probabilities are close to the desired 95% level using either the MI or the ES method, and ASE and ESD estimates are very close to one another. The bias of parameter estimates is generally small, although the ES intercept estimate for model (3.6) shows slightly higher bias than the MI estimate for the same term. As one would expect, the variability of parameter estimates from the model with binary outcomes (model 3.5) is somewhat larger than the variability of parameter estimates from the model with continuous outcomes (model 3.6).

Table 3.2 shows finite sample properties of τ-RMST estimates using both our proposed τ-IBR methodology and the XMT model for sample sizes of $n = 100, 500, 1000$ and 1500. We will first briefly summarize results related to bias and then analyze coverage probabilities for the true τ-RMST, which varied by method and simulation setting. In terms of bias, both τ-IBR MI and XMT modeling approaches yield approximately unbiased τ-RMST estimates. The ES-based τ-

Table 3.1: Finite sample performance of τ-IBR parameter estimates from models (3.5) and (3.6) for n=500 subjects with correlated longitudinal outcomes ($\rho = 0.3$) based on 1000 iterations.

	No Censoring (GEE)				**20% Censoring (MI)**				**20% Censoring (ES)**			
	Bias	**ASE**	**ESD**	**CP**	**Bias**	**ASE**	**ESD**	**CP**	**Bias**	**ASE**	**ESD**	**CP**
Setting 1												
α_0 = -1.023	-0.005	0.106	0.105	0.947	-0.004	0.112	0.114	0.939	0.041	0.119	0.116	0.929
α_1 = 0.654	0.001	0.144	0.143	0.949	0.009	0.152	0.155	0.949	0.011	0.162	0.158	0.954
α_2 = 1.678	0.012	0.150	0.151	0.957	0.002	0.159	0.165	0.940	0.007	0.169	0.167	0.945
β_0 = -1.832	-0.011	0.218	0.216	0.949	0.007	0.231	0.233	0.950	-0.009	0.230	0.235	0.950
β_1 = 0.839	0.008	0.228	0.232	0.948	0.000	0.241	0.235	0.957	0.004	0.239	0.239	0.953
β_2 = 1.980	0.011	0.234	0.236	0.951	0.019	0.248	0.249	0.944	0.023	0.246	0.250	0.940
β_3 = 1.012	0.006	0.229	0.227	0.954	-0.026	0.242	0.245	0.955	-0.010	0.239	0.248	0.948
Setting 2												
α_0 = -1.023	-0.008	0.121	0.126	0.941	-0.004	0.129	0.134	0.933	0.034	0.141	0.134	0.951
α_1 = 0.654	0.011	0.156	0.162	0.938	0.010	0.167	0.171	0.942	0.015	0.179	0.173	0.952
α_2 = 1.678	0.005	0.159	0.167	0.935	0.002	0.170	0.183	0.933	0.012	0.182	0.184	0.950
β_0 = 0.500	0.005	0.202	0.201	0.953	0.003	0.214	0.207	0.961	0.002	0.212	0.210	0.959
β_1 = 0.000	-0.005	0.221	0.218	0.953	-0.002	0.234	0.225	0.957	-0.004	0.232	0.232	0.940
β_2 = 0.000	0.010	0.222	0.220	0.947	0.014	0.234	0.236	0.937	0.011	0.232	0.238	0.937
β_3 = 0.000	-0.010	0.221	0.220	0.956	-0.010	0.235	0.229	0.952	-0.010	0.232	0.237	0.943

Bias is the average difference between the true and estimated parameters across the simulations; ASE is the average of the model-based standard error estimates across the simulations; ESD is empirical standard deviation of the parameter estimates seen in simulation; CP is the empirical coverage probability of the true parameter by the model-based 95% confidence interval seen in simulation.

IBR τ-RMST estimates show slightly more bias than the other estimation methods, likely due to the intercept term estimates from model (3.6) being slightly off using the ES estimation approach.

Coverage probabilities for true τ-RMST values were consistently good for the τ-IBR MI estimation method in all settings. In setting 2, where there are no interesting covariate associations related to the point mass of restricted event times at $\tau = 30$, coverage was very good for both the τ-IBR MI and XMT approaches. Coverage for the τ-IBR ES approach lagged slightly behind when compared to these other two methods, but was adequate.

In setting 1, where covariate associations related to the point mass of restricted event times at $\tau = 30$ are in play, coverage for the τ-IBR ES approach was adequate, but not as good as coverage for the τ-IBR MI estimation method. In contrast, the XMT approach had remarkably poor coverage for the true τ-RMST values that deteriorated further with increasing sample sizes (despite both ASE and EMSE getting smaller with increasing sample sizes). The EMSE was particularly high for the XMT method relative to other methods in Setting 1.

Figure 3.2 displays differences between estimated and true τ-RMST values using τ-IBR MI and XMT methods for representative Setting 1 datasets with n=100, 500, 1000 and 1500, highlighting subjects whose true τ-RMST were not covered by their estimated 95% confidence interval. Al-

Table 3.2: Finite sample performance of τ-RMST estimates using (1) the τ-IBR model and (2) the XMT model based on 1000 iterates

	No censoring (Setting1)		**20% censoring (Setting1)**			**No censoring (Setting2)**		**20% censoring (Setting2)**		
	τ-IBR (GEE)	**XMT**	**τ-IBR (MI)**	**τ-IBR (ES)**	**XMT**	**τ-IBR (GEE)**	**XMT**	**τ-IBR (MI)**	**τ-IBR (ES)**	**XMT**
1500 Subjects										
Bias	0.004	0.004	0.012	0.158	0.009	0.009	0.009	0.013	0.123	0.010
EMSE	0.082	0.491	0.093	0.118	0.503	0.080	0.087	0.087	0.102	0.095
ASE	0.270	0.289	0.286	0.284	0.302	0.265	0.280	0.281	0.278	0.293
CP	0.950	0.568	0.947	0.907	0.582	0.947	0.944	0.950	0.914	0.943
1000 Subjects										
Bias	-0.002	-0.001	0.004	0.149	0.001	-0.002	-0.002	0.000	0.109	-0.003
EMSE	0.129	0.541	0.135	0.156	0.547	0.118	0.129	0.126	0.138	0.139
ASE	0.331	0.353	0.350	0.348	0.370	0.324	0.343	0.344	0.340	0.359
CP	0.945	0.652	0.950	0.924	0.671	0.948	0.946	0.952	0.930	0.947
500 Subjects										
Bias	0.005	0.004	0.014	0.161	0.010	0.003	0.003	0.006	0.118	0.005
EMSE	0.244	0.661	0.275	0.298	0.690	0.230	0.245	0.252	0.265	0.272
ASE	0.466	0.498	0.493	0.490	0.521	0.458	0.484	0.485	0.480	0.506
CP	0.947	0.780	0.947	0.928	0.792	0.948	0.950	0.952	0.935	0.949
100 Subjects										
Bias	0.002	0.006	-0.002	0.148	-0.001	0.017	0.020	0.019	0.133	0.020
EMSE	1.263	1.720	1.336	1.334	1.790	1.150	1.226	1.259	1.262	1.341
ASE	1.019	1.079	1.089	1.082	1.141	1.004	1.054	1.063	1.056	1.105
CP	0.926	0.898	0.939	0.931	0.910	0.931	0.937	0.933	0.924	0.936

Bias: average difference between the true and predicted τ-RMST values across all subjects, windows and simulations; EMSE: empirical mean squared error of τ-RMST values across all subjects, windows and simulations; ASE: average of the model-based standard error estimates corresponding to the τ-RMST estimates across all subjects, windows and simulations; CP: empirical coverage probability of the true τ-RMST value by the model-based 95% confidence interval across all subjects, windows and simulations.

though both methods give unbiased τ-RMST estimates on average in Setting 1, XMT τ-RMST estimates had large differences from their true τ-RMST values observed that did not improve with larger sample sizes and were not correctly accounted for in corresponding model-based variance estimates given by the XMT method. Hence in Setting 1, where there is important statistical information related to the point mass of restricted event times at $\tau = 30$, ignoring this statistical information had a strong impact on the performance of the XMT method.

3.6 Azithromycin for Prevention of COPD Exacerbations Trial

In this section, we apply the proposed τ-IBR model to analyze the effect of azithromycin on the longitudinal time to the first acute exacerbation in participants with chronic obstructive pulmonary disease (COPD). This study enrolled COPD patients with a history of recurrent acute exacerbations and randomized them to take a daily dose of either 250 mg azithromycin or placebo over a one year follow-up period. Primary and secondary analyses were based on logrank test and multivari-

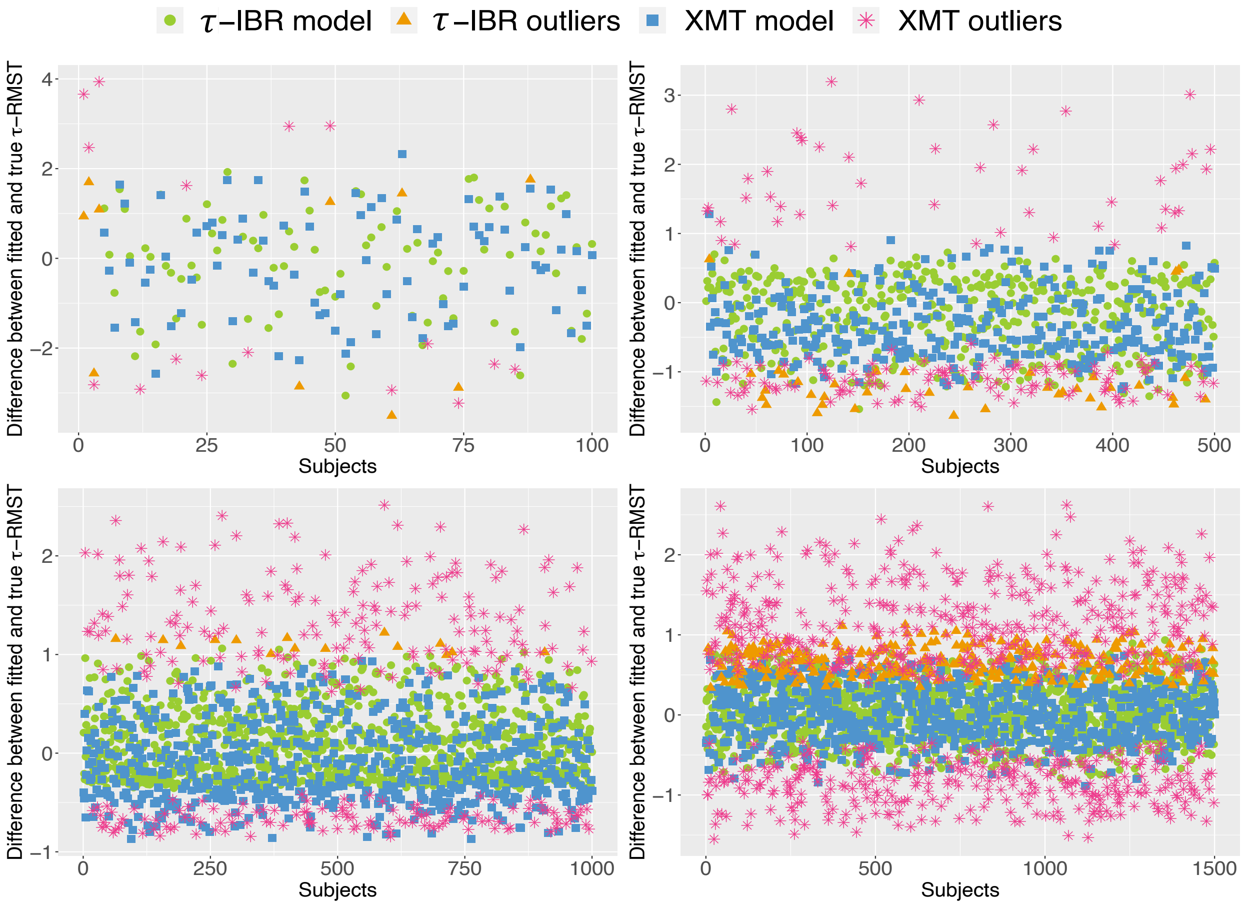

Figure 3.2: Difference between fitted τ-RMST and actual τ-RMST via τ-IBR MI-fitted model and XMT model in Setting 1 based on (a) 100 subjects (b) 500 subjects (c) 1000 subjects and (d) 1500 subjects.

able Cox model analysis of treatment effect, respectively. Azithromycin was found to be beneficial using these standard methods and is in common use today. Our τ-IBR model analysis focuses on the 1112 participants who were available for multivariable modeling with adjustments for age at randomization [in decades and centered at 65 years], sex [male=1, female=0], baseline forced expiratory volume in one second (FEV$_1$) [in percentage of predicted units by tens and centered at 40% of predicted], baseline smoking status [current=1, former=0] and study site. Individual i's baseline covariates as described above are denoted Z_i in what follows. When referring to the average patient profile in results below, this participant was 65 years old, with baseline FEV$_1$%=40%, a 59% probability of being male, and a 22% probability of being a current smoker at baseline.

Recurrent event data was reconfigured into a censored longitudinal dataset with $\tau = 6$ months and follow-up window start times at 0, 60, 120 and 180 days based on advice given in Xia and Murray[67]. Closed-form calculations in the setting with exponential recurrent event times showed that on average 90% of recurrent events are captured in at least one follow-up window when τ is set at the historical mean exacerbation-free time with window spacing every $\tau/3$ units.

Multiply imputed datasets using the approach given in section 3.4.2 with $Z_{\pi i}(t) = Z_{\mu i}(t) = Z_i$ were used to create the heatmaps displayed in Figures 3.3 (a) and 3.3 (b). Heatmap entries in Figures 3.3 (a) and 3.3 (b) are individual 6-month restricted times to first acute exacerbation averaged across 10 multiply imputed datasets, where in Figure 3.3 (a) these entries are additionally averaged over the 4 follow-up windows and in Figure 3.3 (b) the values are shown separately for the 4 follow-up windows. In each heatmap, longer and shorter 6-month restricted exacerbation-free times are in the yellow and purple color ranges, respectively. Additional individual characteristics, Z_i, are color coded along the left side of the heatmap. Descriptively from Figure 3.3 (a), the cluster with the highest average 6-month restricted exacerbation-free time is more likely to be taking azithromycin, has more male participants, and tends to have FEV$_1$%>50% of predicted. We will return to Figure 3.3 (b) later in this section. These heatmaps, which we believe are the first of their kind for censored recurrent event time data, gives us a fairly robust view of the raw τ-restricted outcome data plotted alongside predictors of interest. Parametric assumptions used in imputing values for the censored follow-up windows only appear in the selection of risk set participants used in the otherwise nonparametric inverse transform imputation step.

Table 3.3 displays 6-month-IBR parameter estimates for models (3.5) and (3.6) that allow interpretation of the azithromycin effect adjusted for baseline covariates; parameter estimates for study site have been submerged in this table. Results were similar using either the MI or ES estimation approach, hence, for brevity, we focus on MI estimation results in what follows. The odds ratio of remaining exacerbation-free for any 6-month period when taking azithromycin versus placebo was 1.448 (95% CI: 1.194-1.756, p<0.001), adjusted for other factors in the model. Amongst those who experienced an exacerbation during a 6-month follow-up period, those taking azithromycin had a 1.053 longer exacerbation-free period (95% CI: 0.993-1.123) that was marginally significant (p = 0.073), adjusted for other factors in the model. That is, the 6-month-IBR model suggests that those taking azithromycin were generally less susceptible to having exacerbations in any 6-month period (model 3.5), but that amongst those who had exacerbations the 6-month-restricted exacerbation-free period was marginally longer by approximately 5% (model 3.6). Using equations (3.9) and (3.10) as described in section 3.3, estimated 6-month exacerbation-free times for the average patient profile were 136 days (95% CI 128-144 days) and 125 days (95% CI 117-132 days) for the azithromycin and placebo groups, respectively, with a treatment difference of 11 days (95% CI 6-17 days; p<0.001) favoring azithromycin.

From Figure 3.3 (b), we observe a large number of individuals remaining exacerbation-free during the entire study duration (solid yellow area in blue box) but many different patterns exist, including individuals with progressively shorter and longer exacerbation-free periods as well as

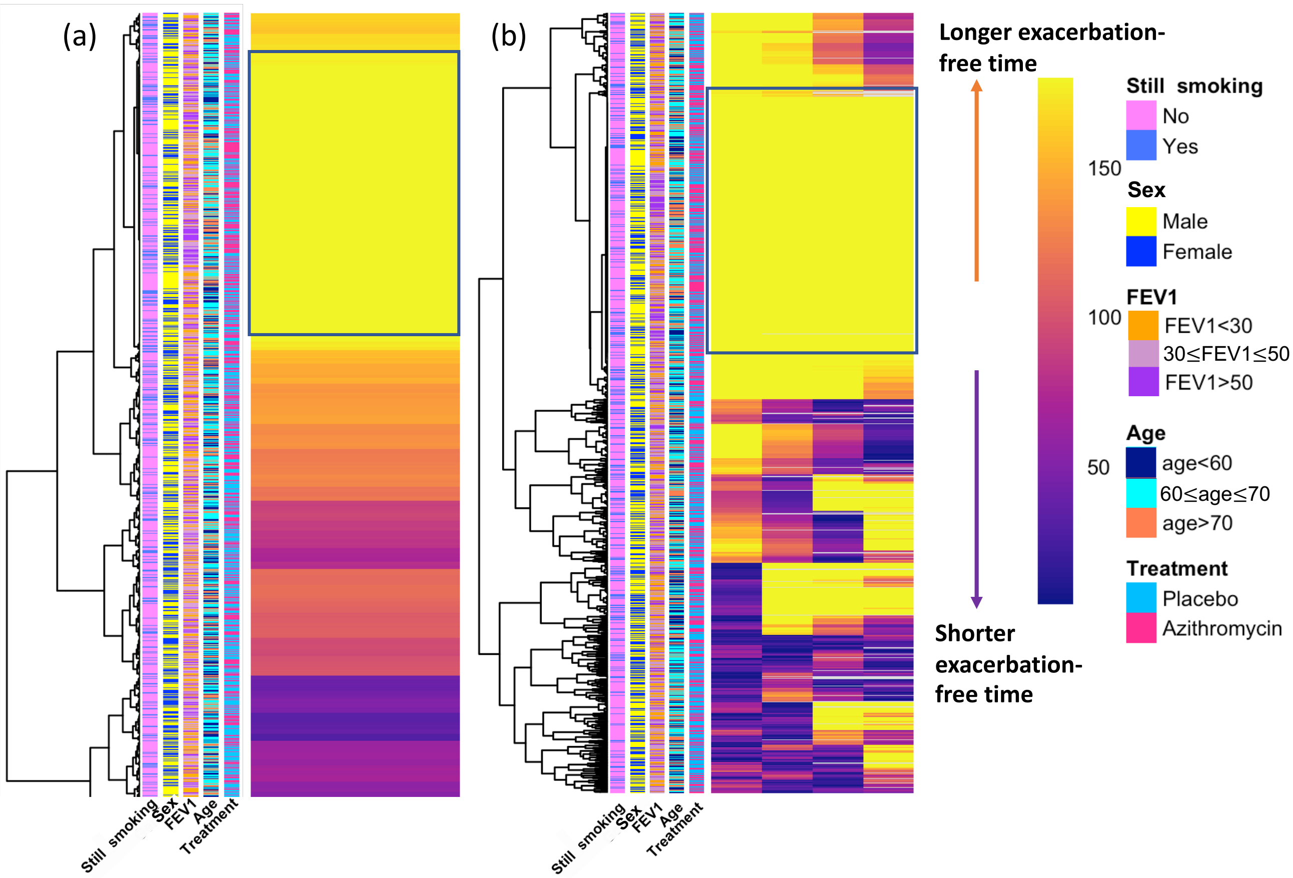

Figure 3.3: Six-month time-to-first-exacerbation for participants in the Azithromycin for Prevention of COPD Exacerbations Trial. Heatmap (a) entries are individual 6-month restricted times-to-first-exacerbation averaged across follow-up windows from 10 multiply imputed datasets using the τ-IBR method of imputation. Heatmap (b) entries show individual 6-month restricted times-to-first-exacerbation by follow-up window averaged across 10 multiply imputed datasets using the τ-IBR method of imputation.

some individuals with consistently short exacerbation-free periods that show little improvement over time. Stability of the treatment effect over the follow-up time windows may be evaluated by introducing window start times as predictors in the models with interaction terms, as appropriate. Figure 3.4, panels (a)-(d), show estimated 6-month restricted times-to-first-exacerbation by follow-up window start time, treatment group and selected subgroups of interest based on FEV$_1$ percent of predicted [panel (a)], age [panel (b)], sex [panel (c) and baseline smoking status [panel (d)]. At the far right of these panels, overall 6-month restricted times-to-first-exacerbation estimates for these subgroups are given from models not including window start times as covariates. Corresponding differences in estimated 6-month restricted times-to-first-exacerbation between azithromycin and placebo groups are shown in panels (e)-(h) of Figure 3.4, along with 95% confidence intervals. Although this study was not powered to detect subgroup differences, some interesting patterns are

Table 3.3: Estimated 6-month-IBR and 6-month-XMT multivariate model parameters with 95% confidence intervals and p-values. All models are additionally adjusted for study site (data not shown).

	Azithromycin	**Age**	**Male**	**FEV$_1$**	**Current Smoker**
	(vs. Placebo)	**(per 10 Years)**	**(vs. Female)**	**(per 10% Predicted)**	**(vs. Ex)**
τ-IBR model (ES) (Beta Regression)					
Fold Change*	1.053	1.041	1.039	1.009	1.025
95% CI	(0.984, 1.122)	(1.000, 1.082)	(0.970, 1.109)	(0.985, 1.032)	(0.937, 1.113)
P-value	0.122	0.053	0.254	0.469	0.577
τ-IBR model (ES) (Logistic Regression)					
Odds Ratio†	1.457	1.104	1.410	1.134	1.107
95% CI	(1.206, 1.760)	(0.978, 1.246)	(1.163, 1.708)	(1.067, 1.207)	(0.862, 1.422)
P-value	<0.001	0.109	<0.001	<0.001	0.426
τ-IBR model (MI) (Beta Regression)					
Fold Change*	1.058	1.041	1.037	1.008	1.020
95% CI	(0.993, 1.123)	(1.001, 1.081)	(0.972, 1.102)	(0.987, 1.030)	(0.939, 1.101)
P-value	0.073	0.045	0.252	0.449	0.628
τ-IBR model (MI) (Logistic Regression)					
Odds Ratio†	1.448	1.112	1.408	1.129	1.117
95% CI	(1.194, 1.756)	(0.986, 1.255)	(1.159, 1.710)	(1.060, 1.202)	(0.867, 1.438)
P-value	<0.001	0.084	<0.001	<0.001	0.392
XMT Model					
Coef/$\tau^\star$	0.060	0.025	0.062	0.018	0.022
95% CI	(0.029, 0.092)	(0.004, 0.045)	(0.029, 0.094)	(0.008, 0.028)	(-0.019, 0.062)
P-value	<0.001	0.016	<0.001	0.001	0.300

*Among those experiencing an exacerbation during the 6 months of follow-up, fold change is the ratio of estimated exacerbation-free time during the year when comparing those with versus without a one unit increase in the predictor, assuming all other predictors are zero. Age is centered at 65 years and percent of predicted FEV$_1$ is centered at 40% to aid in interpreting fold changes.
†Odds ratio for remaining exacerbation-free at 6 months comparing those with versus without a one unit increase in the predictor shown, adjusted for other covariates in the model including treatment group, age, gender, percent of predicted FEV$_1$, smoking status and study site.
*Percentage increase in 6-month-RMST for each unit increase of the predictor, adjusted for other covariates in the model.

revealed by these analyses. The azithromycin effect is most consistent over time in patients with FEV$_1$ percent of predicted$>$50%, those aged 70 years or older and females. Investigators seeing these analyses would proceed by looking for potential explanations for waning treatment effects in some of the other subgroups and seeking additional data and/or analyses to support or further explain these results.

3.7 Discussion

This manuscript offers a suite of new methods for a modern analysis of recurrent events data subject to censoring. Data visualization via heatmaps of τ-restricted times-to-first-event are introduced that handle censored event times through a multiple imputation procedure that has very few parametric assumptions made during its implementation. By converting traditional recurrent event data into a censored longitudinal data structure, we are able to leverage longitudinal data analysis experience in the analysis of this data. The τ-IBR model is a one-stop shop for assessing overall susceptibility to recurrent events (model 3.5 without window start times as covariates), changes to susceptibility to recurrent events over time (Model 3.5 with window start times as covariates and interactions, as appropriate), the influence of predictors on time-to-first-exacerbation amongst those who experience it (model 3.6 with and without window start times as covariates and interactions, as appropriate) and patterns of mean time-to-first-recurrent event in follow-up windows over time via equations (3.9) and (3.10) that combine results from models 3.5 and 3.6. Data visualizations of model results shown in Figure 3.4 are an additional tool for making results interpretable to a medical research readership.

To our knowledge, a τ-inflated beta regression model framework has never been proposed as a way to model recurrent time-to-event data, censored or otherwise. The conversion process from a recurrent event time data structure to a series of τ-restricted times-to-first-event in (potentially overlapping) follow-up windows over time tends to produce a point mass at τ in each follow-up window. This data feature is often associated with patient predictor profiles that are less susceptible to the recurrent events and, as such, they must be addressed in the analysis. Simulations indicate that the τ-IBR model addresses this issue well compared to its nearest competitor model, the XMT model that ignores these point masses in its analyses. Both the XMT model and the τ-IBR model have the ability to model patterns of mean time-to-first-recurrent event in follow-up windows over time. While the XMT model has good coverage rates for true τ-RMST values when there are no predictors associated with the point mass of restricted event times at τ (simulation setting 2), these coverage rates fall apart when predictors related to susceptibility of recurrent events at τ exist (simulation setting 1), even when there are no censored event times. Figure 3.2 explains this phenomena by showing the high average absolute value of bias in this setting.

The added interpretations afforded by the τ-IBR model that correctly models $T_i(t)$ by breaking it down into its mixture components via equation (1) is an additional bonus to correcting this important coverage rate issue. As seen in the COPD example, the relative importance of risk factors in models (3.5) and (3.6) shifted between the two models, with most statistical signal for the treatment effect appearing in model (3.5). The 6-month-IBR model highlighted the interpretation that the azithromycin participants were significantly less susceptible to having exacerbations in any 6-month period, and that there was a marginally significant 5% increase in exacerbation-free time amongst those who had exacerbations during a 6-month follow-up period.

We developed both ES and MI algorithms for fitting and reporting results for the τ-IBR model. In simulation, both estimation methods performed reasonably well in terms of efficiency and bias of parameter estimates, although the MI slightly outperforms the ES approach in estimating the intercept term of model 3.5. Since the intercept term of this model affects estimation of τ-RMST values as well, the MI approach performed a bit better than the ES approach in terms of bias and coverage of the true τ-RMST values. Our MI algorithm does not incorporate a bootstrap step as some MI algorithms recommend [47, 19, 57], a step we elected to skip after noting good coverage of our approach. Those of a more pure Bayesian mindset will likely wish to include a bootstrap step as a matter of principle, but we found the gain in obtaining results more quickly with good coverage rates quite satisfying and continue to recommend skipping the bootstrap step. This is, of course, an easily incorporated change for those who wish to do so.

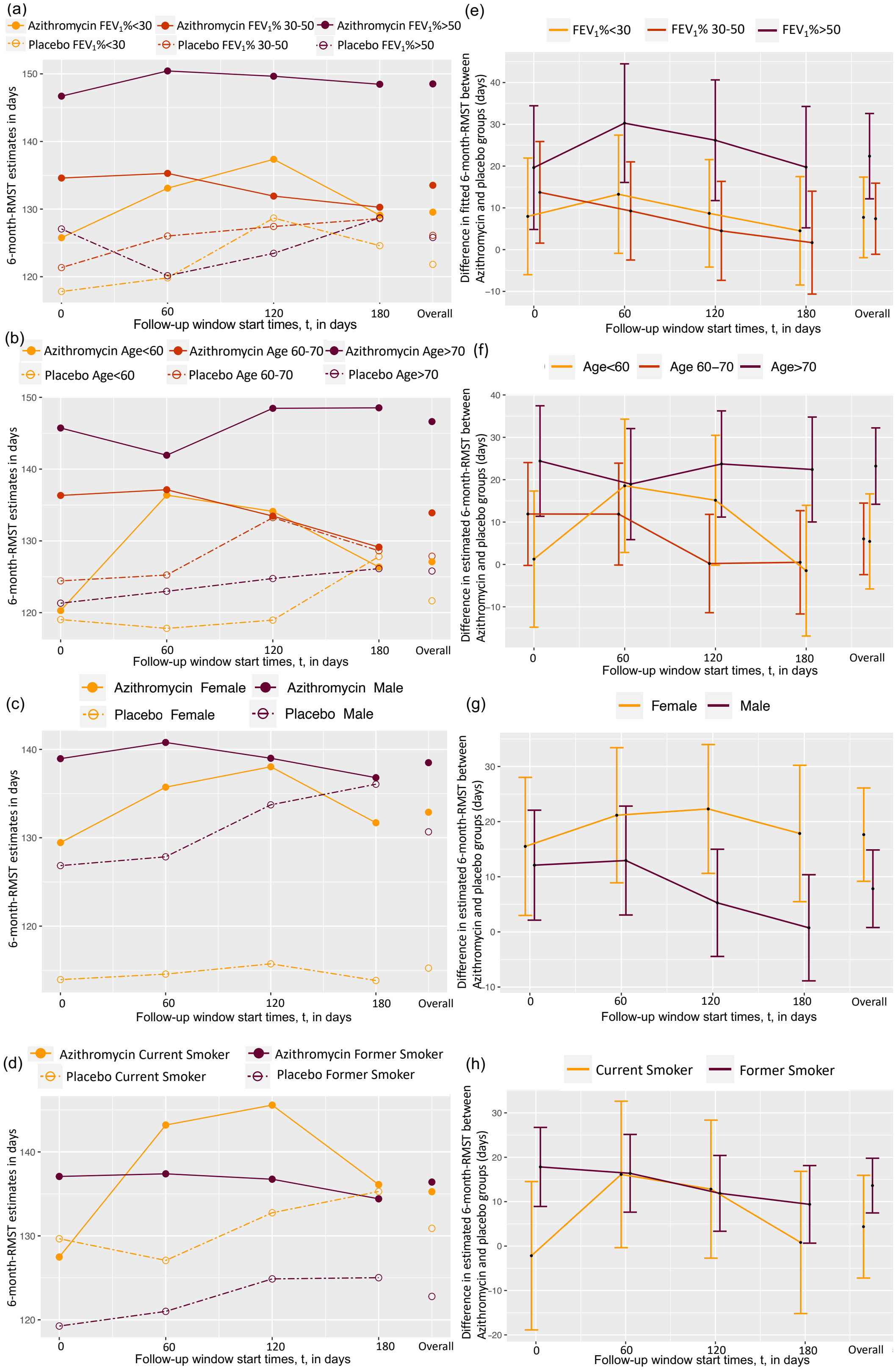

Figure 3.4: (a)-(d) Estimated 6-month-RMST by different combinations of treatment and other predictors via 6-month-IBR MI-fitted model. (e)-(h) Difference in estimated 6-month-RMST between Azithromycin and placebo groups for different levels of predictors.

CHAPTER 4

τ-Inflated Beta Regression Model for Lung Transplant Candidate Urgency Estimation Subject to Dependent Censoring

4.1 Introduction

Lung transplantation has the potential to enhance the lifespan and well-being of patients suffering from end-stage lung disease. However, due to the limited availability of organs, access to this therapy is severely restricted. In 2005, the Organ Procurement and Transplant Network (OPTN) altered the United States lung allocation policy for individuals aged 12 and above to take into account expected mortality projections for each lung transplant candidate in the two scenarios of receiving a transplant or remaining on the waitlist [9]. Each lung transplant candidate is assigned a Lung Allocation Score (LAS) based on these projections, with higher LAS scores leading to offers of deceased donor lung transplants. One important component of the LAS is a measure of transplant urgency that is based on a candidate's one-year restricted mean survival time (RMST) if they remain on the waitlist without transplant. This focus of this manuscript is (1) to describe potential waitlist mortality data biases and heretofore overlooked opportunities relevant to RMST transplant urgency estimation, (2) to provide an improved method for estimating these RMST values, and (3) to improve the ability to understand the relationship between factors used in the LAS and waitlist mortality. The remainder of the introduction will discuss potential waitlist mortality data biases that we plan to address and data structure opportunities that we plan to capitalize upon with our proposed modeling framework.

Because lung waitlist candidates with high LAS values are typically offered lung transplants, their one-year-restricted waitlist survival times are dependently censored at the time of transplant. Hence any method for estimating transplant urgency must address this dependent censoring issue. Inverse-probability-of-censoring weighting (IPCW) methods for time-to-event data are a popular approach, with many varieties of these algorithms appearing in the literature since first proposed by Robins, Rotnitzky and co-authors [43, 41, 42, 52, 51, 15]. In the context of RMST estimation with dependently censored times-to-event, IPCW approaches have been incorporated into modeling censored restricted event times using pseudo-observations (PO) [71, 60], multiple imputation (MI)[60, 72] and generalized estimating equations (GEE)[65]. In this manuscript, we will pursue an MI approach to address dependent censoring of event times.

By definition, although not widely acknowledged, τ-restricted times-to-event, min(T_i, τ), for individuals, $i = 1, \ldots, n$, are mixtures of a time-to-event of interest, T_i and a Bernoulli$[\pi_i = P(T_i \geq \tau)]$ random variable, B_i, via the relationship, min$(T_i, \tau) = \tau B_i + T_i(1 - B_i)$. To date, no existing IPCW RMST estimation approaches acknowledge the random variable, B_i, as having a role in RMST estimation and inference. Yet based on this representation of min(T_i, τ), an individual's τ-RMST becomes:

$$\text{E}[\text{min}(T_i, \tau)] = \tau\pi_i + \text{E}(T_i|T_i < \tau)(1 - \pi_i), \tag{4.1}$$

where the mean of B_i is prominently featured. In many practical applications, predictors associated with π_i and E$(T_i|T_i < \tau)$ will differ in either composition or relative importance. Hence an RMST estimation algorithm that incorporates results from separate models for π_i and E$(T_i|T_i < \tau)$ should offer opportunities for improvement in both inference and statistical efficiency. We vigorously pursue this line of thinking in our manuscript. For instance in our motivating lung waitlist setting, Chronic Obstructive Pulmonary Disease (COPD) lung candidates often desire a transplant to improve quality of life as opposed to length of life, and so these candidates are more likely to remain at risk on the waitlist beyond $\tau = 1$ year compared to other lung diagnoses. Alternatively, Interstitial Pulmonary Fibrosis (IPF) candidates typically enter the waitlist with a very low life expectancy. Hence IPF diagnosed participants are more likely to contribute information towards models of E$(T_i|T_i < 1$ year) than COPD participants. We therefore anticipate that 1-year RMST estimates for COPD participants will be more strongly influenced by the model for π_i versus the model for E$(T_i|T_i < 1$ year) with the reverse likely the case for IPF candidates.

In general, modeling the statistical quantities on the right hand side of equation (4.1) should better allow analysts to evaluate patient profiles that lend themselves to (a) a lower overall chance of dying during a restricted period and (b) a longer survival time amongst patients who died during a restricted period. And resulting RMST estimates based on the right hand side of equation (4.1) are likely to be more precise than estimates based on models that don't take into account the role of the point-mass at min$(T_i, \tau) = \tau$ via the relationship min$(T_i, \tau) = \tau B_i + T_i(1 - B_i)$. This intuition regarding efficiency has been confirmed in other settings where survival quantities were averaged across mutually exclusive and exhaustive groups of inherent interest in the censored time-to-event setting. [13, 34, 35, 32].

Another data opportunity that is often overlooked in τ-RMST estimation, and that we plan to address in our methods development, is the vast availability of statistical information beyond the initial τ years of follow-up in many practical settings. Tayob and Murray [58, 60] demonstrated considerable efficiency gains in τ-RMST estimation merely by incorporating information from several regularly spaced τ-length follow-up windows per individual and appropriately taking into account correlation between information contributed from the same individual in the analysis of the resulting censored longitudinal data structure. Our methods will build upon their suggested representation of censored times-to-event as a longitudinal data structure of censored τ-restricted outcomes observed throughout the follow-up period.

The remainder of the manuscript is structured as follows. Section 4.2 introduces notation and describes how to convert traditional censored time-to-event data into a regularly spaced censored longitudinal data structure suitable for τ-RMST analysis based on the decomposition of its elements given in equation (4.1). In Section 4.3, we introduce our proposed τ-IBR model for the censored longitudinal data described in section 4.2 in the special case with no censoring. An MI approach for addressing dependent censoring is developed in Section 4.4. Section 4.5 reports the results of simulation studies that assess the finite sample properties of our methodology. An analysis of the lung candidate data highlighting advantages of our approach is given in Section 4.6, followed by a discussion in Section 4.7.

4.2 Description of Random Variables and Construction of Censored Longitudinal Data

In this section, we define notation, provide an overview of how to transform traditional time-to-event data into a censored longitudinal data structure, and introduce additional longitudinal data notation that is relevant to the goals of this manuscript.

Let T_i and C_i represent the time-to-event and censoring time, respectively, for patient i, where $i = 1, 2, \dots, n$. As the data is censored, we can only observe $X_i = \min(T_i, C_i)$, along with the corresponding censoring indicator $\Delta_i = I(T_i \leq C_i)$ for $i = 1, \dots, n$. The vectors of covariates related to the event time T_i and censoring time C_i at a specific time point t are represented by $Z_i(t)$ and $V_i(t)$, respectively. The covariate histories up to time t are indicated by $\bar{Z}_i(t) = \{Z_i(u); 0 \leq u \leq t\}$ and $\bar{V}_i(t) = \{V_i(u); 0 \leq u \leq t\}$. We define the counting process for the event of interest as $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, where $dN_i(t) = I(X_i = t, \Delta_i = 1)$, and the counting process for censoring time as $N_{c_i}(t) = I(X_i \leq t, \Delta_i = 0)$, where $dN_{c_i}(t) = I(X_i = t, \Delta_i = 0)$. In addition, we define $Y_i(t) = I(X_i \geq t)$ as the indicator of individual i being at risk for the event at time t.

In this manuscript, we construct a censored longitudinal data structure for the time-to-event data similar to that described in Tayob and Murray [58, 59, 60]. This structure transforms traditional time-to-event data into potentially censored time-to-event in follow-up windows of length τ measured from follow-up times $t \in t_1, ..., t_b$, where $t_1 = 0$ and $t_k = t_{k-1} + a$ for $k = 2, \ldots, b$. The length of the follow-up window, τ, is chosen to be clinically meaningful for the patient population and research question of interest. For instance, the lung transplant candidate urgency estimate described in the introduction is based on an estimate of 1-year restricted lifetime without opportunity for transplant, so that τ = 1 year is appropriate in this case. With the structure of follow-up windows defined in terms of a, b, and τ, we now define notation for τ-restricted time-to-event in each of the follow-up windows mapped from traditional event time random variables.

For each $t \in t_1, \ldots, t_b$, we define:

$$T_i(t) = \text{ min}[T_i - t, \tau]$$

with observed data in the presence of censoring:

$$X_i(t) = \text{ min}[X_i - t, \tau]$$
$$\Delta_i(t) = I[X_i(t) < C_i - t]$$

where $T_i(t)$ is the corresponding τ-restricted time-to-event measured from window start time t, with $X_i(t)$ and $\Delta_i(t)$ being the corresponding τ-restricted time-to-observed-event and censoring indicator, respectively, measured from t. Any individual i who is not at risk at the beginning of a follow-up window starting at t is assumed to have $X_i(t) = \Delta_i(t) = 0$.

To illustrate the data structure transformation and introduce the longitudinal outcome notation used in our modeling framework, we present two examples. Suppose patient 1 died 16 months after listing, and patient 2 was censored 10 months after listing due to receiving a transplant. Using traditional time-to-event notation, the observed data for patient 1 becomes $\{X_1 = T_2 = 16$ months, $\Delta_1 = 1\}$, and the observed data for patient 2 becomes $\{X_2 = C_2 = 10$ months, $\Delta_2 = 0\}$. To define the new longitudinal data structure shown in Figure 1, we set $\{\tau$, a, $b\}$ = {12, 6, 3} months, giving 3 12-month follow-up windows starting at $\{t_1, t_2, t_3\}$={0, 6, 12} months. Hence, patient 1 contributes the following longitudinal outcome data related to the time-to-event of interest:

$$\{T_1(0) = X_1(0) = 12 \text{ months}, \Delta_1(0) = 1, Z_1(0)\}$$
$$\{T_1(6) = X_1(6) = 10 \text{ months}, \Delta_1(6) = 1, Z_1(6)\},$$
$$\{T_1(12) = X_1(12) = 4 \text{ months}, \Delta_1(12) = 1, Z_1(12)\},$$

and patient 2 contributes:

$$\{X_2(0) = 10 \text{ months}, \Delta_1(0) = 0, Z_2(0)\}$$
$$\{X_2(6) = 4 \text{ months}, \Delta_2(6) = 0, Z_2(6)\}.$$

Note that $X_1(0) = 12$ months means that the time-to-event within the first window attains τ for patient 1. In the lung transplant candidate data described in the introduction, approximately 1679 individuals at time t = 0 have $X_i(0) = 12$ months, leading to a point mass of follow-up windows achieving this same value of $X_i(0)$.

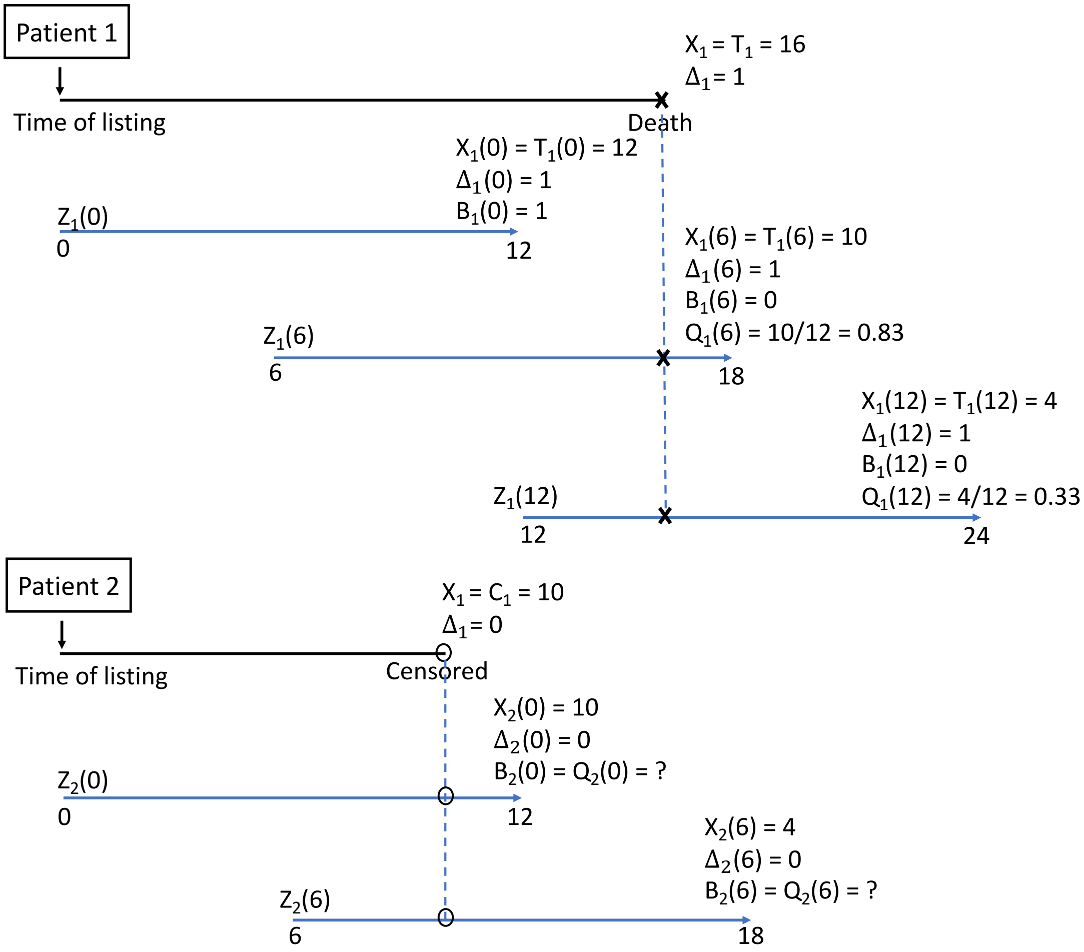

Figure 4.1: Two examples of how to construct censored longitudinal data from an individual's traditional event-time data.

This manuscript leverages information inherent in the data about this point mass for the ultimate goal of better understanding properties of $T_i(t)$ via the following relationship:

$$T_i(t) = \tau B_i(t) + T_i(t)[1 - B_i(t)] = \tau\big\{B_i(t) + Q_i(t)[1 - B_i(t)]\big\}, \tag{4.2}$$

where $B_i(t) = I[T_i(t) = \tau]$ is a Bernoulli random variable with mean $\pi_i(t) = \Pr[T_i(t) = \tau]$ and $Q_i(t) = \tau^{-1}T_i(t)$ defined conditionally for $T_i(t) < \tau$ (i.e., $B_i(t) = 0$), is a continuous random variable on the sample space between zero and one that is statistically independent of $B_i(t)$. We may model $B_i(t), t \in t_1, \ldots, t_b$ and $Q_i(t), t \in t_1, \ldots, t_b$ to obtain inferences on $T_i(t)$.

Adding these additional random variables to our example patients in Figure 4.1, patient 1 (who died at 16 months post listing) contributes longitudinal outcomes $\{[B_1(0) = 1], [B_1(6) = 0, Q_1(6) = 0.83], [B_1(12) = 0, Q_1(12) = 0.33]$. Patient 2 (who was censored due to receiving a transplant at 10 months post listing) has missing data for $B_2(0), B_2(6), Q_2(0)$ and $Q_2(6)$ due to the patient being at risk at both $t = 0$ and 6 months post listing and censored within each of these follow-up windows. In Section 4, we describe a inverse probability weighted multiple imputation approach that takes into account the partially observed at-risk data in each of these follow-up windows, as well as covariate risk profiles of individuals in the dataset, to impute missing $B_i(t)$ and $Q_i(t)$ outcomes that will be used in our models in the presence of potentially dependent censoring.

The following two sections of the manuscript develop methodology first in the uncensored data setting (section 4.3) and then in the censored data setting (section 4.4), where our multiple imputation approach is introduced for fitting the model.

4.3 τ-Inflated Beta Regression Model Specification in the Special Case with No Censoring

One of the primary goals of this manuscript is to improve estimation of τ-restricted mean event times in the presence of dependent censoring, since that is an essential element of U.S. Lung Allocation Score urgency estimate. In this section, we establish some important preliminary model definitions and methods for estimation and inference in the no censoring case that will be useful when describing multiple imputation methods for the case with dependent censoring in section 4.4.

Prior τ-restricted mean models that address potential dependent censoring have either focused solely in the follow-up window starting at $t = 0$ [71], or in the case of Tayob and Murray [60], modeled the restricted mean event-time of interest over follow-up windows starting at $t \in \{t_1, ..., t_b\}$ assuming a mean structure for $\log[T_i(t)]$ or $T_i(t)$ such as:

$$\mathrm{E}[T_i(t)|Z_i(t)] = \tilde{\beta}^T Z_i(t), \tag{4.3}$$

$i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, where $\tilde{\beta}$ is a vector of parameters corresponding to a vector of predictors, $Z_i(t)$. Although Tayob and Murray developed their approach with $\log[T_i(t)]$ as the focus of their model and estimation methods, for the purposes of this manuscript we apply their approach (with minor modifications) to the original scale of $T_i(t)$ as in model (4.3) so that results from our proposed method will be more directly comparable to theirs. Hereafter, we will refer to model (4.3) as the Tayob and Murray τ-Restricted Mean Survival (TM τ-RMST) model.

As mentioned in section 4.2, we view $T_i(t)$ through the lens of the random variables $B_i(t)$ and $Q_i(t)$ via the algebraic relationship expressed in equation (4.2). Based on equation (1), the following algebraic relationship for the mean of $T_i(t)$ must also hold:

$$\mathrm{E}[T_i(t)] = \mathrm{E}\big\{\tau\big\{B_i(t) + Q_i(t)[1 - B_i(t)]\big\}\big\} = \tau\big\{\pi_i(t) + \mu_i(t)[(1 - \pi_i(t)]\big\}, \tag{4.4}$$

where $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$, $\pi_i(t) = \mathrm{E}[B_i(t)]$ and $\mu_i(t) = \mathrm{E}[Q_i(t)]$. It makes sense to estimate transplant urgency based on the right hand side of equation (4.4) if we suspect that there are different associations between covariates and $B_i(t)$ versus $Q_i(t)$. Multivariable models for $\pi_i(t)$ will more clearly identify characteristics associated with surviving for the full τ-duration follow-up period. In evaluating lung transplant candidates the $\pi_i(t)$ model will identify predictors associated with low risk of mortality during the next τ duration follow-up period if a donor lung is not made available for transplant. Multivariable models for $\mu_i(t)$ will more clearly identify predictors associated with higher and lower transplant urgency amongst those who are not anticipated to survive the subsequent τ-duration follow-up period. To reflect different potential covariate associations at work for outcomes $B_i(t)$ and $Q_i(t)$, we introduce notation for the relevant subsets of $Z_i(t)$ that pertain to these models, $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$, respectively.

Our proposed τ-inflated beta regression model applied to the censored longitudinal dataset laid out in section 4.2 uses generalized estimation equation (GEE) methods to fit models for $\pi_i(t)$ and $\mu_i(t)$ based on data $\{B_i(t), Z_{\pi i}(t), i = 1, \ldots, n, t \in (t_1, ..., t_b)\}$ and $\{Q_i(t), Z_{\mu i}(t), i = 1, \ldots, n, t \in (t_1, ..., t_b)\}$, respectively. Inspired by equation (4.4), the underlying assumption of the τ-IBR approach is that patient i's τ-RMST for the follow-up window starting at time t satisfies

$$\mathrm{E}[T_i(t)|Z_i(t)] = \tau\big\{\mathrm{E}[B_i(t)|Z_{\pi i}(t)] + E[Q_i(t)|Z_{\mu i}(t)]\{(1 - \mathrm{E}[B_i(t)|Z_{\pi i}(t)]\}\big\}. \tag{4.5}$$

For simplicity, we will continue to use the notation $\pi_i(t)$ for $\mathrm{E}[B_i(t)|Z_{\pi i}(t)]$ and the notation $\mu_i(t)$ for $\mathrm{E}[Q_i(t)|Z_{\mu i}(t)]$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$. In vector form these become $\pi_i = [\pi_i(t_1), \ldots, \pi_i(t_b)]^T$ and $\mu_i = [\mu_i(t_1), \ldots, \mu_i(t_b)]^T$, $i = 1, \ldots, n$. As in standard longitudinal analysis, $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$ can include window start times t, time-dependent covariates that change at the window start times, and interactions between t and other covariates.

For each model, GEE methodology requires specification of: (1) the mean structure being modeled, (2) the assumed variance function, and (3) the assumed working correlation matrix for outcomes taken from the same individual.

We assume that the mean structure for $B_i(t)$ given $Z_{\pi i}(t), i = 1, \ldots, n, t \in (t_1, ..., t_b)$, follows the model:

$$g[\pi_i(t)] = \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = \beta_0 + \beta_1^T Z_{\pi i}(t). \tag{4.6}$$

Later we express $\pi_i(t)$ in terms of $\beta = (\beta_0, \beta_1^T)^T$ and $Z_i^\pi(t) = [1, Z_{\pi i}^T(t)]^T$, so that $\pi_i(t) = 1/\left[1 + e^{-\beta^T Z_i^\pi(t)}\right]$. The corresponding variance function for $B_i(t)$ given $Z_{\pi i}(t)$ is taken from the Bernoulli[$\pi_i(t)$] distribution, i.e., Var$[B_i(t)|Z_{\pi i}(t)] = \pi_i(t)[1-\pi_i(t)], i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$. Suppose that $R_{\pi i}$ is the working correlation matrix for $B_i = [B_i(t_1), \ldots, B_i(t_b)]^T$, and that $A_{\pi i}$ is a diagonal matrix with Var$[B_i(t)] = \pi_i(t)[1-\pi_i(t)]$, $t = t_1, \ldots, t_b$, along the diagonal so that the covariance matrix for B_i becomes $V_{\pi_i} = A_{\pi i}^{\frac{1}{2}} R_{\pi i} A_{\pi i}^{\frac{1}{2}}$, $i = 1, \ldots, n$.

According to Liang and Zeger's GEE methodology [76], we may estimate parameters, β, in model (4.6) by solving the estimating equation,

$$\sum_{i=1}^{n} \frac{\partial \pi_i}{\partial \beta}^T V_{\pi_i}^{-1}(B_i - \pi_i) = 0, \tag{4.7}$$

where $(\partial\pi_i/\partial\beta)^T = \{[\partial\pi_i(t_1)/\partial\beta]^T, \ldots, [\partial\pi_i(t_b)/\partial\beta]^T\}$, $i = 1, \ldots, n$, with components $[\partial\pi_i(t)/\partial\beta]^T = Z_i^\pi(t)e^{\beta^T Z_i^\pi(t)}/\left[1 + e^{\beta^T Z_i^\pi(t)}\right]^2$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$.

We assume that the mean structure for $Q_i(t)$ given $Z_{\mu i}(t), i = 1, \ldots, n, t \in (t_1, ..., t_b)$, follows the model:

$$g[\mu_i(t)] = \log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = \alpha_0 + \alpha_1^T Z_{\mu i}(t). \tag{4.8}$$

Later we express $\mu_i(t)$ in terms of $\alpha = (\alpha_0, \alpha_1^T)^T$ and $Z_i^\mu(t) = [1, Z_{\mu i}^T(t)]^T$, so that $\mu_i(t) = 1/\left[1 + e^{-\alpha^T Z_i^\mu(t)}\right]$. The corresponding variance function for $Q_i(t)$ given $Z_{\mu i}(t)$ is taken from the beta$\{\mu_i(t)\nu, [1-\mu_i(t)]\nu\}$ distribution, i.e., Var$[Q_i(t)|Z_{\mu i}(t)] = (\nu+1)^{-1}\mu_i(t)[1-\mu_i(t)]$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$. Suppose that $R_{\mu i}$ is the working correlation matrix for $Q_i = [Q_i(t_1), \ldots, Q_i(t_b)]^T$, and that $A_{\mu i}$ is a diagonal matrix with Var$[Q_i(t)] = (\nu+1)^{-1}\mu_i(t)[1-\mu_i(t)]$, $t = t_1, \ldots, t_b$, along the diagonal so that the covariance matrix for Q_i becomes $V_{\mu_i} = A_{\pi i}^{\frac{1}{2}} R_{\pi i} A_{\pi i}^{\frac{1}{2}}$, $i = 1, \ldots, n$. Let $U_{\mu_i} = \text{diag}\{[1-B_i(t_1)], ..., [1-B_i(t_b)]\}$, $i = 1, \ldots, n$. The estimating equation for α in model (4.8) can be written as:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \alpha}^T V_{\mu_i}^{-1} U_{\mu_i}(Q_i - \mu_i) = 0, \tag{4.9}$$

where $(\partial\mu_i/\partial\alpha)^T = \{[\partial\mu_i(t_1)/\partial\alpha]^T, \ldots, [\partial\mu_i(t_b)/\partial\alpha]^T\}$ with components $[\partial\mu_i(t)/\partial\alpha]^T = Z_i^\mu(t)e^{\alpha^T Z_i^\mu(t)}/\left[1 + e^{\alpha^T Z_i^\mu(t)}\right]^2$, $i = 1, \ldots, n, t \in \{t_1, ..., t_b\}$. To interpret this model, consider the k^{th} element of $Z_{\mu i}(t)$, $Z_{\mu ik}(t)$, with corresponding parameter α_{1k}. For a one unit increase of $Z_{\mu ik}(t)$ from z to z+1, the fold change for $\mu_i(t)$ becomes $e^{\alpha_{1k}}(1+e^{\alpha_0+z\alpha_{1k}})/(1+e^{\alpha_0+\alpha_{1k}+z\alpha_{1k}})$ assuming that all other predictors in the model are zero. When interpreting the model in subject area manuscripts, we center continuous predictors and use zero values for reference groups of categorical predictors.

We consider unstructured working correlation matrix to easily handle the correlation between overlapping and non-overlapping follow-up windows. Solutions to equations (4.7) and (4.9) are obtained using the geem2 function from the mmmgee package. This yields estimated values for $\hat{\alpha}$ and $\hat{\beta}$, as well as robust sandwich estimates for the corresponding parameter covariance matrices, $\hat{V}_\alpha$ and $\hat{V}_\beta$, respectively. As long as the mean functions are correctly specified, $\hat{\alpha}$ and $\hat{\beta}$ are consistent and asymptotically normal[75, 76].

Based on estimated parameters and equation (4.4), we may estimate τ-RMST for each individual and follow-up window, $\mathrm{E}[T_i(t)|Z_i]$, $i = 1,\ldots,n$, $t = t_0,\ldots,t_b$. Defining $\hat{\mu}_i(t) = 1/\big[1+e^{-\hat{\alpha}^T Z_i^\mu(t)}\big]$ and $\hat{\pi}_i(t) = 1/\big[1+e^{-\hat{\beta}^T Z_i^\pi(t)}\big]$, the estimated $\mathrm{E}[T_i(t)]$ for subject i and window t becomes

$$\hat{\mathrm{E}}[T_i(t)] = \tau\hat{\mu}_i(t)[1-\hat{\pi}_i(t)] + \tau\hat{\pi}_i(t). \tag{4.10}$$

After some algebraic manipulation,

$$\begin{aligned}\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[T_i(t)]\} &= \tau^2\left[1-\frac{1}{1+e^{-\hat{\alpha}^T Z_i^\mu(t)}}\right]^2 Z_i^\pi(t)^T\hat{V}_\beta Z_i^\pi(t)\frac{\big[e^{-\hat{\beta}^T Z_i^\pi(t)}\big]^2}{\big[1+e^{-\hat{\beta}^T Z_i^\pi(t)}\big]^4}\\ &+\tau^2\left[1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^\pi(t)}}\right]^2 Z_i^\mu(t)^T\hat{V}_\alpha Z_i^\mu(t)\frac{\big[e^{-\hat{\alpha}^T Z_i^\mu(t)}\big]^2}{\big[1+e^{-\hat{\alpha}^T Z_i^\mu(t)}\big]^4}.\end{aligned} \tag{4.11}$$

4.4 Multiple Imputation Algorithm For Dependently censored data

In this section, we describe how to multiply impute dependently censored outcomes in the context of fitting our proposed τ-inflated beta regression model. For each censored individual, the general idea of our imputation procedure is to (1) construct a risk set of individuals with similar covariate histories to the censored individual, where this risk set takes into account similarity in predicted $\mu_i(t)$ and $\pi_i(t)$ values over time, and then (2) use an inverse transform (IT) imputation algorithm to multiply impute outcomes for the censored individual, a technique that has been developed and modified by many authors[30, 57, 20, 21, 70, 60].

In our setting a censored patient i requires imputation of $T_i(t)$ in the set, S_i, of τ-length follow-up windows starting at $\{t \in \{t_1, \ldots, t_b\} : 0 < X_i(t) < \tau, \Delta_i(t) = 0\}$. If S_i contains more than one window, we only need to impute the time-to-event for the last window in the set, denoted as b_i^{th} window starting at t_i^{sup}. Once we have imputed the τ-restricted time-to-event, denoted as $\tilde{T}_i(t_i^{sup})$ for the last window starting at t_i^{sup}, we can impute the time-to-event for other windows in S_i starting at $t < t_i^{sup}$ as $\tilde{T}_i(t) = \min[\tilde{T}_i(t_i^{sup}) + t_i^{sup} - t, \tau]$. Using the imputed $\tilde{T}_i(t)$ for each window in S_i, we can impute the outcomes of models (4.6) and (4.8). Specifically, we impute $\tilde{B}_i(t)$ as $\mathrm{I}[\tilde{T}_i(t) = \tau]$, and $\tilde{Q}_i(t)$ as $\tilde{T}_i(t)$ if $\tilde{B}_i(t) = 0$.

An outline of the steps needed to generate our imputed datasets follows:

1. Fit models (4.6) and (4.8) using data from follow-up windows that don't require imputation to obtain initial model parameter estimates, $\hat{\theta}^{(0)} = (\hat{\alpha}^{(0)}, \hat{\beta}^{(0)})$.
2. Form risk set $\mathcal{R}_i$ for each subject i requiring imputation by including candidates with similar histories of covariates to patient i based on initial estimates of model parameters $\hat{\theta}^{(0)}$. If there are particularly important covariates related to either the outcome of interest and/or the censoring mechanism over time, these may be used as additional criteria for entering the risk set for patient i. In the lung waitlist setting, time-dependent LAS would be appropriate to single out as such a covariate.
3. Impute $B_i(t)$ and $Q_i(t)$ for each window in S_i.
   (a) Compute the IPCW (Inverse Probability of Censoring Weighted) survival estimate within the risk set $\mathcal{R}_i$, denoted as $\hat{S}^W_{R_i}$.
   (b) Sample a valid impute using the IPTI approach.
4. Repeat step 3 (b) M times, resulting in M imputed data sets to be analyzed, and then combine the results from M imputed datasets using method proposed by Rubin[46, 45] to get final estimates of θ and its corresponding covariance matrix.

We now provide details of steps 2-4 of the algorithm.

Step 2: (Risk set definition step) To impute $\tilde{T}_i(t_i^{sup})$ in the sup window for each individual $i = 1, \ldots, N_c$ requiring imputation, we define a risk set $\mathcal{R}_i$ consisting of individuals who are similar to i. To be included in $\mathcal{R}_i$, a candidate $j = 1, \ldots, N_i$ must satisfy two constraints: (a) $\max[|\hat{\mu}_i(t) - \hat{\mu}_j(t)|, |\hat{\pi}_i(t) - \hat{\pi}_j(t)|] < \epsilon$ for $t = 0, \ldots, t_i^{sup}$, where $\hat{\mu}_i(t)$ and $\hat{\pi}_i(t)$ are estimated from $\hat{\theta}^{(0)}$ and (b) $X_j(t_i^{sup}) > X_i(t_i^{sup})$. Condition (b) in this step ensures that all subjects in $\mathcal{R}_i$ are at risk when individual i is censored. Condition (a) requires that individuals in R_i have similar predicted model outcomes to the individual i at each window start time up to t_i^{sup}. Suppose that $Z_i^*(t)$ is a particularly important covariate related to the outcome of interest and/or the dependent censoring mechanism. Then criteria (a) may be modified to require that $\max[|\hat{\mu}_i(t) - \hat{\mu}_j(t)|, |\hat{\pi}_i(t) - \hat{\pi}_j(t)|, |Z_i^*(t) - Z_j^*(t)|] < \epsilon$ for $t = 0, \ldots, t_i^{sup}$. We set $\epsilon = 0.05$ in condition (a) and increase ϵ by 0.005 until either $N_i \geq 10$ or $\epsilon > 0.5$.

Step 3 (a): (IPCW survival estimation) The IPTI imputation approach requires a consistent marginal survival curve estimate to be calculated for each risk set, $\mathcal{R}_i$, which in the dependent censoring setting can be obtained via the IPCW survival estimate[41]. IPCW related survival, hazard and cumulative hazard estimates in this step are defined from the start of the first follow-up window, so that the time argument, u, should be interpreted on that time scale in this step. Inverse weights to adjust for dependent censoring are estimated via a Cox model for the dependent censoring time of the form:

$$\lambda_C[u|\bar{V}_i(u)] = \lambda_C^0(u)\text{exp}[\gamma^T V_i(u)],$$

where $\lambda_C[u|\bar{V}_i(u)]$ is the hazard function for the censoring distribution conditional on the history of covariates $\bar{V}_i(u)$, $\lambda_C^0(u)$ is an baseline hazard function for the censoring distribution, and γ is a vector of parameters corresponding to a vector of predictors, $V_i(u)$ available at time u. For each individual $i = 1, \ldots, n$, we calculate

$$\hat{W}_i(u) = \hat{P}^{-1}[C_i > u|\bar{V}_i(u)] = \text{exp}\Big[\sum_{k=1}^{n} \int_0^u \frac{e^{\hat{\gamma}^T V_i(v)} dN_{C_k}(v)}{\sum_{k'=1}^{n} Y_{k'}(v) e^{\hat{\gamma}^T V_{k'}(v)}}\Big].$$

We then define the estimated IPCW for each individual j within the risk set $\mathcal{R}_i$ as:

$$\hat{W}_j^{\mathcal{R}_i}(u) = \hat{P}^{-1}[C_j > u|C_j > C_i, \bar{V}_j(u)] = \frac{\hat{P}[C_j > C_i|\bar{V}_j(u)]}{\hat{P}[C_j > u|\bar{V}_j(u)]} = \frac{\hat{W}_j(u)}{\hat{W}_j(C_i)}.$$

Based on $\hat{W}_j^{\mathcal{R}_i}(u)$, the inverse weighted cumulative hazard within the risk set $\hat{\Lambda}_{\mathcal{R}_i}^W(u)$ can be calculated using:

$$\hat{\Lambda}_{\mathcal{R}_i}^W(u) = \sum_{j \in \mathcal{R}_i} \int_{C_i}^u \frac{dN_j(v) \hat{W}_j^{\mathcal{R}_i}(v)}{\sum_{j' \in \mathcal{R}_i} Y_{j'}(v) \hat{W}_{j'}^{\mathcal{R}_i}(v)},$$

so that the corresponding inverse weighted survival function becomes $\hat{S}_{\mathcal{R}_i}^W(u) = \text{exp}[-\hat{\Lambda}_{\mathcal{R}_i}^W(u)]$. For the IPTI approach described in step 3 (b), it is convenient to define $\hat{S}_{\mathcal{R}_i}^W(u) = 0$ for $u \geq \tau + t_i^{sup}$.

Step 3 (b): (IPTI approach) In this step, we define the IPTI approach for imputing $B_i(t_i^{sup})$ and $Q_i(t_i^{sup})$ for each censored individual i that requires imputation. First, we generate a uniform(0,1) random variable, $U_i = u_i$, and find the smallest observed event time t^* where $\hat{S}_{\mathcal{R}_i}^W(t^*) \leq u_i$. If $t^* \geq t_i^{sup} + \tau$, then impute 1 for $B_i(t_i^{sup})$; no further imputation for $Q_i(t_i^{sup})$ is required. Otherwise, impute 0 for $B_i(t_i^{sup})$ and $t^* - t_i^{sup}$ for $Q_i(t_i^{sup})$. Imputes for the remaining censored outcomes in S_i are then determined from the imputes for $B_i(t_i^{sup})$ and $Q_i(t_i^{sup})$, as appropriate. Completing this step results in a fully imputed dataset.

Step 4: Repeat step 3 (b) M times, resulting in M imputed data sets to be analyzed. By fitting model (4.6) and model (4.8) using each imputed data set, we obtain M parameter estimates $\hat{\theta}_m^{MI}$ with corresponding estimated covariance matrix $\widehat{\text{Var}}(\hat{\theta}_m^{MI})$, $m = 1, \ldots, M$. We combine the results from $M = 10$ imputed data sets using the approach proposed bu Rubin [45, 46]. The final estimate of θ based on M parameter estimates becomes $\hat{\theta}^{MI} = \sum_{m=1}^{M} \hat{\theta}_m^{MI}/M$ with corresponding estimated covariance matrix $\widehat{\text{Var}}(\hat{\theta}^{MI}) = \bar{U} + (1 + M^{-1})\bar{B}$, where $\bar{U} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\text{Var}}(\hat{\theta}_m^{MI})$ and $\bar{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m^{MI} - \hat{\theta}^{MI})(\hat{\theta}_m^{MI} - \hat{\theta}^{MI})^T$. The terms, $\hat{V}_\alpha^{MI}$ and $\hat{V}_\beta^{MI}$, can be extracted from $\widehat{\text{Var}}(\hat{\theta}^{MI})$, as appropriate.

4.5 Simulation Study

In this section we evaluate the performance of our proposed τ-IBR model for dependently censored data in terms of : (i) the quality of parameter estimates obtained from models (4.6) and (4.8) and (ii) the quality of longitudinal τ-RMST estimates in follow-up windows starting at $t = t_1 \ldots, t_b$ for individuals $i = 1, \ldots, n$ using our modeling framework versus the TM model approach. The data generation details for each of the 500 iterations simulated per scenario are described in Section 4.5.1 and the results of our simulations are summarized in Section 4.5.2.

The quality of parameter estimates for models (4.6) and (4.8) is assessed using several metrics, including bias, the average estimated standard error (ASE), the empirical standard deviation of the parameter estimates (ESD), and the average coverage probability of model-based 95% confidence intervals (CP). To evaluate the quality of longitudinal τ-RMST estimates, we consider bias, absolute value of bias (Abs-bias), ASE, and CP for n = 200, 500, and 700 subjects. Since estimated τ-RMST values are unique to each follow-up window contributed by individuals $i = 1, \ldots, n$, summary statistics for the performance of longitudinal τ-RMST estimates across the $j = 1, \ldots, 500$ simulations are defined as

$$\text{Bias} = \sum_{j=1}^{500}\sum_{i=1}^{n}\sum_{t=t_1}^{t_b} \frac{\hat{\text{E}}[T_{ij}(t)] - \text{E}[T_{ij}(t)]}{n \times b \times 500},$$

$$\text{Abs-bias} = \sum_{j=1}^{500}\sum_{i=1}^{n}\sum_{t=t_1}^{t_b} \frac{|\hat{\text{E}}[T_{ij}(t)] - \text{E}[T_{ij}(t)]|}{n \times b \times 500},$$

$$\text{ASE} = \sum_{j=1}^{500}\sum_{i=1}^{n}\sum_{t=t_1}^{t_b} \frac{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}{n \times b \times 500} \text{ and}$$

$$\text{CP} = \sum_{j=1}^{500}\sum_{i=1}^{n}\sum_{t=t_1}^{t_b} \frac{I(\text{Lower}_{ij} < \hat{\text{E}}[T_{ij}(t)] < \text{Upper}_{ij})}{n \times b \times 500},$$

where $\text{Lower}_{ij} = \hat{\text{E}}[T_{ij}(t)] - 1.96 \times \sqrt{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}$ and $\text{Upper}_{ij} = \hat{\text{E}}[T_{ij}(t)] + 1.96 \times \sqrt{\widehat{\text{Var}}\{\hat{\text{E}}[T_{ij}(t)]\}}$.

4.5.1 Data Generation

The simulated censored longitudinal data structure assumes $b = 3$ follow-up windows initiated at $\{t_1, t_2, t_3\} = \{0, 6, 12\}$ months with $\tau = 6$ months. For each individual, $i = 1, ..., n$, and time, $t \in \{0, 6, 12\}$, a time-dependent covariate $Z_{1i}(t)$ is independently generated from a uniform(0,1) distribution. An additional time-independent covariate, Z_{2i}, is generated from a Bernoulli(0.5) distribution. We consider two settings for the true τ-IBR model:

$$\text{Setting 1: } \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = -1.5 + 1.0Z_{1i}(t) + 2Z_{2i},$$
$$\log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = -1.0 + 1.0Z_{1i}(t) + 1.5Z_{2i};$$

$$\text{Setting 2: } \log\left[\frac{\pi_i(t)}{1-\pi_i(t)}\right] = 0.5,\ \log\left[\frac{\mu_i(t)}{1-\mu_i(t)}\right] = -1.0 + 1.0Z_{1i}(t) + 1.5Z_{2i}.$$

The model for $\pi_i(t)$ varies between the two settings. In Setting 1, the covariates $Z_{\pi_i} = \{Z_{1i}(t), Z_{2i}\}$ are considered important, whereas in Setting 2, there are no important covariates included in the model for $\pi_i(t)$. In terms of RMST (transplant urgency) estimation, our approach is not expected to outperform the TM model in Setting 2 since there are no interesting covariate features related to the point mass of τ-restricted event times at τ. However, in Scenario 1, our approach is expected to demonstrate advantages.

Outcomes for each subject $i, i = 1, \ldots, n$ are simulated to follow the assumed relationships for $\pi_i(t)$ and $\mu_i(t)$, $t \in \{0, 6, 12\}$ laid out in Settings 1 and 2. First, correlated Bernoulli$\{\pi_i(t)\}$ random variables for $t \in \{0, 6, 12\}$ are generated for each individual, giving us $\{B_i(0), B_i(6), B_i(12)\}, i = 1, \ldots, n$. These correlated Bernoulli outcomes are generated using an algorithm proposed by Emrich and Piedmonte [10] that is implemented using the function rmvbin from the R package bindata. If for individual, i, any follow-up windows are observed to have $B_i(t) = 0$ for $t \in \{0, 6, 12\}$, then we index the first of these follow-up windows with $t_i^{sup} = \min\{t : B_i(t) = 0\}$. Outcomes $B_i(t) : t > t_i^{sup}$ are removed from individual i's censored longitudinal data. For the follow-up window starting at t_i^{sup}, an outcome, $Q_i(t_i^{sup})$, is added to the censored longitudinal dataset for individual i that is independently generated from a Beta$[\mu_i(t_i^{sup})\nu, (1 - \mu_i(t_i^{sup}))\nu]$ distribution with $\nu = 3$.

We assess model performance for Settings 1 and 2 in the case of 25% dependent censoring before $t_b + \tau$, and compare it to a benchmark without censoring. For the 25% censoring case, we assume that the censoring time, C_i, has a piecewise exponential distribution that depends on covariates, $V_i(t) = \{Z_{1i}(t), Z_{2i}\}$. In particular, censoring hazards follow $\lambda_{C_i} = \lambda_C^0(u)\exp\{[0.4Z_{1i}(0) + 0.35Z_{2i}] \times I(0 \leq u < 6) + [0.4Z_{1i}(6) + 0.35Z_{2i}] \times I(6 \leq u < 12) + [0.4Z_{1i}(12) + 0.35Z_{2i}] \times I(u \geq$

Table 4.1: Finite sample performance of τ-IBR parameter estimates from models (4.6) and (4.8) for n=500 subjects with correlated longitudinal outcomes ($\rho = 0.2$) based on 500 iterations.

	No Censoring				**25% Censoring**			
	Bias	**ASE**	**ESD**	**CP**	**Bias**	**ASE**	**ESD**	**CP**
Setting 1								
α_0 = -1.0	0.002	0.119	0.124	0.944	0.004	0.126	0.132	0.946
α_1 = 1.0	0.001	0.197	0.203	0.944	-0.004	0.212	0.219	0.938
α_2 = 1.5	0.003	0.119	0.119	0.954	0.000	0.132	0.134	0.938
β_0 = -1.5	-0.045	0.190	0.200	0.944	-0.036	0.212	0.213	0.944
β_1 = 1.0	-0.010	0.252	0.278	0.924	-0.017	0.290	0.301	0.936
β_2 = 2.0	0.002	0.170	0.181	0.944	0.008	0.188	0.201	0.930
Setting 2								
α_0 = -1.0	0.003	0.134	0.143	0.938	0.006	0.143	0.152	0.932
α_1 = 1.0	-0.004	0.206	0.207	0.958	-0.012	0.222	0.225	0.950
α_2 = 1.5	0.004	0.119	0.126	0.944	0.006	0.128	0.134	0.930
β_0 = 0.5	-0.059	0.147	0.159	0.920	-0.052	0.159	0.169	0.928
β_1 = 0.0	-0.013	0.215	0.208	0.952	-0.003	0.236	0.233	0.958
β_2 = 0.0	-0.004	0.142	0.144	0.942	-0.001	0.153	0.158	0.946

Bias is the average difference between the true and estimated parameters across the simulations; ASE is the average of the model-based standard error estimates across the simulations; ESD is empirical standard deviation of the parameter estimates seen in simulation; CP is the empirical coverage probability of the true parameter by the model-based 95% confidence interval seen in simulation.

12)}, where $\lambda_C^0(u) = 0.021\, I(0 \leq u < 6)$ $+0.022\, I(6 \leq u < 12)$ $+0.021\, I(u \geq 12)$ in Setting 1, and $\lambda_C^0(u) = 0.016\, I(0 \leq u < 6)$ $+0.015\, I(6 \leq u < 12)$ $+0.015\, I(u \geq 12)$ in Setting 2, producing approximately 25% censoring in each setting.

4.5.2 Simulation Results

Table 4.1 displays finite sample properties of τ-IBR model parameter estimates based on $n = 500$ simulated individuals for both the uncensored case and the case with 25% censoring. The bias of parameter estimates is generally small, although the intercept estimate for model (4.6) shows slightly higher bias than estimates for other model parameters. As one would expect, the variability of each parameter estimate from the model with binary outcomes (model 4.6) is somewhat larger than the variability of each parameter estimate from the model with continuous outcomes (model 4.8). In general, simulation results show good performance of the proposed estimation and inferential procedures as outlined in this manuscript with ASE and ESD estimates close to one another and coverage probabilities close to the desired 95% level.

Table 4.2: Finite sample performance of τ-RMST estimates using (1) the τ-IBR model and (2) the TM model based on 500 iterates.

	No censoring (Setting1)		**25% censoring (Setting1)**		**No censoring (Setting2)**		**25% censoring (Setting2)**	
	τ-IBR	**TM**	**τ-IBR**	**TM**	**τ-IBR**	**TM**	**τ-IBR**	**TM**
700 Subjects								
Bias	-0.021	0.003	-0.019	0.006	0.031	0.069	0.029	0.060
Abs Bias	0.075	0.150	0.078	0.152	0.073	0.090	0.075	0.088
ASE	0.081	0.081	0.089	0.087	0.074	0.074	0.080	0.079
CP	0.909	0.600	0.927	0.632	0.877	0.796	0.895	0.840
500 Subjects								
Bias	-0.022	0.006	-0.019	0.010	0.032	0.069	0.031	0.062
Abs Bias	0.087	0.154	0.091	0.155	0.084	0.098	0.087	0.097
ASE	0.096	0.096	0.106	0.103	0.087	0.087	0.095	0.093
CP	0.909	0.678	0.929	0.713	0.892	0.832	0.914	0.864
200 Subjects								
Bias	-0.025	0.001	-0.014	0.007	0.021	0.057	0.021	0.050
Abs Bias	0.134	0.182	0.140	0.187	0.131	0.136	0.137	0.140
ASE	0.151	0.150	0.166	0.162	0.139	0.138	0.151	0.147
CP	0.918	0.821	0.932	0.837	0.894	0.877	0.904	0.889

Bias: average difference between the true and predicted τ-RMST values across all subjects, windows and simulations; Abs Bias: average of absolute bias across all subjects, windows and simulations; ASE: average of the model-based standard error estimates corresponding to the τ-RMST estimates across all subjects, windows and simulations; CP: empirical coverage probability of the true τ-RMST value by the model-based 95% confidence interval across all subjects, windows and simulations.

Table 4.2 shows finite sample properties of τ-RMST estimates using both our proposed τ-IBR methodology and the TM model for sample sizes of n = 200, 500 and 700. In Setting 2, where there are no interesting covariate associations related to the point mass of restricted event times at $\tau = 6$, both the τ-IBR and the TM perform well in terms of bias and absolute bias, although coverage probabilities are slightly lower for the TM versus the τ-IBR method. ASE values are also comparable between methods.

Setting 1 reflects the case where covariate associations related to the point mass of restricted event times at $\tau = 6$ are in play. Coverage probabilities for the true τ-RMST values using the τ-IBR method are very reasonable, while the TM approach had remarkably poor coverage that deteriorated further with increasing sample sizes (despite ASE getting smaller with increasing sample sizes). And although average bias is low for both methods in this case, the absolute bias was higher for the TM method relative to τ-IBR method.

Figure 4.2 displays differences between estimated and true τ-RMST values using τ-IBR MI and TM methods for representative Setting 1 datasets with n=200, 500, 700 and 900 highlighting subjects whose true τ-RMST were not covered by their estimated 95% confidence interval. Although both methods give unbiased τ-RMST estimates on average, TM τ-RMST estimates had

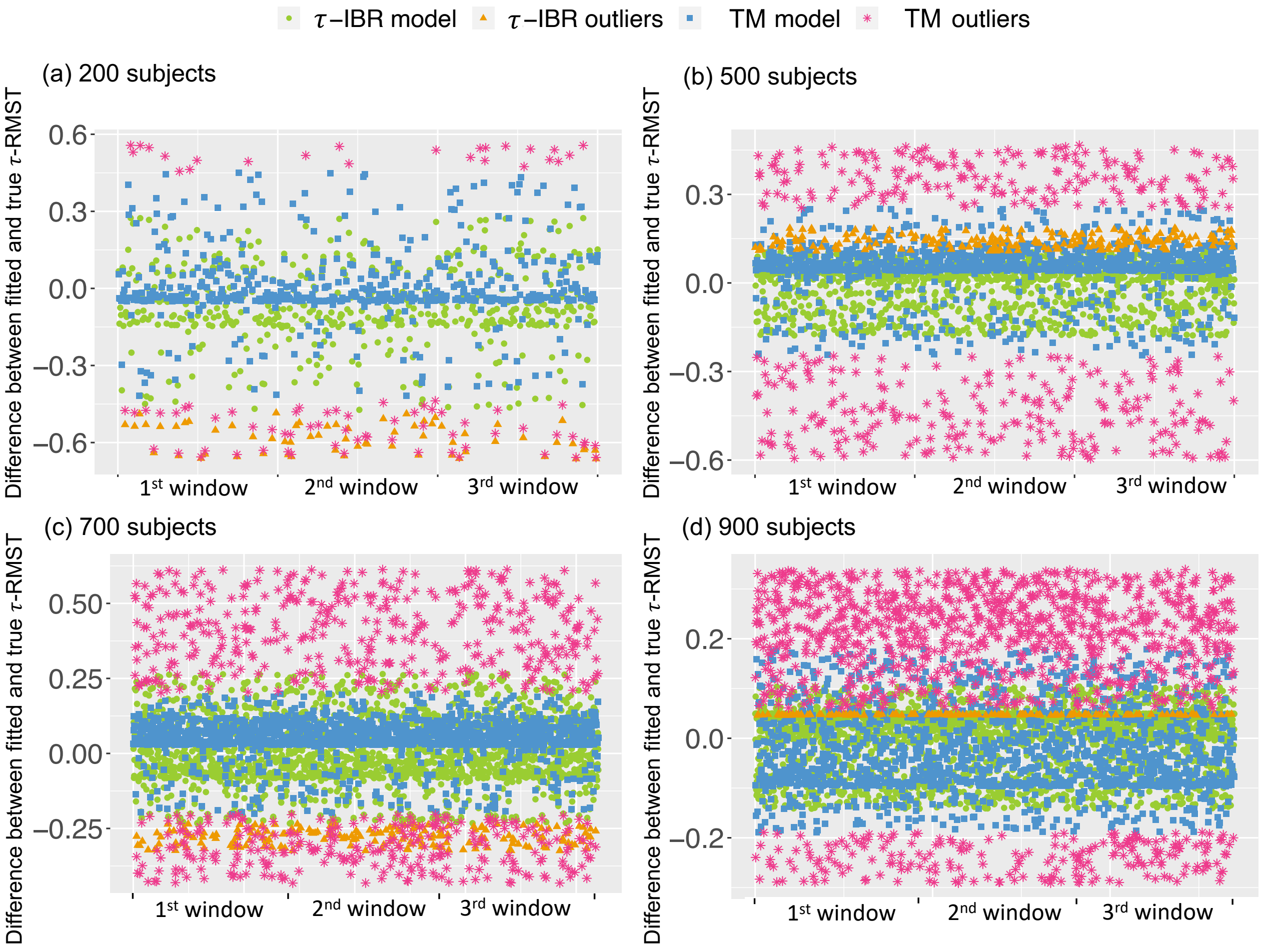

Figure 4.2: Differences between estimated and true τ-RMST values from τ-IBR and TM models based on (a) 200 subjects (b) 500 subjects (c) 700 subjects and (d) 900 subjects in Setting 1.

larger observed differences from their true τ-RMST values that did not improve with larger sample sizes. This larger number of outliers for the TM method in Setting 1 helps explain the higher absolute bias seen for that model in Table 4.2. Hence in Setting 1, where there is important statistical information related to the point mass of restricted event times at $\tau = 6$, ignoring this statistical information had a strong impact on the performance of the TM method.

4.6 Lung transplant candidate 1-year urgency analyses

In this section, we apply the proposed τ-IBR model to estimate 1-year transplant urgency (1-year RMST) in 10,396 lung transplant candidates aged 12 years and above who were newly listed after September 1, 2006, and followed through March 2, 2012. Our data was obtained via a formal data request in December 2022 to the Scientific Registry of Transplant Recipients (SRTR) that maintains data collected by the Organ Procurement and Transplantation Network (OPTN). Mortality information for our requested cohort was augmented by the United States Social Security Death Master file. Candidates are required by the OPTN to update LAS predictors at least once every 6 months. In addition to mortality and transplantation information, our dataset included daily updates on candidate Lung Allocation Scores (LAS) and risk factors used to calculate the LAS urgency score. For participants with multiple listings from the same or different centers, we combined information across these listings. Of the 10,396 candidates included in our data, 7421 received a transplant, 918 died without a transplant, and 2057 were still alive as of March 2, 2012. The interpretation and reporting of these data are the responsibility of the author(s) and in no way should be seen as an official policy of or interpretation by the SRTR or the U.S. Government.

The longitudinal data structure for our analysis has 1-year follow-up windows with starting times of $\{t_1, \ldots, t_6\} = \{0, 6, 12, 18, 24, 30\}$ months. Dependent censoring of waitlist survival during follow-up (due to transplant) was modeled via a Cox model with time-dependent LAS values as well as other factors previously identified by the SRTR as associated with time-to-transplant: gender, race, height, blood type and time-dependent active versus inactive waitlist status. Results from this dependent censoring model are displayed in Supplemental Table C.1 in Appendix C.1. Corresponding IPCW values were calculated from this model as described in Section 4.4. The impact of adjusting for dependent censoring via these IPCW values in transplant urgency calculations is substantial, as can be seen in Supplemental Figure C.1 in Appendix C.2, which displays unadjusted (Kaplan-Meier) versus IPCW-adjusted waitlist survival curve estimates. The IPCW-adjusted waitlist survival curve in this figure is used to implement the MI algorithm described in Section 4.4, resulting in $M = 10$ imputed longitudinal data structures for analysis.

Our analysis evaluates the same predictors and interaction terms used in LAS urgency estimation, with no additional model selection performed. Some of the most important terms in the urgency models have been the major diagnosis grouping variables, where Group A candidates were predominantly diagnosed with chronic obstructive pulmonary disease (COPD), group B candidates with pulmonary hypertension, group C candidates with cystic fibrosis, and group D candidates with idiopathic pulmonary fibrosis. Several terms related to smaller diagnosis groups were originally included in the LAS algorithm because these groups felt strongly about having their diagnoses individually represented, Other clinical features have been maintained as part of the LAS algorithm due to showing statistical significance in earlier analysis cohorts. Table 4.3 summarizes parameter estimates, 95% confidence intervals and p-values from the τ-IBR model and its nearest competitor, the TM model, where both of these models used data imputed using our proposed MI algorithm. The majority of LAS features retained statistical significance in this cohort. For the purpose of this example, we'll focus on some of the strongest predictors that have been historically important in urgency estimation: O_2 requirement at rest (O_2), 6-minute walk distance (6MWD), continuous mechanical ventilation (CMV), diagnosis group (A, B, C or D) and age. The majority of these are statistically significant in the TM model as well as the τ-IBR models for μ_i and π_i. Diagnosis group has more complexity because of its many interaction terms. The clinical expectation is that diagnosis group A will have less urgency than the other diagnosis groups since many COPD candidates pursue a transplant to enhance their quality of life rather than to increase their lifespan. Consequently, they are more likely to survive beyond a year on the waitlist, but with increasingly poor quality of life. This clinical expection is statistically supported by results of the τ-IBR model for π_i, model (4.6). But this intuition is not seen as clearly when viewing the TM model diagnosis group results, with main effects for diagnosis groups B and D missing statistical significance.

To assist in assessing model fit, the τ-IBR MI approach for dependently censored outcomes allows us to visualize individual-level 1-year restricted lifetimes using popular data views for uncensored data. The heatmap in Figure 4.3 focuses on lung candidates with either observed or multiply imputed lifetimes less than 1 year and gives an observed versus expected view of these individuals' first follow-up year. The three heatmap columns represent (1) 1-year restricted lifetimes with censored values replaced with multiply imputed outcomes averaged across 10 imputed datasets (those with observed data augmented via imputation are designated as Imputed=Yes in the bar immediately to the left of the heatmap columns), (2) 1-year τ-IBR RMST estimates and (3) 1-year TM RMST estimates, respectively. Longer and shorter survival times on the heatmap scale are shown in yellow and purple color ranges, respectively. Additional individual characteristics, including O_2 requirement at rest, 6-min walk distance, diagnosis group, and age, are color-coded along the left side of the heatmap. Descriptively, the cluster with the lowest 1-year RMST values is more likely to have 6-min walk distance less than 1000 feet and continuous mechanical ventilation, and is predominantly composed of patients in diagnosis groups B, C, and D. In contrast, patients in group A, with 6-min walk distance larger than 1000 feet and no continuous mechanical ventilation, tend to have the highest 1-year RMST values. For the most urgent lung candidates in the purple range of column one of the heatmap, where good model fit is particularly important for ranking candidates for transplant, there are clusters where urgency is captured well by both the τ-IBR and TM methods, with the τ-IBR estimates slightly closer to the augmented raw data than the TM estimates. However, many individuals in the dark purple range of column one do not have adequate urgency estimates using either method, suggesting room for improvement in the lung allocation urgency model as newer predictors become available. This particular data view has not previously been available to analysts of dependently censored lung candidate data.

4.7 Discussion

This manuscript offers a new modeling framework for a modern analysis of dependently censored time-to-event data within the context of restricted mean regression models. By converting traditional time-to-event data into a censored longitudinal data structure, we are able to leverage longitudinal data analysis experience in the analysis of restricted mean survival time (RMST) and utilize information across multiple follow-up windows to enhance the efficiency of RMST estimation. Breaking down τ-RMST into its mixture components through equation (1) enables us to examine the nuanced effects of factors. In paying close attention to modeling individuals with a τ-RMST = τ, we have found that much of the statistical signal for τ-RMST estimation lies in those individuals who have not experienced the event by time τ. Simulation results demonstrate that our proposed τ-IBR model, which separately models components of the mixture distribution laid out in equation (1), achieves higher coverage probabilities for τ-RMST estimation compared to the TM model, which models RMST directly. Our lung urgency data example confirms that predictors for models of μ_i and π_i can have varying degrees of clinical and statistical signficance that are not captured by the TM model. The τ-IBR model was particularly helpful when validating the contribution of diagnosis group to urgency estimation in this cohort.

This manuscript further contributes to multiple imputation approaches for dependently censored data. In addition to enhancing risk set definitions and using IPCW adjustment throughout the otherwise non-parametric inverse-transform imputation procedure, we develop a graphical procedure for assessing model fit via looking at observed versus expected τ-RMST (as in Figure 4.3). This diagnostic has not previously been seen in dependently censored τ-RMST estimation literature.

Table 4.3: Lung transplant candidate 1-year urgency model parameters with 95% confidence intervals and p-values.

	τ-IBR Model for μ_i			τ-IBR Model for π_i			TM Model		
	Fold Change*	95% CI	P-value	Odds Ratio[†]	95% CI	P-value	Coef/τ	95% CI	P-value
Age (per 20 years increase)	1.00	(0.96,1.04)	0.919	0.74	(0.65,0.84)	<0.001	-0.02	(-0.03,-0.01)	<0.001
BMI (per 1 kg/m decrease if BMI<20kg/m)	0.99	(0.98,1.01)	0.585	0.88	(0.81,0.93)	<0.001	-0.01	(-0.02,-0.01)	<0.001
Cardic Index<2.0 (ref=CI>2)	0.92	(0.85,1.00)	0.044	0.69	(0.53,0.91)	0.009	-0.04	(-0.07,-0.02)	0.003
CVP in Group B (per 10 mm Hg)	0.92	(0.79,1.05)	0.212	0.78	(0.51,1.19)	0.254	-0.04	(-0.09,0.01)	0.116
Continuous Mechanical Ventilation (ref=other/no Ventilation)	0.52	(0.42,0.62)	<0.001	0.05	(0.03,0.07)	<0.001	-0.52	(-0.58,-0.47)	<0.001
Creatinine (Age>18 years)	0.97	(0.90,1.03)	0.310	1.01	(0.83,1.21)	0.956	-0.01	(-0.02,0.01)	0.351
Diabetes (ref=no diabetes)	0.99	(0.94,1.03)	0.552	0.86	(0.74,1.00)	0.045	-0.02	(-0.03,0.00)	0.027
Diagnosis group (ref=group A)									
Group B	1.00	(0.86,1.15)	0.968	0.29	(0.16,0.52)	<0.001	-0.01	(-0.06,0.05)	0.808
Group C	1.07	(0.97,1.17)	0.176	0.19	(0.14,0.26)	<0.001	-0.05	(-0.08,-0.03)	<0.001
Group D	1.07	(0.97,1.16)	0.168	0.40	(0.29,0.56)	<0.001	0.00	(-0.03,0.03)	0.844
Bronchiectasis	1.07	(0.94,1.21)	0.291	0.59	(0.38,0.91)	0.017	-0.01	(-0.04,0.02)	0.516
Eisenmenger Syndrome	0.89	(0.62,1.15)	0.387	2.62	(0.82,8.40)	0.105	0.06	(-0.02,0.14)	0.137
Lymphangioleiomyomatosis	1.17	(0.97,1.36)	0.110	0.41	(0.18,0.93)	0.032	-0.02	(-0.06,0.02)	0.299
Obliterative Bronchiolitis	1.21	(0.99,1.43)	0.094	2.58	(1.21,5.48)	0.014	0.09	(0.05,0.14)	<0.001
Pulmonary Fibrosis Other	0.99	(0.92,1.06)	0.802	1.32	(1.07,1.62)	0.009	0.02	(0.00,0.05)	0.043
Sarcoidosis with PA mean>30mm Hg in group D	1.05	(0.94,1.15)	0.409	2.34	(1.61,3.41)	<0.001	0.09	(0.05,0.13)	<0.001
Sarcoidosis with PA mean$\leq$ 30mm Hg in group A	1.01	(0.86,1.17)	0.893	0.34	(0.22,0.55)	<0.001	-0.04	(-0.07,-0.01)	0.003
FVC (per 10% decrease if FVC % predicted<80%) in group D	0.99	(0.97,1.00)	0.089	0.82	(0.78,0.87)	<0.001	-0.02	(-0.03,-0.02)	<0.001
No assistance with ADL (ref=some/total assisstance)	1.00	(0.92,1.07)	0.904	1.50	(1.24,1.81)	<0.001	0.01	(0.00,0.02)	0.013
O_2 requirement at rest in group B	0.98	(0.97,1.00)	0.013	0.93	(0.89,0.97)	<0.001	-0.02	(-0.02,-0.01)	<0.001
O_2 requirement at rest in Groups A, C, or D	0.98	(0.98,0.99)	<0.001	0.83	(0.81,0.85)	<0.001	-0.02	(-0.02,-0.02)	<0.001
PCO_2 (per 10 mm Hg)	0.99	(0.97,1.01)	0.287	0.82	(0.77,0.88)	<0.001	-0.01	(-0.02,-0.01)	<0.001
PCO_2 increase of $\geq$ 15% (ref = PCO_2 increase of<15%)	1.00	(0.92,1.09)	0.930	1.11	(0.82,1.50)	0.517	0.00	(-0.02,0.03)	0.943
PA systolic in group A (per 10 mm Hg increase if PA systolic>40 mm Hg)	1.01	(0.98,1.05)	0.456	0.68	(0.61,0.76)	<0.001	-0.01	(-0.02,0.00)	0.010
PA systolic in group B, C or D (per 10 mm Hg increase)	0.99	(0.98,1.00)	0.252	0.89	(0.85,0.93)	<0.001	-0.01	(-0.02,-0.01)	<0.001
6-min walk distance (per 1000 feet)	1.10	(1.06,1.14)	<0.001	1.67	(1.48,1.88)	<0.001	0.05	(0.04,0.06)	<0.001

*Amongst those who die during a follow-up year, fold change for predictors is the ratio of estimated survival times comparing those with predictor equal to 1 versus predictor equal to 0, assuming all other predictors are zero; [†]Odds ratio for predictors compares the odds of remaining alive at 1 year for those with versus without a one unit increase in the predictor (unless otherwise stated), adjusted for other covariates in the model; [⋆]Percentage increase in 1-year-RMST for each unit increase of the predictor, adjusted for other covariates in the model.

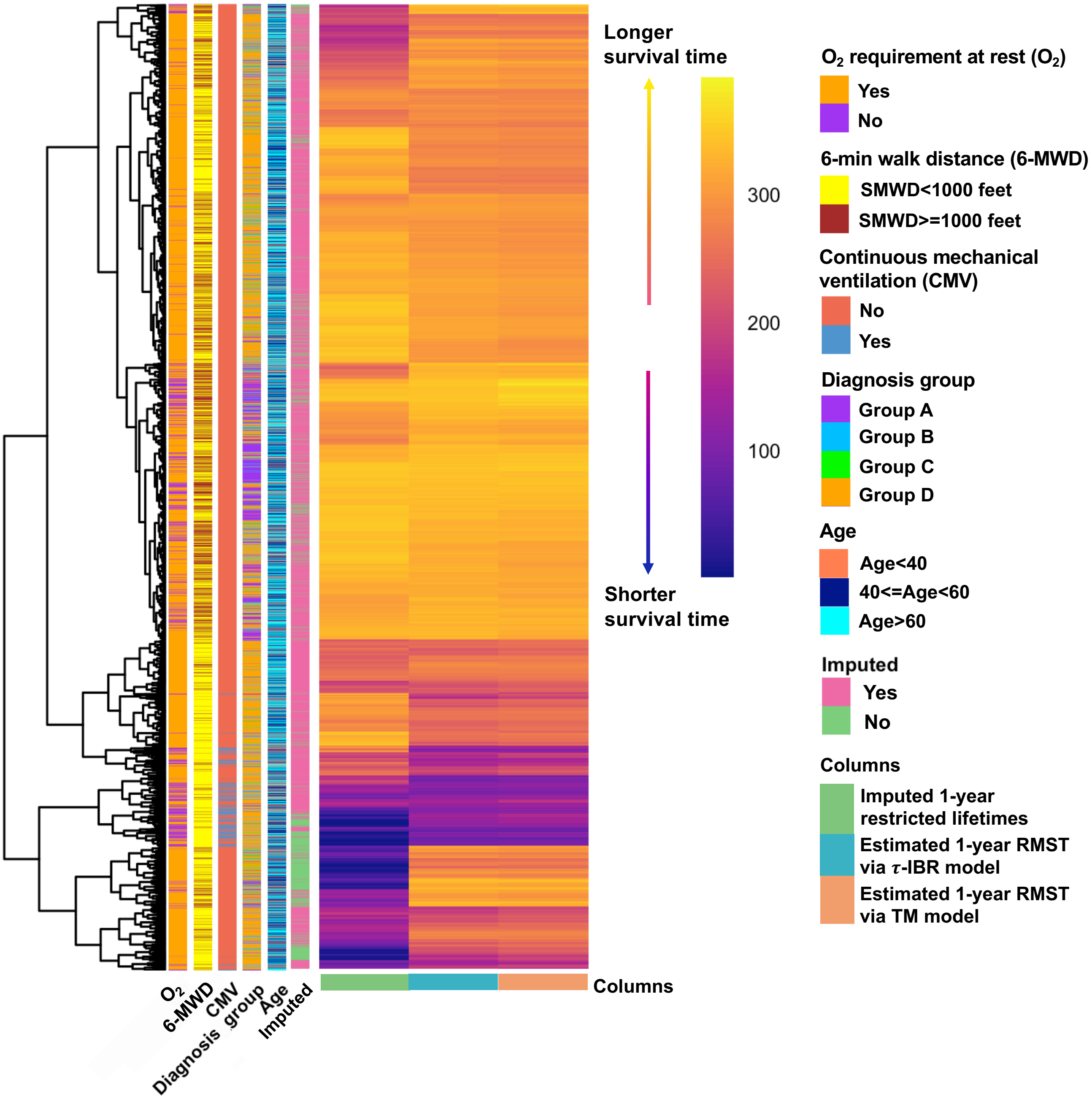

Figure 4.3: Observed versus expected one-year restricted lifetimes for lung candidates with either observed or multiply imputed lifetimes less than one year.

CHAPTER 5

Conclusion

The aim of this dissertation is to develop statistical methodologies that enhance the estimation of τ-restricted event-free time and deepen the understanding of its relationship with predictors. To this end, we emphasize the point mass at τ caused by the mixture of individuals that achieve an active event endpoint during a τ-restricted follow-up period as opposed to individuals who are event-free during that follow-up period. By modeling these two groups separately, we can examine the nuanced effects of various factors and attain greater precision in statistical inferences regarding τ-restricted event-free time compared to approach modeling τ-restricted event-free time directly. Applying this novel modeling framework to single time-to-event data, recurrent events data, and single time-to-event data subject to dependent censoring, along with the corresponding estimation methods proposed, can confirm the benefits of our proposed model in various settings.

The first setting is clinical trial analysis where the outcome of interest is a single time-to-event. Chapter 2 develops a τ-IBR model framework with an EM algorithm and a semi-parametric MI algorithm for fitting and reporting results for the τ-IBR model. Our simulation results confirm better precision of τ-RMST estimates and better corresponding confidence interval coverage rates when the point mass at τ is more appropriately modeled regardless of estimation approach (EM, MI) or censoring mechanism. The benefits of our proposed τ-IBR model also manifest in the COPD example. We noticed that the relative importance of risk factors in beta model and logistic model shifted between the two models, with most statistical signal appearing in the logistic model. In particular, the treatment effect manifested significantly in the logistic model and not the beta model. The standard τ-RMST model identifies a significant treatment effect as well, but is not able to distinguish the nature of the treatment effect as clearly. To our knowledge, a τ-inflated beta regression model has never been proposed as a way to model time-to-event data, censored or otherwise. The key advantages of this method are (1) a better understanding of predictors associated with no events in the τ-restricted period of interest, as opposed to predictors associated with shorter expected event-free time amongst those who experienced the event, (2) more efficient estimation of restricted means due to properly modeling the point mass of min(τ, T) events at τ and (3) parametric assumptions that are flexible for modeling the respective means of min(τ, T) given min(τ, T) $<$ τ and binary variable indicating if min(τ, T) = τ without making unnecessarily restrictive assumptions about proportional hazards.

In the second setting, the primary endpoints are recurrent instead of a single time-to-event endpoint. Chapter 3 offers a suite of new methods for a modern analysis of recurrent events data subject to censoring. By converting traditional recurrent event data into a censored longitudinal data structure, we are able to apply our τ-IBR model framework to recurrent events data and leverage longitudinal data analysis experience in the analysis of this data. The τ-IBR model is a one-stop shop for assessing overall susceptibility to recurrent events via model (3.5), changes to susceptibility to recurrent events over time via model (3.5) with window start times as covariates and interactions, the influence of predictors on time-to-first-event amongst those who experience it via model (3.6) with and without window start times as covariates and interactions, as appropriate, and patterns of mean time-to-first-recurrent event in follow-up windows over time by combining results from models (3.5) and (3.6). Simulations indicate excellent performance of the τ-IBR model in various settings and show that the XMT model, the nearest competitor model that ignores point masses, also has good coverage rates for true τ-RMST values when there are no predictors associated with the point mass of restricted event times at τ. These coverage rates fall apart when predictors related to susceptibility of recurrent events at τ exist (simulation setting 1), even when there are no censored event times. Data visualizations of model results shown in section 3.6 are an additional tool for making results interpretable to a medical research readership.

In our third setting, we develop a τ-IBR model of dependently censored time-to-event data based on the censored longitudinal data framework in Chapter 4. Transforming traditional single time-to-event data into a censored longitudinal data structure enables (a) easily incorporating information from time-dependent risk factors past baseline by including covariates updated at each start time of follow-up window in the model and (b) utilizing information across multiple follow-up windows to enhance the efficiency of RMST estimation. We also aim to address dependent censoring, a feature of the data that requires consideration. To accomplish this, we follow the multiple imputation approach proposed by Tayob and Murray [60]. This approach imputes censored values from a risk set of patients with similar characteristics and censoring probability. In the case of dependent censoring, MI approaches that select candidates from an appropriate risk set are attractive. The idea behind this is that survival within the risk set is homogeneous, providing an unbiased imputation. We incorporate the inverse probability of censoring weighting (IPCW) method to further correct dependent censoring within the risk set when estimating survival. This double correction makes our approach sufficiently address dependent censoring in our application. Compared to the model used in Tayob and Murray (2017)[60], which models RMST directly, our proposed τ-IBR model, which separately models components of the mixture distribution, achieves higher coverage probabilities for τ-RMST estimation in simulations. Our lung urgency data example confirms that predictors for models of μ_i and π_i can have varying degrees of clinical and statistical significance that are not captured by the TM model. The τ-IBR model was particularly helpful when validating the contribution of diagnosis group to urgency estimation in this cohort.

In Chapters 3 and 4, our τ-IBR models are developed based on a censored longitudinal data structure. The objective is to estimate patient i's $\pi_i(t)$ and $\mu_i(t)$ for a τ-duration follow-up window starting at time t, given the corresponding covariates $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$, respectively. The underlying assumptions of how these terms relate to τ-RMST for the follow-up window starting at time t are given in equations (3.3) and (4.4) in Chapter 3 and Chapter 4, respectively. Similar to standard longitudinal analysis, we can incorporate window start times t, time-dependent covariates that change at the window start times, and interactions between t and other covariates in $Z_{\pi i}(t)$ and $Z_{\mu i}(t)$. For instance, in the example section of Chapter 3, Figure 3.3 displays variation in time-to-first-recurrent event patterns across different follow-up windows. Therefore, we assessed the stability of the treatment effect over the follow-up time windows by including window start times as predictors in the models, along with appropriate interaction terms. Just like in standard longitudinal analysis, our modeling approach also allows for statistical tests to evaluate whether patient i's τ-RMST over the period starting at follow-up time t depends on t or is stable across all follow-up windows.

The decision to include time-dependent covariates in longitudinal models must always be made in the context of the research question. For instance, in the assessing the effect of azithromycin, covariates that change after baseline and that can be affected by treatment would be inappropriate to adjust for. In Chapter 4, the designers of the LAS had a specific objective to create a score unaffected by the duration of time spent on the waitlist. Consequently, window start times were excluded from consideration in these models. As such, resulting urgency estimates reflect a mixture of urgency across the various follow-up windows included in the τ-IBR analyses.

Computational times can become significant when dealing with large sample sizes or a substantial number of follow-up windows in the analyses. Regarding the selection of follow-up window parameters, several suggestions are provided in the literature [58, 67, 70]. In the example section of Chapter 4, a spacing of 6 months between follow-up windows is reasonable, considering that patients are required to update their medical information at least once every 6 months. However, in settings where covariate data is updated more frequently, the cost of including more frequent follow-up windows is computational time. Similar considerations apply to the setting of recurrent events. While it is theoretically possible to initiate follow-up windows every day to capture every recurrent event and increase efficiency, this would result in a significant computational burden. The paper by Xia and Murray [67] demonstrates that the computational burden significantly increases as the number of follow-up windows incorporated into the analyses increases. Hence, it is crucial to strike a balance between the number of follow-up windows and computational feasibility when designing the analysis strategy.

For instance, when analyzing restructured azithromycin trial data with 1112 participants and 4 6-month follow-up windows starting every 2 months, our method's computation took 4 seconds using the ES algorithm and 30 seconds using the MI approach. However, when analyzing restructured lung transplant waitlist data with 10,396 candidates and 6 1-year follow-up windows starting every 6 months, the computation of MI-fitted τ-IBR models took 110 minutes due to the large sample sizes and the setting of follow-up windows. The most time-consuming part of the MI approach is the construction of the risk set for each individual requiring imputation. The requirement of matching histories of predicted outcomes in Chapter 4 significantly increases the computational time involved in selecting suitable candidates within the risk set. It's also worth noting that we did not encounter any convergence issues throughout the fitting procedure of τ-IBR models using either the ES algorithm or the MI approach.

When applying τ-IBR models to complex real-world data using data restructuring techniques, aligning the start times of windows with the dates of updated covariates can sometimes pose challenges. Various imputation approaches can be employed to define covariates at the start times of windows. For instance, missing covariates can be replaced with the closest available value or imputed based on their association with outcomes through fitting a predictive model. Exploring these imputation methods and their application in different data settings could lead to further enhancements of our method. Moreover, in scenarios where the data dimension is large and complex relationships exist between predictors and outcomes, our methods may encounter computational and fitting difficulties. To address this, one promising avenue for future research is to incorporate machine learning methodologies with our approaches to analyze τ-RMST, with a particular focus on capturing the point mass at τ. This combination of techniques has the potential to advance the analysis of complex data and improve the performance of our models.

Overall, this dissertation gives a useful set of tools for the analysis of single and recurrent time-to-event data. These methods fill a nice gap in the literature for τ-restricted event time estimation and inference.

APPENDIX A

Supplementary Materials for Chapter 2

A.1 Computation of the estimate of the variance of τ-RMST

We calculate the estimate of the variance of restricted mean survival time of each subject i as follows:

$$\begin{aligned}\text{Var}\{\hat{\text{E}}[\min(T_i,\tau)]\} &= \tau^2\text{Var}\big[\hat{\mu}_i(1-\hat{\pi}_i)+\hat{\pi}_i\big]\\ &= \tau^2\big\{\text{E}\big\{\text{Var}\big[\big(\hat{\mu}_i(1-\hat{\pi}_i)+\hat{\pi}_i\big)|\hat{\mu}_i\big]\big\} + \tau^2\text{Var}\big\{\text{E}\big[\big(\hat{\mu}_i(1-\hat{\pi}_i)+\hat{\pi}_i\big)|\hat{\mu}_i\big]\big\}\\ &= \tau^2\text{E}\big[(1-\hat{\mu}_i)^2\text{Var}(\hat{\pi}_i)\big] + \tau^2\text{Var}\big[\hat{\mu}_i+(1-\hat{\mu}_i)\text{E}(\hat{\pi}_i)\big],\end{aligned}$$

which is asymptotically equivalent to

$$\begin{aligned}&= \tau^2(1-\mu_i)^2\text{Var}(\hat{\pi}_i) + \tau^2\text{Var}\big[\hat{\mu}_i+(1-\hat{\mu}_i)\pi_i\big]\\ &= \tau^2(1-\mu_i)^2\text{Var}(\hat{\pi}_i) + \tau^2\text{Var}\big[(1-\pi_i)\hat{\mu}_i+\pi_i)\big]\\ &= \tau^2(1-\mu_i)^2\text{Var}(\hat{\pi}_i) + \tau^2(1-\pi_i)^2\text{Var}(\hat{\mu}_i),\end{aligned}$$

then the asymptotic estimate of the variance of restricted mean survival time for subject i becomes:

$$\widehat{\text{Var}}\{\hat{\text{E}}[\min(T_i,\tau)]\} = \tau^2(1-\hat{\mu}_i)^2\widehat{\text{Var}}(\hat{\pi}_i) + \tau^2(1-\hat{\pi}_i)^2\widehat{\text{Var}}(\hat{\mu}_i). \tag{A.1}$$

Based on model (2.2) and model (2.3), by using delta method the first term of equation (A.1) is seen to be:

$$\begin{aligned}\tau^2(1-\hat{\mu}_i)^2\widehat{\text{Var}}(\hat{\pi}_i) &= \tau^2\left(1-\frac{1}{1+e^{-\hat{\alpha}^T Z_i^{\mu}}}\right)^2\widehat{\text{Var}}\left(\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}}}\right)\\ &= \tau^2\left(1-\frac{1}{1+e^{-\hat{\alpha}^T Z_i^{\mu}}}\right)^2 Z_i^{\pi T}\hat{V}_\beta Z_i^{\pi}\frac{(e^{-\hat{\beta}^T Z_i^{\pi T}})^2}{(1+e^{-\hat{\beta}^T Z_i^{\pi}})^4},\end{aligned}$$

the second term of equation (A.1) is given by:

$$\begin{aligned}\tau^2(1-\hat{\pi}_i)^2\widehat{\text{Var}}(\hat{\mu}_i) &= \tau^2\left(1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}}}\right)^2 \widehat{\text{Var}}\left(\frac{1}{1+e^{-\hat{\alpha}^T Z_i^{\mu}}}\right)\\ &= \tau^2\left(1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}}}\right)^2 Z_i^{\mu T}\hat{V}_{\alpha}Z_i^{\mu}\frac{(e^{-\hat{\alpha}^T Z_i^{\mu}})^2}{(1+e^{-\hat{\alpha}^T Z_i^{\mu}})^4}.\end{aligned}$$

where $Z_i^{\mu T} = (1, {Z_{\mu i}}^T)$, $Z_i^{\pi T} = (1, {Z_{\pi i}}^T)$, $\hat{V}_{\beta}$ and $\hat{V}_{\alpha}$ are the estimates of covariance matrix of coefficients in model model (2.2) and model (2.3) respectively.

A.2 Calculation of the variance-covariance matrix of estimated parameters using Louis method

Recall the section 2.3.1,

$$\widehat{\text{Var}}(\hat{\theta}^{EM}) = \left[-\frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\theta\partial\theta^T} - \text{Var}\left(\frac{\partial l_c(\theta)}{\partial\theta}\right)\right]^{-1}\Bigg|_{\theta=\hat{\theta}^{EM}},$$

where

$$\frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\theta\partial\theta^T} = \begin{bmatrix} \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\beta\partial\beta^T} & 0 & 0 \\ 0 & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\alpha^T} & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\nu} \\ 0 & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\nu} & \frac{\partial^2\mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\nu^2} \end{bmatrix},$$

and

$$\text{Var}\left(\frac{\partial l_c(\theta)}{\partial\theta}\right) = \begin{bmatrix} \text{Var}(U_{\beta}) & 0 & 0 \\ 0 & \text{Var}(U_{\alpha}) & \text{Cov}(U_{\alpha},U_{\nu}) \\ 0 & \text{Cov}(U_{\alpha},U_{\nu}) & \text{Var}(U_{\nu}) \end{bmatrix}.$$

We first calculate each components of $\text{Var}\left(\frac{\partial l_c(\theta)}{\partial\theta}\right)$ as follows: The first step is to obtain the derivatives of the complete log-likelihood $l_c(\theta)$:

$$U_{\beta} = \frac{\partial l_c(\theta)}{\partial\beta} = \sum_{i=1}^{n_1}[b_i - \pi_i(\beta)]Z_i^{\pi} + \sum_{i=n_1+1}^{n}[B_i - \pi_i(\beta)]Z_i^{\pi}$$

$$
\begin{aligned}
U_\alpha = \frac{\partial l_c(\theta)}{\partial \alpha} &= \sum_{i=1}^{n_1}(1-b_i)\nu_i(y_i^* - \psi_i^*)\mu_i(\alpha)[1-\mu_i(\alpha)]Z_i^\mu \\
&+ \sum_{i=n_1+1}^{n} \frac{B_i - 1}{1 - F_{Y_i}(y_i;\mu_i(\alpha),\nu)} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \alpha} \\
U_\nu = \frac{\partial l_c(\theta)}{\partial \nu} &= \sum_{i=1}^{n_1}(1-b_i)\big[\mu_i(\alpha)(y_i^* - \psi_i^*) + \phi_i^* + \log\big(1-y_i\big)\big] \\
&+ \sum_{i=n_1+1}^{n} \frac{B_i - 1}{1 - F_{Y_i}(y_i;\mu_i(\alpha),\nu)} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \nu};
\end{aligned}
$$

here $y_i^* = \log[y_i/(1-y_i)]$, $\psi_i^* = \psi[\mu_i(\alpha)\nu_i] - \psi\{[1-\mu_i(\alpha)]\nu_i\}$, $\phi_i^* = \psi(\nu_i) - \psi\{[1-\mu_i(\alpha)]\nu_i\}$, where $\psi(x) = d\log\Gamma(x)/dx$, later it's also convenient to define $\psi'(x) = [1/\Gamma(x)][d\Gamma(x)/dx]$, and $F_{Y_i}(y_i;\mu_i(\alpha),\nu)$ is the cumulative distribution for Y_i evaluated at y_i.

Elements of Var$\left(\frac{\partial l_c(\theta)}{\partial \theta}\right)$ are listed next, where all variance and covariance terms corresponding to B_i are conditioned on the observed data and $\hat{\theta}^{EM}$:

$$
\text{Var}(U_\beta) = \sum_{i=n_1+1}^{n} w_i(1-w_i)Z_i^\pi Z_i^{\pi T}
$$

$$
\text{Var}(U_\alpha) = \sum_{i=n_1+1}^{n} \frac{w_i(1-w_i)}{[1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]^2} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \alpha} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \alpha^T}
$$

$$
\text{Cov}(U_\alpha, U_\nu) = \sum_{i=n_1+1}^{n} \frac{w_i(1-w_i)}{[1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]^2} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \alpha} \frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial \nu}.
$$

Here,

$$
\begin{aligned}
w_i &= \text{E}(B_i|Y_i \geq y_i, i = n_1+1,\ldots,n,\hat{\theta}^{EM}) \\
&= \frac{\pi_i(\hat{\beta}^{EM})}{\pi_i(\hat{\beta}^{EM}) + [1-\pi_i(\hat{\beta}^{EM})][1-F_{Y_i}(y_i;\mu_i(\hat{\alpha}^{EM}),\hat{\nu}^{EM})]}.
\end{aligned}
$$

Components of $\frac{\partial^2 \mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\theta\partial\theta^T}$ become:

$$
\frac{\partial^2 \mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} Z_i^\pi Z_i^{\pi T}\pi_i(\beta)[1-\pi_i(\beta)],
$$

$$\begin{aligned}\frac{\partial^2 \mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\alpha^T} &= \sum_{i=1}^{n_1+1}(1-b_i)[\nu(y_i^* - \phi_i^*)\mu(\alpha)[1-\mu(\alpha)][1-2\mu(\alpha)] \\ &- \nu^2\mu_i(\alpha)^2[1-\mu(\alpha)]^2\psi_i^\dagger]Z_i^\mu Z_i^{\mu T} \\ &+ \sum_{i=n_1+1}^{n}\left\{\frac{w_i-1}{1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)}\frac{\partial^2 F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\alpha\partial\alpha^T}\right. \\ &+ \left.\frac{w_i-1}{[1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]^2}\frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\alpha}\frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\alpha^T}\right\}, \\ \frac{\partial^2 \mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\nu^2} &= \sum_{i=1}^{n_1}(1-b_i)\{-\mu_i^2(\alpha)\psi_i^\dagger + 2\mu_i(\alpha)\psi_i'\{[1-\mu_i(\alpha)]\nu_i\} + \phi_i^\dagger\} \\ &+ \sum_{i=n_1+1}^{n}\left\{\frac{w_i-1}{1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)}\frac{\partial^2 F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\nu^2}\right. \\ &+ \left.\frac{w_i-1}{[1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]^2}\left(\frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\nu}\right)^2\right\}\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{Q}(\theta;\hat{\theta}^{EM})}{\partial\alpha\partial\nu} &= \sum_{i=1}^{n_1}(1-b_i)\{y_i^* - \psi_i^* - \nu_i\mu_i(\alpha)\psi_i^\dagger + \nu_i\psi_i'[\nu_i - \mu_i(\alpha)\nu_i]\}\mu_i(\alpha)[1-\mu_i(\alpha)]Z_i^\mu \\ &+ \sum_{i=n_1+1}^{n}\left\{\frac{w_i-1}{1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)}\frac{\partial^2 F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\nu\partial\alpha}\right. \\ &+ \left.\frac{w_i-1}{[1-F_{Y_i}(y_i;\mu_i(\alpha),\nu)]^2}\frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\nu}\frac{\partial F_{Y_i}(y_i;\mu_i(\alpha),\nu)}{\partial\alpha}\right\}.\end{aligned}$$

with $\psi_i^\dagger = \psi'[\mu_i(\alpha)\nu] + \psi'\{[1-\mu_i(\alpha)]\nu\}$ and $\phi_i^\dagger = \psi'(\nu) - \psi'[(1-\mu_i)\nu]$. We used the grad and hessian functions from the numDeriv R package to approximate the first derivative and the second derivative of the cumulative distribution function of beta distribution $F_{Y_i}(y_i;\mu_i(\alpha),\nu)$ with respect to α and ν using Richardson's extrapolation in the calculation the variance-covariance matrix of $\hat{\theta}$.

A.3 Supplementary Figure of Simulation Section 2.4.2

In this section we show supplemental simulation results comparing our proposed τ-IBR model and traditional τ-RMST model. This figure displays differences between estimated and true τ-RMST values for representative datasets with n=100, 500, 1000 and 1500 subjects for the dependent censoring setting, highlighting subjects whose true τ-RMST values were not covered by each method's estimated 95% confidence interval. τ-IBR estimates are based on the MI algorithm described in section 2.3.2.

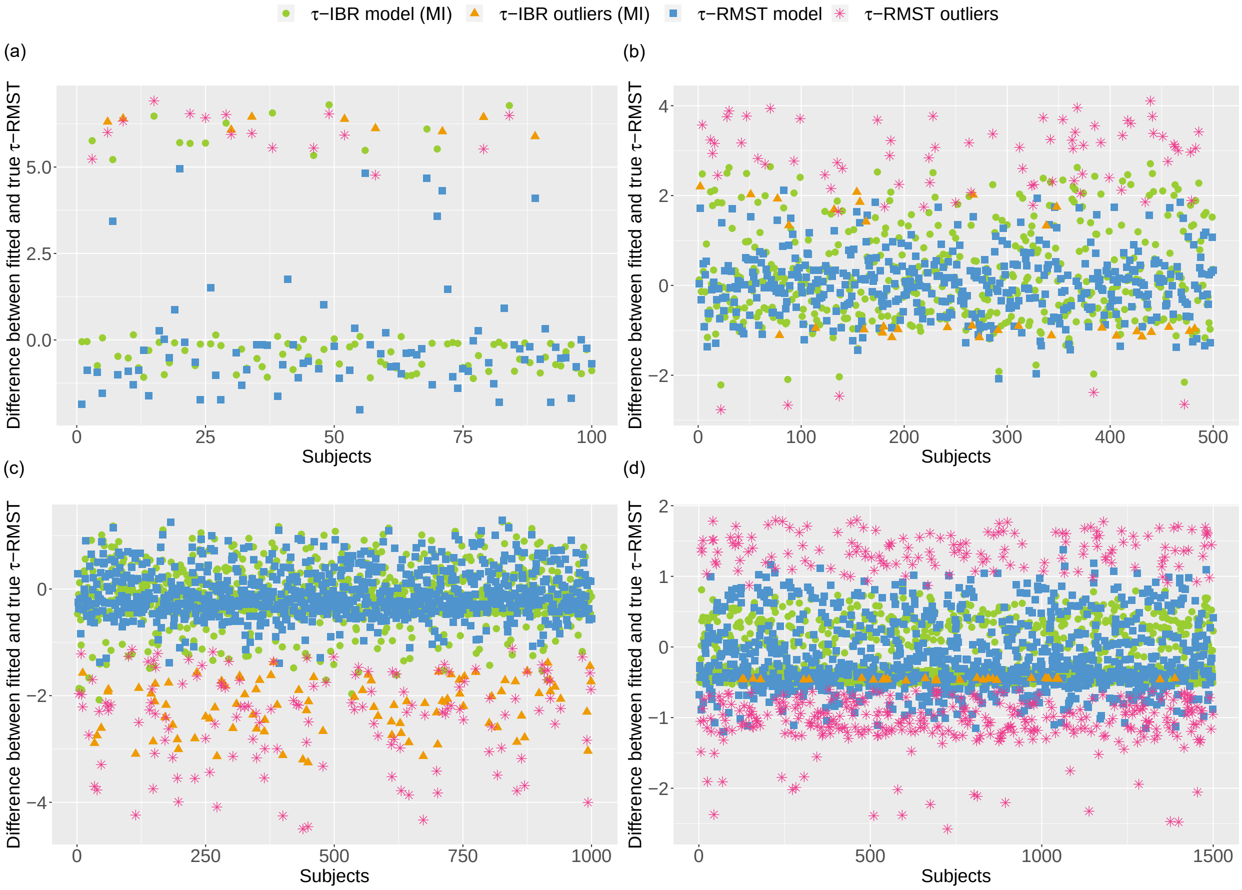

Figure A.1: Difference between MI-fitted τ-RMST and actual τ-RMST based on (a) 100 subjects (b) 500 subjects (c) 1000 subjects and (d) 1500 subjects.

A.4 Supplementary Tables of Example Section 2.5

Table A.1: Estimated effect of azithromycin for varying τ in τ-IBR and τ-RMST multivariable models.

	τ= 3 months	**τ= 6 months**	**τ= 9 months**	**τ= 12 months**
τ-IBR model (MI) (Beta Regression)				
Fold Change*	0.904	1.020	0.996	1.025
95% CI	(0.790, 1.018)	(0.903, 1.137)	(0.899, 1.092)	(0.926, 1.123)
P-value	0.112	0.734	0.987	0.617
τ-IBR model (MI) (Logistic Regression)				
Odds Ratio†	1.489	1.559	1.692	1.766
95% CI	(1.143, 1.938)	(1.218, 1.997)	(1.320, 2.168)	(1.358, 2.296)
P-value	0.003	<0.001	<0.001	<0.001
τ-IBR model (EM) (Beta Regression)				
Fold Change*	0.905	1.021	0.998	1.026
95% CI	(0.791, 1.020)	(0.905, 1.137)	(0.902, 1.094)	(0.927, 1.125)
P-value	0.116	0.724	0.972	0.606
τ-IBR model (EM) (Logistic Regression)				
Odds Ratio†	1.488	1.552	1.676	1.748
95% CI	(1.143, 1.937)	(1.213, 1.986)	(1.307, 2.149)	(1.345, 2.271)
P-value	0.003	<0.001	<0.001	<0.001
τ-RMST Model				
Coef/$\tau^\star$	0.032	0.070	0.081	0.091
95% CI	(-0.004, 0.069)	(0.028, 0.112)	(0.037, 0.126)	(0.047, 0.136)
P-value	0.084	0.001	<0.001	<0.001

* Among those experiencing an exacerbation during the τ follow-up duration, fold change is the ratio of estimated exacerbation-free time when comparing those taking versus not taking azithromycin, assuming all other predictors in the model are zero. Age is centered at 65 years and percent of predicted FEV_1 is centered at 40% to aid in interpreting fold changes.

† Odds ratio for remaining exacerbation-free at time τ comparing those taking versus not taking azithromycin, adjusted for age, gender, percent of predicted FEV_1, smoking status and study site.

* Percentage increase in τ-RMST for those taking versus not taking azithromycin, adjusted for other covariates in the model.

Table A.2: Summary of interaction tests between treatment group and age, gender, percent of predicted FEV_1 and current smoking status.

	P-value	**Description**
Azithromycin * Age		
τ-IBR model (MI) (Beta Regression)	0.754	
τ-IBR model (MI) (Logistic Regression)	0.013	Older ->more effect
τ-IBR model (EM) (Beta Regression)	0.644	
τ-IBR model (EM) (Logistic Regression)	0.014	Older ->more effect
τ-RMST Model	0.041	Older ->more effect
Azithromycin * Gender		
τ-IBR model (MI) (Beta Regression)	0.983	
τ-IBR model (MI) (Logistic Regression)	0.266	
τ-IBR model (EM) (Beta Regression)	0.925	
τ-IBR model (EM) (Logistic Regression)	0.245	
τ-RMST Model	0.500	
Azithromycin * FEV_1		
τ-IBR model (MI) (Beta Regression)	0.125	
τ-IBR model (MI) (Logistic Regression)	0.400	
τ-IBR model (EM) (Beta Regression)	0.161	
τ-IBR model (EM) (Logistic Regression)	0.379	
τ-RMST Model	0.073	Higher FEV_1 ->more effect
Azithromycin * Current smoking status		
τ-IBR model (MI) (Beta Regression)	0.205	
τ-IBR model (MI) (Logistic Regression)	0.089	Current smokers ->less effect
τ-IBR model (EM) (Beta Regression)	0.197	
τ-IBR model (EM) (Logistic Regression)	0.083	Current smokers ->less effect
τ-RMST Model	0.030	Current smokers ->less effect

APPENDIX B

Supplementary Materials for Chapter 3

B.1 Computation of the estimate of the variance of τ-RMST

We calculate the estimate of the variance of restricted mean survival time of each subject i as follows:

$$\begin{aligned}\mathrm{Var}\{\hat{\mathrm{E}}[T_i(t)]\} &= \tau^2 \mathrm{Var}\big[\hat{\mu}_i(t)[1-\hat{\pi}_i(t)] + \hat{\pi}_i(t)\big] \\ &= \tau^2 \big\{\mathrm{E}\big\{\mathrm{Var}\big[\big(\hat{\mu}_i(t)[1-\hat{\pi}_i(t)] + \hat{\pi}_i(t)\big)|\hat{\mu}_i(t)\big]\big\} \\ &+ \tau^2 \mathrm{Var}\big\{\mathrm{E}\big[\big(\hat{\mu}_i(t)[1-\hat{\pi}_i(t)] + \hat{\pi}_i(t)\big)|\hat{\mu}_i(t)\big]\big\} \\ &= \tau^2 \mathrm{E}\big\{[1-\hat{\mu}_i(t)]^2 \mathrm{Var}[\hat{\pi}_i(t)]\big\} + \tau^2 \mathrm{Var}\big[\hat{\mu}_i(t) + [1-\hat{\mu}_i(t)]\mathrm{E}[\hat{\pi}_i(t)]\big],\end{aligned}$$

which is asymptotically equivalent to

$$\begin{aligned}&= \tau^2 [1-\mu_i(t)]^2 \mathrm{Var}[\hat{\pi}_i(t)] + \tau^2 \mathrm{Var}\big\{\hat{\mu}_i(t) + [1-\hat{\mu}_i(t)]\pi_i(t)\big\} \\ &= \tau^2 [1-\mu_i(t)]^2 \mathrm{Var}[\hat{\pi}_i(t)] + \tau^2 \mathrm{Var}\big\{[1-\pi_i(t)]\hat{\mu}_i(t) + \pi_i(t)]\big\} \\ &= \tau^2 [1-\mu_i(t)]^2 \mathrm{Var}[\hat{\pi}_i(t)] + \tau^2 [1-\pi_i(t)]^2 \mathrm{Var}[\hat{\mu}_i(t)],\end{aligned}$$

so that the asymptotic estimate of the variance of restricted mean survival time for subject i becomes:

$$\widehat{\mathrm{Var}}\{\hat{\mathrm{E}}[(T_i(t)]\} = \tau^2 [1-\hat{\mu}_i(t)]^2 \widehat{\mathrm{Var}}[\hat{\pi}_i(t)] + \tau^2 [1-\hat{\pi}_i(t)]^2 \widehat{\mathrm{Var}}[\hat{\mu}_i(t)]. \tag{B.1}$$

Based on model (3.5) and model (3.6) and an application of the delta method, the first term of equation (B.1) becomes:

$$\begin{aligned}\tau^2[1-\hat{\mu}_i(t)]^2 \widehat{\mathrm{Var}}[\hat{\pi}_i(t)] &= \tau^2 \left[1 - \frac{1}{1+e^{-\hat{\alpha}^T Z_i^\mu(t)}}\right]^2 \widehat{\mathrm{Var}}\left[\frac{1}{1+e^{-\hat{\beta}^T Z_i^\pi(t)}}\right] \\ &= \tau^2 \left[1 - \frac{1}{1+e^{-\hat{\alpha}^T Z_i^\mu(t)}}\right]^2 Z_i^\pi(t)^T \hat{V}_\beta Z_i^\pi(t) \frac{[e^{-\hat{\beta}^T Z_i^\pi(t)^T}]^2}{[1+e^{-\hat{\beta}^T Z_i^\pi(t)}]^4},\end{aligned}$$

and the second term of equation (B.1) becomes:

$$\begin{aligned}\tau^2[1-\hat{\pi}_i(t)]^2\widehat{\text{Var}}[\hat{\mu}_i(t)] &= \tau^2\left[1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}(t)}}\right]^2\widehat{\text{Var}}\left[\frac{1}{1+e^{-\hat{\alpha}^T Z_i^{\mu}(t)}}\right]\\ &= \tau^2\left[1-\frac{1}{1+e^{-\hat{\beta}^T Z_i^{\pi}(t)}}\right]^2 Z_i^{\mu T}(t)\hat{V}_\alpha Z_i^{\mu}(t)\frac{[e^{-\hat{\alpha}^T Z_i^{\mu}(t)}]^2}{[1+e^{-\hat{\alpha}^T Z_i^{\mu}(t)}]^4}.\end{aligned}$$

where $Z_i^{\mu T} = [1, Z_{\mu i}(t)^T]$, $Z_i^{\pi T} = [1, Z_{\pi i}(t)^T]$, and $\hat{V}_\beta$ and $\hat{V}_\alpha$ are the estimates of covariance matrix of coefficients in model (3.5) and model (3.6) respectively.

B.2 Calibrate user-specified correlation to the correlation of the multivariate normal random variables used in the algorithm

To perform this algorithm proposed by Emrich and Piedmonte[10], we should obtain the correlation matrix R_{Ni} of the multivariate standard normal distribution. By definition, the correlation coefficient $\rho_{t_a t_b}$ for any pair of correlated binary variables, $B_i(t_a)$ and $B_i(t_b)$, where $t_a, t_b \in \{0, 30, 60, 90\}$, can be calculated as:

$$\begin{aligned}\rho_{t_a t_b} &= \text{E}\{[B_i(t_a)-\pi_i(t_a)][B_i(t_b)-\pi_i(t_b)]\}/\sqrt{\text{Var}[B_i(t_a)]\text{Var}[B_i(t_b)]}\\ &= [P_{t_a t_b}-\pi_i(t_a)\pi_i(t_b)]/\sqrt{\pi_i(t_a)[1-\pi_i(t_a)]\pi_i(t_b)[1-\pi_i(t_b)]}.\end{aligned} \tag{B.2}$$

The equation (B.2) can be rewritten as:

$$\text{P}_{t_a t_b} = \rho_{t_a t_b}\sqrt{\pi_i(t_a)[1-\pi_i(t_a)]\pi_i(t_b)[1-\pi_i(t_b)]}+\pi_i(t_a)\pi_i(t_b) \tag{B.3}$$

where $P_{t_a t_b} = \text{Pr}[B_i(t_a) = 1, B_i(t_b) = 1]$. Because $B_i(t_a)$ and $B_i(t_b)$ are transformed from two standard normal variables $N_i(t_a)$ and $N_i(t_b)$, $P_{t_a t_b}$ can also be written as:

$$P_{t_a t_b} = \text{Pr}[N_i(t_a) > a(t_a), N_i(t_b) > a(t_b)] = \int_{a(t_a)}^{\infty}\int_{a(t_b)}^{\infty}\phi[n_i(t_a), n_i(t_b), \tilde{\rho}_{t_a t_b}]dn_i(t_a)dn_i(t_b), \tag{B.4}$$

where $\phi[n_i(t_a), n_i(t_b), \tilde{\rho}_{t_a t_b}]$ is the probability density function of the bivariate standard normal distribution with correlation coefficient $\tilde{\rho}_{t_a t_b}$. Thus, with $\pi_i(t_a)$, $\pi_i(t_b)$ and the desired correlation coefficient $\rho_{t_a t_b}$, we can now solve for $P_{t_a t_b}$ by equation (B.3) and then solve for the correlation coefficient $\tilde{\rho}_{t_a t_b}$ in R_{N_i} for any pair of correlated standard normal variables by equation (B.4).

APPENDIX C

Supplementary Materials for Chapter 4

C.1 Supplementary Table of Example Section 4.6

In this section, we show the results of the Cox model for dependent censoring time with time-dependent LAS values as well as other factors previously identified by the SRTR as associated with time-to-transplant: gender, race, height, blood type and time-dependent active versus inactive waitlist status. Corresponding IPCW values were calculated from this model as described in Section 4.4.

Table C.1: Cox dependent censoring model: estimated hazard ratios, 95% confidence intervals and p-values for n=10396 lung waitlist candidates.

	Hazard ratio	95% CI	P-value
Time-independent characteristics			
Female (vs Male)	0.76	(0.71, 0.80)	<0.001
Black (vs White)	0.78	(0.73, 0.84)	<0.001
Other (vs White)	0.85	(0.74, 0.97)	0.020
Height: <5’3” (versus >5’9’)	0.62	(0.57, 0.67)	<0.001
Height: 5’3”-5’6” (versus >5’9”)	0.76	(0.71, 0.82)	<0.001
Height: 5’6”-5’9” (versus >5’9”)	0.87	(0.82, 0.92)	<0.001
Blood type: B (versus A)	1.00	(0.93, 1.07)	0.997
Blood type: O (versus A)	0.94	(0.90, 0.98)	0.003
Blood type: AB (versus A)	1.16	(1.04, 1.30)	0.008
Time-dependent characteristics			
LAS>0 (versus LAS= 0)	138	(1.24, 15415)	0.041
Unit increase in LAS: 0<LAS≤30	0.84	(0.71, 0.98)	0.027
Unit increase in LAS: 30<LAS≤35	1.12	(1.09, 1.15)	<0.001
Unit increase in LAS: 35<LAS≤40	1.10	(1.09, 1.12)	<0.001
Unit increase in LAS: 40<LAS≤60	1.04	(1.04, 1.05)	<0.001
Unit increase in LAS: LAS>60	1.03	(1.03, 1.03)	<0.001
Active vs inactive status	3.72	(3.28, 4.21)	<0.001

C.2 Supplementary Figure of Example Section 4.6

The impact of adjusting for dependent censoring via IPCW values in transplant urgency calculations is substantial, as can be seen in this figure, which displays unadjusted (Kaplan-Meier) versus IPCW-adjusted waitlist survival curve estimates. The IPCW-adjusted waitlist survival curve in this figure is used to implement the MI algorithm described in Section 4.4.

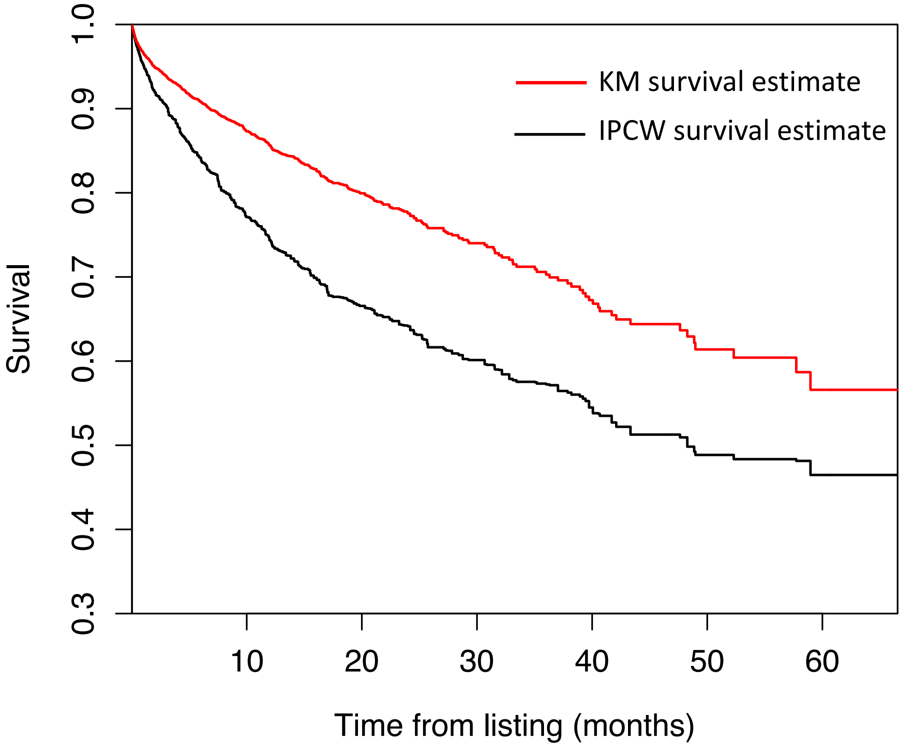

Figure C.1: The unadjusted (Kaplan-Meier) versus IPCW-adjusted waitlist survival curve estimates.

BIBLIOGRAPHY

[1] Hatice Tul Kubra Akdur. Gee-based bell model for longitudinal count outcomes. *Communications in Statistics-Theory and Methods*, pages 1–15, 2022.

[2] Richard K Albert, John Connett, William C Bailey, Richard Casaburi, J Allen D Cooper Jr, Gerard J Criner, Jeffrey L Curtis, Mark T Dransfield, MeiLan K Han, Stephen C Lazarus, et al. Azithromycin for prevention of exacerbations of copd. *New England Journal of Medicine*, 365(8):689–698, 2011.

[3] Per Kragh Andersen and Richard D Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

[4] Per Kragh Andersen, Mette Gerster Hansen, and John P Klein. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime data analysis*, 10(4):335–350, 2004.

[5] James R Anderson, Kevin C Cain, Richard D Gelber, et al. Analysis of survival by tumor response. *J Clin Oncol*, 1(11):710–719, 1983.

[6] Adin-Cristian Andrei and Susan Murray. Regression models for the mean of the quality-of-life-adjusted restricted survival time using pseudo-observations. *Biometrics*, 63(2):398–404, 2007.

[7] Joseph Berkson and Robert P Gage. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515, 1952.

[8] John W Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society: Series B (Methodological)*, 11(1):15–44, 1949.

[9] Thomas M Egan, S Murray, Rami T Bustami, Tempie H Shearon, Keith P McCullough, LB Edwards, MA Coke, ER Garrity, Stuart C Sweet, DA Heiney, et al. Development of the new lung allocation system in the united states. *American Journal of Transplantation*, 6(5):1212–1227, 2006.

[10] Lawrence J Emrich and Marion R Piedmonte. A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304, 1991.

[11] Vern T Farewell. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, pages 1041–1046, 1982.

[12] Cheryl L Faucett, Nathaniel Schenker, and Jeremy MG Taylor. Survival analysis using auxiliary variables via multiple imputation, with application to aids clinical trial data. *Biometrics*, 58(1):37–47, 2002.

[13] Dianne M Finkelstein and David A Schoenfeld. Analysing survival in the presence of an auxiliary variable. *Statistics in medicine*, 13(17):1747–1754, 1994.

[14] David M Gay. Usage summary for selected optimization routines. *Computing science technical report*, 153:1–21, 1990.

[15] Qi Gong and Douglas E Schaubel. Partly conditional estimation of the effect of a time-dependent factor in the presence of dependent censoring. *Biometrics*, 69(2):338–347, 2013.

[16] William H Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. 1994.

[17] Joel B Greenhouse and Robert A Wolfe. A competing risks derivation of a mixture model for the analysis of survival data. *Communications in Statistics-Theory and Methods*, 13(25):3133–3154, 1984.

[18] Daniel B Hall and Zhengang Zhang. Marginal models for zero inflated clustered data. *Statistical modelling*, 4(3):161–180, 2004.

[19] Daniel F Heitjan and Roderick JA Little. Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 40(1):13–29, 1991.

[20] Chiu-Hsieh Hsu, Jeremy M. G. Taylor, Susan Murray, and Daniel Commenges. Survival analysis using auxiliary variables via non-parametric multiple imputation. *Statistics in Medicine*, 25(20):3503–3517, 2006.

[21] Chiu-Hsieh Hsu and Jeremy MG Taylor. Nonparametric comparison of two survival functions with dependent censoring via nonparametric multiple imputation. *Statistics in medicine*, 28(3):462–475, 2009.

[22] Theodore Karrison. Restricted mean life with adjustment for covariates. *Journal of the American Statistical Association*, 82(400):1169–1176, 1987.

[23] John P Klein, Mette Gerster, Per Kragh Andersen, Sergey Tarima, and Maja Pohar Perme. Sas and r functions to compute pseudo-values for censored data regression. *Computer methods and programs in biomedicine*, 89(3):289–300, 2008.

[24] Maiying Kong, Sheng Xu, Steven M Levy, and Somnath Datta. Gee type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational statistics & data analysis*, 85:54–66, 2015.

[25] Jerald F Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225, 1987.

[26] Jerald F Lawless and Claude Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168, 1995.

[27] Chin-Shang Li and Jeremy MG Taylor. A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21(21):3235–3247, 2002.

[28] Danyu Y Lin, Lee-Jen Wei, I Yang, and Zhiliang Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[29] DY Lin, W Sun, and Zhiliang Ying. Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, 86(1):59–70, 1999.

[30] Lyrica Xiaohong Liu, Susan Murray, and Alex Tsodikov. Multiple imputation based on restricted mean model for censored data. *Statistics in medicine*, 30(12):1339–1350, 2011.

[31] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, 1982.

[32] Todd Mackenzie and Michal Abrahamowicz. Using categorical markers as auxiliary variables in log-rank tests and hazard ratio estimation. *Canadian Journal of Statistics*, 33(2):201–219, 2005.

[33] Lee S McDaniel, Nicholas C Henderson, and Paul J Rathouz. Fast pure r implementation of gee: application of the matrix package. *The R journal*, 5(1):181, 2013.

[34] Susan Murray and Anastasios A Tsiatis. Nonparametric survival estimation using prognostic longitudinal covariates. *Biometrics*, pages 137–151, 1996.

[35] Susan Murray and Anastasios A Tsiatis. Using auxiliary time-dependent covariates to recover information in nonparametric testing with censored data. *Lifetime data analysis*, 7(2):125–141, 2001.

[36] MA Nicolaie, JC Van Houwelingen, TM de Witte, and H Putter. Dynamic pseudo-observations: a robust approach to dynamic prediction in competing risks. *Biometrics*, 69(4):1043–1052, 2013.

[37] Raydonal Ospina and Silvia LP Ferrari. Inflated beta distributions. *Statistical Papers*, 51(1):111, 2010.

[38] Raydonal Ospina and Silvia LP Ferrari. A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, 56(6):1609–1623, 2012.

[39] Layla Parast and Beth Ann Griffin. Landmark estimation of survival and treatment effects in observational studies. *Lifetime data analysis*, 23:161–182, 2017.

[40] Margaret Sullivan Pepe and Jianwen Cai. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American statistical Association*, 88(423):811–820, 1993.

[41] James M Robins. Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association*, volume 24, page 3. San Francisco CA, 1993.

[42] James M Robins and Dianne M Finkelstein. Correcting for noncompliance and dependent censoring in an aids clinical trial with inverse probability of censoring weighted (ipcw) log-rank tests. *Biometrics*, 56(3):779–788, 2000.

[43] James M Robins and Andrea Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS epidemiology: methodological issues*, pages 297–331, 1992.

[44] Ori Rosen, Wenxin Jiang, and Martin A Tanner. Mixtures of marginal models. *Biometrika*, 87(2):391–404, 2000.

[45] Donald B Rubin. Multiple imputation for survey nonresponse. *New York: Wiley*, 1987.

[46] Donald B Rubin. Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489, 1996.

[47] Donald B Rubin and Nathaniel Schenker. Multiple imputation in health-are databases: An overview and some applications. *Statistics in medicine*, 10(4):585–598, 1991.

[48] Issaka Sagara, Roch Giorgi, Ogobara K Doumbo, Renaud Piarroux, and Jean Gaudart. Modelling recurrent events: comparison of statistical models with continuous and discontinuous risk intervals on recurrent malaria episodes data. *Malaria journal*, 13(1):1–9, 2014.

[49] Fatemeh Sarvi, Abbas Moghimbeigi, and Hossein Mahjub. Gee-based zero-inflated generalized poisson model for clustered over or under-dispersed count data. *Journal of Statistical Computation and Simulation*, 89(14):2711–2732, 2019.

[50] Glen A Satten and Somnath Datta. The su algorithm for missing data problems. *Computational Statistics*, 15(2):243–277, 2000.

[51] Glen A Satten and Somnath Datta. The kaplan–meier estimator as an inverse-probability-of-censoring weighted average. *The American Statistician*, 55(3):207–210, 2001.

[52] Daniel Scharfstein, James M Robins, Wesley Eddings, and Andrea Rotnitzky. Inference in randomized studies with informative censoring and discrete time-to-event endpoints. *Biometrics*, 57(2):404–413, 2001.

[53] Haolun Shi and Guosheng Yin. Landmark cure rate models with time-dependent covariates. *Statistical methods in medical research*, 26(5):2042–2054, 2017.

[54] Xu Shu and Douglas E Schaubel. Semiparametric methods to contrast gap time survival functions: Application to repeat kidney transplantation. *Biometrics*, 72(2):525–534, 2016.

[55] Judy P Sy and Jeremy MG Taylor. Estimation in a cox proportional hazards cure model. *Biometrics*, 56(1):227–236, 2000.

[56] Jeremy MG Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, pages 899–907, 1995.

[57] Jeremy MG Taylor, Susan Murray, and Chiu-Hsieh Hsu. Survival estimation and testing via multiple imputation. *Statistics & probability letters*, 58(3):221–232, 2002.

[58] Nabihah Tayob and Susan Murray. Nonparametric tests of treatment effect based on combined endpoints for mortality and recurrent events. *Biostatistics*, 16(1):73–83, 2015.

[59] Nabihah Tayob and Susan Murray. Nonparametric restricted mean analysis across multiple follow-up intervals. *Statistics & probability letters*, 109:152–158, 2016.

[60] Nabihah Tayob and Susan Murray. Statistical consequences of a successful lung allocation system–recovering information and reducing bias in models for urgency. *Statistics in medicine*, 36(15):2435–2451, 2017.

[61] Lu Tian, Hua Jin, Hajime Uno, Ying Lu, Bo Huang, Keaven M Anderson, and LJ Wei. On the empirical choice of the time window for restricted mean survival time. *Biometrics*, 76(4):1157–1166, 2020.

[62] Lu Tian, Lihui Zhao, and Lee-Jen Wei. Predicting the restricted mean event time with the subject's baseline covariates in survival analysis. *Biostatistics*, 15(2):222–233, 2014.

[63] Alexander Tsodikov. Semi-parametric models of long-and short-term survival: an application to the analysis of breast cancer survival in utah by age and stage. *Statistics in medicine*, 21(6):895–920, 2002.

[64] Hans C Van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

[65] Xin Wang and Douglas E Schaubel. Modeling restricted mean survival time under general censoring mechanisms. *Lifetime data analysis*, 24:176–199, 2018.

[66] YC Wang, L Meyerson, YQ Tang, and N Qian. Statistical methods for the analysis of relapse data in ms clinical trials. *Journal of the neurological sciences*, 285(1-2):206–211, 2009.

[67] Meng Xia and Susan Murray. Commentary on tayob and murray (2014) with a useful update pertaining to study design. *Biostatistics*, 20(3):542–545, 2019.

[68] Meng Xia, Susan Murray, and Nabihah Tayob. Nonparametric group sequential methods for evaluating survival benefit from multiple short-term follow-up windows. *Biometrics*, 75(2):494–505, 2019.

[69] Meng Xia, Susan Murray, and Nabihah Tayob. Nonparametric group sequential methods for recurrent and terminal events from multiple follow-up windows. *Statistics in medicine*, 38(30):5657–5669, 2019.

[70] Meng Xia, Susan Murray, and Nabihah Tayob. Regression analysis of recurrent-event-free time from multiple follow-up windows. *Statistics in medicine*, 39(1):1–15, 2020.

[71] Fang Xiang and Susan Murray. Restricted mean models for transplant benefit and urgency. *Statistics in Medicine*, 31(6):561–576, 2012.

[72] Fang Xiang, Susan Murray, and Xiaohong Liu. Analysis of transplant urgency and benefit via multiple imputation. *Statistics in Medicine*, 33(26):4655–4670, 2014.

[73] Xu Steven Xu, Mahesh Samtani, Min Yuan, and Partha Nandy. Modeling of bounded outcome scores with data on the boundaries: Application to disability assessment for dementia scores in alzheimer's disease. *The AAPS journal*, 16(6):1271–1281, 2014.

[74] Kazuo Yamaguchi. Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in japan. *Journal of the American Statistical Association*, 87(418):284–292, 1992.

[75] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

[76] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.

[77] David M. Zucker. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association*, 93(442):702–709, 1998.