October 2022

MIXTURE MODELS FOR INTERVAL CENSORED OUTCOMES

Yibai Zhao
University of Massachusetts Amherst

MIXTURE MODELS FOR INTERVAL CENSORED OUTCOMES

A Dissertation Presented

by

YIBAI ZHAO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2022

Biostatistics and Epidemiology

MIXTURE MODELS FOR INTERVAL CENSORED OUTCOMES

A Dissertation Presented

by

YIBAI ZHAO

Approved as to style and content by:

Raji Balasubramanian, Chair

Chi Hyun Lee, Member

David Shapiro, Member

Ruth Etzioni, Member

Lisa Chasan-Taber, Chair of the Faculty
Biostatistics and Epidemiology

DEDICATION

This dissertation is dedicated to the people who have supported me throughout my education.

Thanks for making me see this adventure through to the end.

ACKNOWLEDGMENTS

Chapter 1

The NHANES I Epidemiologic Followup Study (NHEFS) used in our analysis is a national longitudinal study that was jointly initiated by the National Center for Health Statistics and the National Institute on Aging in collaboration with other agencies of the Public Health Service.

We thank the participants and the staff of the NHANES I Epidemiologic Followup Study (NHEFS) for their valuable contributions.

This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Health Statistics and the National Institute on Aging. The authors assume full responsibility for analyses and interpretation of these data.

Chapter 2

The data source of our analysis are the Women and Infants Transmission Study (WITS) and Perinatal AIDS Collaborative Transmission Study (PACTS). We gratefully acknowledge the participants and the staff in the WITS and PACTS.

Chapter 3

Data for this study come from Canary Prostate cancer Active Surveillance Study (PASS). We thank all PASS participants for their dedicated contributions. We also thank a large and dedicated team of coordinating center staff, coordinators, lab staff, and physicians who have made this study possible.

ABSTRACT

MIXTURE MODELS FOR INTERVAL CENSORED OUTCOMES

SEPTEMBER 2022

YIBAI ZHAO

M.B., CHINA MEDICAL UNIVERSITY

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Raji Balasubramanian

Silent events such as the first detectable HIV infection, the onset of Type 2 diabetes and prostate cancer progression are often ascertained by diagnostic tests and/or self-reports that are scheduled periodically. In such applications, we only observe the time to the event of interest to lie between the times of last negative and the first positive tests, resulting in interval-censored observations. In addition, in some medical studies, a substantial proportion of participants may experience the events before the study, so-called prevalent cases, or participants may never experience the event, that is regarded as non-susceptible cases (or indolent cancer or long-term survivor). In this dissertation, I develop mixture models for the analysis of heterogeneous survival data subject to interval-censoring.

In Chapter 1, we propose a parametric mixture model for interval censored time to event outcomes, while relaxing the commonly used proportional hazards assumption. The proposed model is applied to data collected in the National Health and Nutrition Examination Survey to evaluate risk factors of Type 2 diabetes.

The second chapter of this dissertation is motivated by a study of the effects of maternal and infant antiretroviral therapy on the sensitivity of DNA PCR diagnostic tests in detecting HIV infection in infants born to HIV-positive mothers. We apply a mixture model to evaluate the association of a set of predictors with an interval-censored time to first detectable DNA PCR test, while accounting for the subset of infants who test positive at birth. The mixture model is applied to data from the Pediatric AIDS Collaborative Transmission Study and the Women and Infants Transmission Study to evaluate the effects of maternal/infant antiretroviral therapy in HIV subtype B infected mother-infant pairs.

Chapter 3 is motivated by a Canary Prostate Active Surveillance Study (PASS) where the time to cancer progression (i.e., biopsy upgrade) is of primary interest. In this paper, we assume a mixture model for progressive and indolent cancers as well as the prevalent cases where the proportional hazards model incorporates the effect of either time-independent or varying covariates on cancer progression and the mixing parameter modelled with logistic regression corresponds to the fraction of indolent cancer. We propose a semiparametric likelihood-based approach to handle interval-censored observations while accounting for the misclassification rates of biopsy. We present simulation studies to investigate the performance of the proposed approach under various settings. The proposed approach is applied to the Canary Prostate Active Surveillance Study to evaluate the effects of factors on the risk of cancer progression and estimate the indolent fraction under a range of sensitivity rates of biopsy.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1

FLEXIBLE, PARAMETRIC MIXTURE MODELS FOR TIME TO EVENT OUTCOMES, WITH INFLATION OF ZEROES AT BASELINE

1.1 Introduction

In longitudinal cohort settings such as the National Health and Nutrition Examination Survey, detecting the onset of a silent event such as onset of Type 2 diabetes is challenging. Since diagnostic tests are scheduled at periodic clinic visits, the exact time of onset of the disease is unknown. If tests are perfect, the time to event of interest is interval censored and known only up to the interval from the last negative and the first positive diagnostic test. A variety of methods have been proposed to analyze interval censored outcomes [42, 20]. We refer the reader to the tutorial by Gomez, G. et al. (2009) for a review of statistical models appropriate for interval censored outcomes [24].

Moreover, in longitudinal cohort settings, there is often a non-ignorable proportion of subjects who have already had the event of interest at study onset or baseline, representing left censored observations. [19] pointed out that the use of a Weibull distribution assumption leads to biased estimates when used to model time to death in studies that included a significant proportion of long-term survivors. Instead, the authors proposed using a mixture model to model two distinct populations, short term versus long term survivors. In an analysis of colon cancer patients that included a proportion of cured participants, [13] applied a parametric mixture relative survival model which combined a logistic function to incorporate the cure proportion and an

exponential or Weibull distribution to model to the survival times for those who died during the study. [50] introduced flexible mixture models, where one part was for zero inputs, the other part was usually a continuous distribution of nonzero values. Models discussed in the paper included hurdle models, zero-inflated models and two-part semi-continuous models. [9] presents a parametric mixture model for undiagnosed prevalent disease and interval censored outcomes with application to data from electronic health records. The authors propose a parametric logistic-Weibull mixture model and assume proportional hazards to incorporate the effect of covariates on the time to event outcome.

In this paper, we implement a mixture Logistic-stratified Weibull model for modeling the effect of covariates on the risk of incident disease, applicable to settings that include undiagnosed prevalent cases as well as right censored/interval-censored outcomes, while relaxing the PH assumption. The paper is organized as follows: In Section 2, we present the proposed mixture model and discuss the estimation of unknown parameters of interest using the Expectation-Maximization (EM) algorithm. In Section 3, we present results from simulation studies to test the performance of the proposed model. We compare relative bias and coverage probability estimates for the proposed model to other competing approaches ([27] [9]). In Section 4, we apply the proposed model to NHANES data to find the risk factors of type 2 diabetes. In Section 5, we conclude with a summary of this paper and directions for future research.

1.2 Model

1.2.1 Notation

For subject $i = 1, 2, \ldots, n$, let T_i denote the time to the event of interest, which is never directly observed. Let c_i be a subject-specific class indicator of $T_i = 0$, where:

$$c_i = \begin{cases} 1 & \text{if } T_i = 0 \\ 0 & \text{if } T_i > 0 \end{cases} \tag{1.1}$$

c_i is unobserved for those subjects for whom the first test at $\tau > 0$ is positive. Let (L_i, R_i) denote the last negative and the first positive test times for the ith subject, respectively. In a dataset of n subjects, we denote the sequence of observed test times $0 = \tau_0 < \tau_1 < \cdots < \tau_m$. Let k_i be subject-specific censoring indicator of whether c_i is observed. If c_i is observed, then $k_i = 1$, $k_i = 0$ otherwise. Let z_i denote a binary covariate for subject i. Let $\mathbf{x}_i$ denote a vector of covariates for subject i.

1.2.2 Likelihood, Assumptions

Let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote disjoint vectors of parameters, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Let $Pr(c = 1) = \pi(\boldsymbol{z}; \boldsymbol{\theta_1})$. The mixture model can be expressed as:

$$\begin{aligned} Pr(T \leq t; \boldsymbol{z}, \boldsymbol{\theta}) &= Pr(T \leq t; \boldsymbol{z}, \boldsymbol{\theta} \mid c = 1) \times Pr(c = 1) \\ &\quad + Pr(T \leq t; \boldsymbol{z}, \boldsymbol{\theta} \mid c = 0) \times Pr(c = 0) \\ &= \pi(\boldsymbol{z}; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{z}; \boldsymbol{\theta}_1))[1 - S(t; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z})]. \end{aligned} \tag{1.2}$$

Let $\boldsymbol{K}_1$ include the set of subjects for whom $\boldsymbol{k}_i = 1$, when c_i is observed. Let $\boldsymbol{K}_0$ denote the set of subjects for whom $\boldsymbol{k}_i = 0$, when c_i is missing at random (MAR).

When $k_i = 1$ and $c_i = 1$, the corresponding log likelihood contribution is $\log\{\pi(\boldsymbol{z}; \boldsymbol{\theta}_1)\}$. When $k_i = 1$ and $c_i = 0$, the corresponding log likelihood contribution is $\log\{(1 - \pi(\boldsymbol{z}; \boldsymbol{\theta}_1))[S(L_i; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z}) - S(R_i; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z})]\}$. When $k_i = 0$, the subject contributes $\log\{\pi(\boldsymbol{z}; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{z}; \boldsymbol{\theta}_1))[S(L_i; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z}) - S(R_i; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z})]\}$ to the observed data log-likelihood. Thus,:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i\in \boldsymbol{K}_1} [c_i \log\{\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)\} + (1-c_i)\log\{(1-\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)) \times \\ &\quad (S(L_i;\boldsymbol{\theta}_2|c_i=0,z_i) - S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i))\}] + \\ &\quad \sum_{i\in \boldsymbol{K}_0} \log[\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1) + \{1-\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)\}\{1-S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i)\}]\end{aligned} \tag{1.3}$$

,where $S(L_i;\boldsymbol{\theta}_2|c_i=0,z_i)$ and $S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i))$ are the survival functions evaluated at times L_i and R_i, respectively.

Parametric assumptions and incorporating covariates:

Suppose the time to event $T \sim Weibull(\gamma,\lambda)$, where γ and λ are the shape and scale parameters of a Weibull distribution. The density function and survival functions of T can be expressed as:

$$\begin{aligned} f(t;\gamma,\lambda) &= \lambda\gamma t^{\gamma-1}\exp(-\lambda t^{\gamma}) \\ S(t;\gamma,\lambda) &= \exp(-\lambda t^{\gamma}) \end{aligned}$$

We further assume that the effects of covariates $\boldsymbol{z}$ on T are incorporated via a proportional hazards(PH) model, where $\boldsymbol{\beta}$ is the corresponding covariate coefficient. The survival function at t can be expressed as:

$$S(t;\gamma,\lambda,\boldsymbol{\beta}) = \exp(-\lambda t^{\gamma}\exp(\boldsymbol{z}^T\boldsymbol{\beta}))$$

To model the effect of covariates on $\pi(\boldsymbol{z};\boldsymbol{\theta}_1)$, we assume the logistic model as follows:

$$\pi(\boldsymbol{z};\boldsymbol{\theta}_1) = \frac{e^{\alpha_0+\alpha_1 z_i}}{1+e^{\alpha_0+\alpha_1 z}}$$

, or equivalently $\text{logit}(\pi(\boldsymbol{z};\boldsymbol{\theta}_1)) = \alpha_0 + \alpha_1 z$.

Then the $\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1)$ and $S(t_i;\boldsymbol{\theta}_2|c_i=0,z_i)$ in Equation (1.3) can be expressed as:

$$\begin{aligned}\pi(\boldsymbol{z}_i;\boldsymbol{\theta}_1) &= \frac{\exp\left(\alpha_0\mathbf{1}+\boldsymbol{z_i}^T\boldsymbol{\alpha_j}\right)}{1+\exp\left(\alpha_0\mathbf{1}+\boldsymbol{z_i}^T\boldsymbol{\alpha_j}\right)}\\ S(t_i;\boldsymbol{\theta}_2|c_i=0,z_i) &= \exp\left(-\lambda t_i^{\gamma}\exp(\boldsymbol{z_i}^T\boldsymbol{\beta})\right)\end{aligned}$$

, where $\boldsymbol{z_i}$ is a $(n\times p)$ matrix of p covariates, $\boldsymbol{\alpha_j}$ is a vector of p logistic regression coefficients and $\boldsymbol{\beta}$ is a vector of p Weibull regression coefficients.

To relax the proportional hazards assumption, we consider the stratified Cox model, where λ_z and γ_z depend on covariate $\boldsymbol{z}$ as shown below:

$$S(t;\boldsymbol{\beta},\gamma_z,\lambda_z|\boldsymbol{x},\boldsymbol{z})=\exp(-\lambda_{\boldsymbol{z}}t^{\gamma_{\boldsymbol{z}}}\exp(\boldsymbol{x}^T\boldsymbol{\beta}))$$

So $S(t_i;\boldsymbol{\theta}_2|c_i=0,z_i)$ in Equation (1.3) can be expressed as:

$$\begin{aligned}S(t_i;\boldsymbol{\theta}_2|c_i=0,z_i=j) &= \exp\left(-\lambda_j t_i^{\gamma_j}+\boldsymbol{x_i}^T\boldsymbol{\beta}\right)\\ S(t_i;\boldsymbol{\theta}_2|c_i=0,z_i=j,\boldsymbol{x_i}) &= \exp\left(-\exp(\log\lambda_j+\gamma_j\log t_i+\boldsymbol{x_i}^T\boldsymbol{\beta})\right)\end{aligned}$$

1.2.3 Estimation for Survival Models

Following the approach in cheung, we propose the following EM algorithm for maximizing the log likelihood in Equation (1.3).

Initialization Set initial values for $\boldsymbol{\theta}^{(0)}=\{\theta_1^{(0)}=(\alpha_0^{(0)},\alpha_1^{(0)}),\{\theta_2^{(0)}=(\beta^{(0)},\gamma^{(0)},\lambda^{(0)})\}$.

$\boldsymbol{\theta}^{(0)}=\{\beta^{(0)},\gamma^{(0)},\lambda^{(0)},\alpha_0^{(0)},\alpha_1^{(0)}\}$.

Iterate E-step and M-step until convergence, that is $|\ell(\boldsymbol{\theta}^{(l+1)})-\ell(\boldsymbol{\theta}^{(l)})|<10e^{-6}$

E-step Conditional on $\theta = \theta^{(l)}$, compute the expected log-likelihood given by

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(l)}) &= \sum_{i=1}^{n}[E(c_i;\boldsymbol{\theta}^{(l)})\log\{\pi(\boldsymbol{x}_i;\boldsymbol{\theta}_1)\} \\ &\quad +(1-E(c_i;\boldsymbol{\theta}^{(l)}))\log\{(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}_1)) \\ &\quad \times(S(L_i;\boldsymbol{\theta}_2|c_i=0,z_i)-S(R_i;\boldsymbol{\theta}_2|c_i=0,z_i))\}] \end{aligned} \tag{1.4}$$

, where

$$E(c_i;\boldsymbol{\theta}^{(l)}) = \begin{cases} c_i & \text{if } i \in \boldsymbol{K}_1 \\ \frac{\pi(\boldsymbol{x}_i;\boldsymbol{\theta}_1^{(l)})}{\pi(\boldsymbol{x}_i;\boldsymbol{\theta}_1^{(l)})+\{1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}_1^{(l)})\}\{1-S(R_i;\boldsymbol{\theta}_2^{(l)}|c_i=0,z_i)\}} & \text{if } i \in \boldsymbol{K}_0 \end{cases}$$

M-step The updated $\boldsymbol{\theta}^{(l+1)}$ are the values of $\boldsymbol{\theta}$ maximizes the expected log-likelihood in E-step.

For simplicity, We use Newton's method to find the optimal estimates, the formula is

$$\theta^{(l)} = \theta^{(l-1)} - \frac{Q'(\theta^{(l-1)})}{Q''(\theta^{(l-1)})}$$

, where $Q'(\theta^{(l-1)})$ and $Q''(\theta^{(l-1)})$ are the first and second derivative of Q evaluated at $\theta^{(l-1)}$.

However, Newton's method does not behave well when $\boldsymbol{\theta}^{(0)}$ are far away from the true value. In order to solve this problem, we find two extreme situation, where all subjects are from $\boldsymbol{K}_0$ and all subjects are from $\boldsymbol{K}_1$. Our initial values are in between this two extremes, so we let $\boldsymbol{\theta}^{(0)}$ to be estimates from case where half of the subjects are from $\boldsymbol{K}_0$ and rest of them are from $\boldsymbol{K}_1$.

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, we can derive the variance of $\hat{\boldsymbol{\theta}}$ using observed fisher information matrix,

$$I_{obs}(\hat{\boldsymbol{\theta}}) = -\ell''(\hat{\boldsymbol{\theta}})$$

The the variance is the diagonal elements of $I_{obs}^{-1}(\hat{\boldsymbol{\theta}})$.

1.3 Simulation

The results presented in this section are averages obtained across 1000 simulated datasets, each including 2500 subjects. For the i^{th} subject, we simulated Z, a binary covariate of interest, with $Pr(Z = 0) = Pr(Z = 1) = 0.5$. As shown below, we assumed the logistic model to incorporate the effect of Z on $\pi(Z) = P(T = 0 \mid Z)$.

$$\pi(Z) = \frac{e^{\alpha_0+\alpha_1 Z}}{1 + e^{\alpha_0+\alpha_1 Z}}$$

For each subject, we simulated the time to event random variable T from a mixture distribution with $\pi(Z)$ probability of $T = 0$ and $1 - \pi(Z)$ probability with T distributed as a Weibull distribution. We simulate X from a normal distribution with mean 0 and variance 1. In addition, to incorporate the effects of Z and X on $T \mid T > 0$ we assume a Cox PH model or a stratified Cox model as defined in Section 1.2:

$$\begin{aligned} h(t \mid Z) &= \lambda\gamma t^{\gamma-1} e^{\beta Z} \\ h(t \mid Z, X) &= \lambda_Z \gamma_Z t^{\gamma_Z - 1} e^{\beta X} \end{aligned}$$

where $h(t \mid Z)$ and $h(t \mid Z, X)$ denotes the hazard function at time t for a subject with covariate Z and X, respectively. The set $\beta = 0.7$ and selected $\alpha_0, \alpha_1, \gamma, \lambda$

to satisfy pre-specified values of π and the failure rate; the later is defined as the $\Pr(T \leq \tau_3)$. π was varied between $(0.1, 0.2, 0.4)$ and failure rate between $(0.3, 0.5, 0.9)$. To simulate test results at pre-specified visit times, we set the vector of test times to be $\tau_0 = 0, \tau_1 = 1, \tau_2 = 4$, and $\tau_3 = 10$. At each test time, the corresponding test results are set to be negative if $\boldsymbol{\tau} \leq T$, positive if $\boldsymbol{\tau} > T$. Finally, for each subject, we simulate an independent binary random variable K_i where $K_i = 1$ with probability 0.5 and 0 otherwise. Those subjects for whom $K_i = 1$ are assumed to be missing the test result at $\tau = 0$. The distribution of T averaged over 1000 simulated datasets is shown within groups defined by Z in Figure (1.1) and Figure (1.2).

We quantified model performance based on coverage probability and relative bias. We estimated coverage probability as the proportion of simulated datasets in which the 95% confidence interval includes the true parameter value. We estimated relative bias as $\frac{E(\hat{\theta})-\theta}{\theta}$.

1.3.1 Assuming proportional hazards

In this setting, we let the time to event T follow a Weibull PH model. The distribution of T averaged over 1000 simulated datasets is shown within groups defined by Z in shown Figure (1.1), where each observed value of T is rounded up to the closest integer (for visualization).

Estimates of the parameters in the mixture Logistic-Weibull PH model are obtained by maximizing the log likelihood in Equation (1.3) and compared to estimates obtained from the Weibull PH model that ignores the prevalent cases at baseline (i.e. subjects for whom $T = 0$). The results from this simulation are shown in Figure (1.3) and Table (1.5). As π increases from 0.1 to 0.4, we can see sharp increases in the relative bias of the estimates of β and λ from the Weibull PH model that ignores the prevalent cases - for example, when the failure rate is 0.5, π is 0.4, relative bias of $\hat{\lambda}$ is 15, while the relative bias of the estimates from the mixture Logistic-Weibull PH

model is approximately zero. In the coverage probability plots in Figure (1.3), we see that the mixture model coverage probabilities are centered around the nominal level of 95%, whereas the model that ignores the prevalent cases has significant under coverage.

1.3.2 Relaxing the proportional hazards assumption

In this scenario, we relax the proportional hazards assumption in the Weibull regression model, so that the baseline hazard depends on the levels of covariate $\boldsymbol{z}$. For strata j, our mixture model can be expressed as

$$P(T \leq t; \boldsymbol{z} = j, \boldsymbol{\theta}) = \pi(\boldsymbol{z}; \boldsymbol{\theta}_1) + (1 - \pi(\boldsymbol{z}; \boldsymbol{\theta}_1))[1 - S(t; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z} = j)]$$

, where $S(t; \boldsymbol{\theta}_2 | \boldsymbol{c} = 0, \boldsymbol{z} = j) = \exp(-\lambda_j \exp(\boldsymbol{x}^T \boldsymbol{\beta}) t^{\gamma_j})$. The distribution of T within groups defined by Z and averaged over 1000 simulated datasets is shown in Figure (1.2), where each observed value of T is rounded up to the closest integer for visualization.

Estimates of the parameters in the mixture Logistic-Weibull stratified model are obtained by maximizing the log likelihood in Equation (1.3) and compared to estimates obtained from the Weibull stratified Cox model implemented in the R package **straweib**[25]. We note that the Weibull stratified Cox model ignores the prevalent cases at baseline (i.e. subjects for whom $T = 0$). The results from this simulation are shown in Figure (1.4) and Table (1.5). The relative bias of the parameter estimates from the mixture model are all close to zero. When failure rate is large (failure rate = 0.9) and π is small ($\pi = 0.1$ or 0.2), all estimates are nearly unbiased in the Weibull stratified Cox model. As π increases from 0.1 to 0.4, the estimates of γ_1 and λ_1 from Weibull stratified Cox model move further away from their true values.

In the coverage probability plot in Figure (1.4), we see all estimates from mixture Logistic-Weibull stratified model are approximately 95% (nominal level). In the

Weibull stratified Cox model that ignores prevalent cases, γ_1 and γ_2 are always overestimated, and λ_1 is underestimated. When failure rate is fixed and as π increases, the coverage probabilities of β and λ_2 decrease. λ_1 in all settings are underestimated with coverage probability around 0%.

1.4 Application

Based on the current National Diabetes Statistics Report, 37.3 million Americans, or 11.3% of the population, had diabetes with type 2 diabetes making up about 90% to 95% of diabetes cases. With incidence of type 2 diabetes increasing worldwide, metabolic diseases represent a major public health burden [46]. Epidemiological investigations into risk factors of incident type 2 diabetes in various populations is an active area of research.

We apply our proposed method to data from the NHANES I (First National Health and Nutrition Examination Survey) Epidemiologic Follow-up Study (NHEFS). Data are publicly available via the CDC at `https://wwwn.cdc.gov/nchs/nhanes/nhefs/default.aspx/`. NHEFS is a national longitudinal study aimed to investigate the relationships between clinical, nutritional, and behavioral factors and subsequent morbidity, mortality, and hospital utilization. A subset of individuals who participated in the original NHANES I survey in the early 1970s are included in NHEFS. Participants of NHEFS cohorts were followed during the periods 1982-1984 (data collected over a two-year period), 1986, 1987, and 1992 with either a personal interview or phone interview. The analysis dataset included 9974 participants, of whom 639 (6.41%) reported being diabetic at baseline. Of those who did not report a diabetes diagnosis at baseline, 634 reported an incident diabetes diagnosis during follow up. In this group, the mean time to diabetes was 6.43 with a range (IQR) of 4 to 9.5 years. The dataset included 6374 women and 3600 men, with a mean (IQR) age at

baseline of 54 (44-68) years, mean (IQR) weight of 157 (136-181) lb and 3431 (34%) who reported hypertension at baseline.

We applied the mixture Weibull stratified model to evaluate risk factors associated with prevalent and incident Type 2 diabetes, including weight (weight $<$ 150 lb, weight in [150, 250) lb and weight $\geq$ 250 lb), hypertension (1 if has hypertension, 0 if not), age (age $<$ 60, age in [60, 80) and age $\geq$ 80) and biological sex (1 for male, 0 for female). We tested the PH assumption for weight in a likelihood ratio test by comparing the mixture stratified Weibull model (full model) to the nested mixture PH Weibull model (reduced model). In the mixture PH Weibull model, we included weight as a categorical covariate in both the logistic and Weibull components. In the mixture stratified Weibull model, we adjusted for weight as a categorical predictor in the logistic component, and fit a stratified Weibull model with weight category as a stratification factor. The likelihood ratio test comparing the full model to the reduced model resulted in a p value of 0.004, indicating a violation of the PH assumption. In Figure (1.5), we see that the hazard ratio estimates from mixture stratified Weibull model change with time during follow up, which is consistent with the violation of the PH assumption in the likelihood ratio test.

We fit the mixture stratified Weibull model including weight, age, hypertension and biological sex as simultaneous predictors in both the logistic and Weibull components. In this dataset, all included participants had self-reports at baseline. As a result, there were no unreported diabetics at baseline ($T = 0$). To simulate the scenario in which a proportion of subjects are unreported prevalent cases at baseline, a random subset of 20% of subjects who reported diabetes at baseline were selected and their baseline self-report was set to be missing. Table (1.5) presents the model results, showing that all predictors in the model are significantly associated with both the rate of prevalent diabetes (at baseline) and with incident diabetes. Figure (1.6) illustrates the effects of each covariate in different weight groups. We see that survival

rates are all above 95% within the first 10 years for weight $\leq$ 150 (lb). However, the survival probabilities drop dramatically in the stratum of participants with weight $\geq$ 250 (lb). Participants who are 55-years-old older or with hyertension are more likely to have type 2 diabetes. Men are less likely to have type 2 diabetes. These results are consistent with the established literature on risk factors of Type 2 diabetes.

1.5 Discussion

Silent events, like incidence of Type 2 diabetes, cannot be observed directly and can be diagnosed only when tests or questionnaires are given. For the analysis of data collected in longitudinal cohorts that include a non-ignorable proportion of participants with a prevalent diagnosis at baseline, we demonstrate that traditional parametric survival models are inappropriate and propose a flexible mixture model that also relaxes the oft used PH assumption. We use EM algorithm to fit our proposed model and obtain estimates by numeric optimization of the log likelihood.

We compare our proposed model with time to event analyses that ignore the participants with a prevalent diagnosis at baseline. We find that when failure rate is large (0.9) and prevalent proportion is small ($\pi = 0.1$), both approaches yield estimates with a relative bias around 0 and with coverage probability around 95%. As the prevalent proportion increases, models that exclude observations with $T = 0$ result in severely biased estimates. Our approach made the strong assumption of a Weibull distribution for $T \mid T > 0$. Alternative solutions that relax this parametric assumption will be useful for general settings.

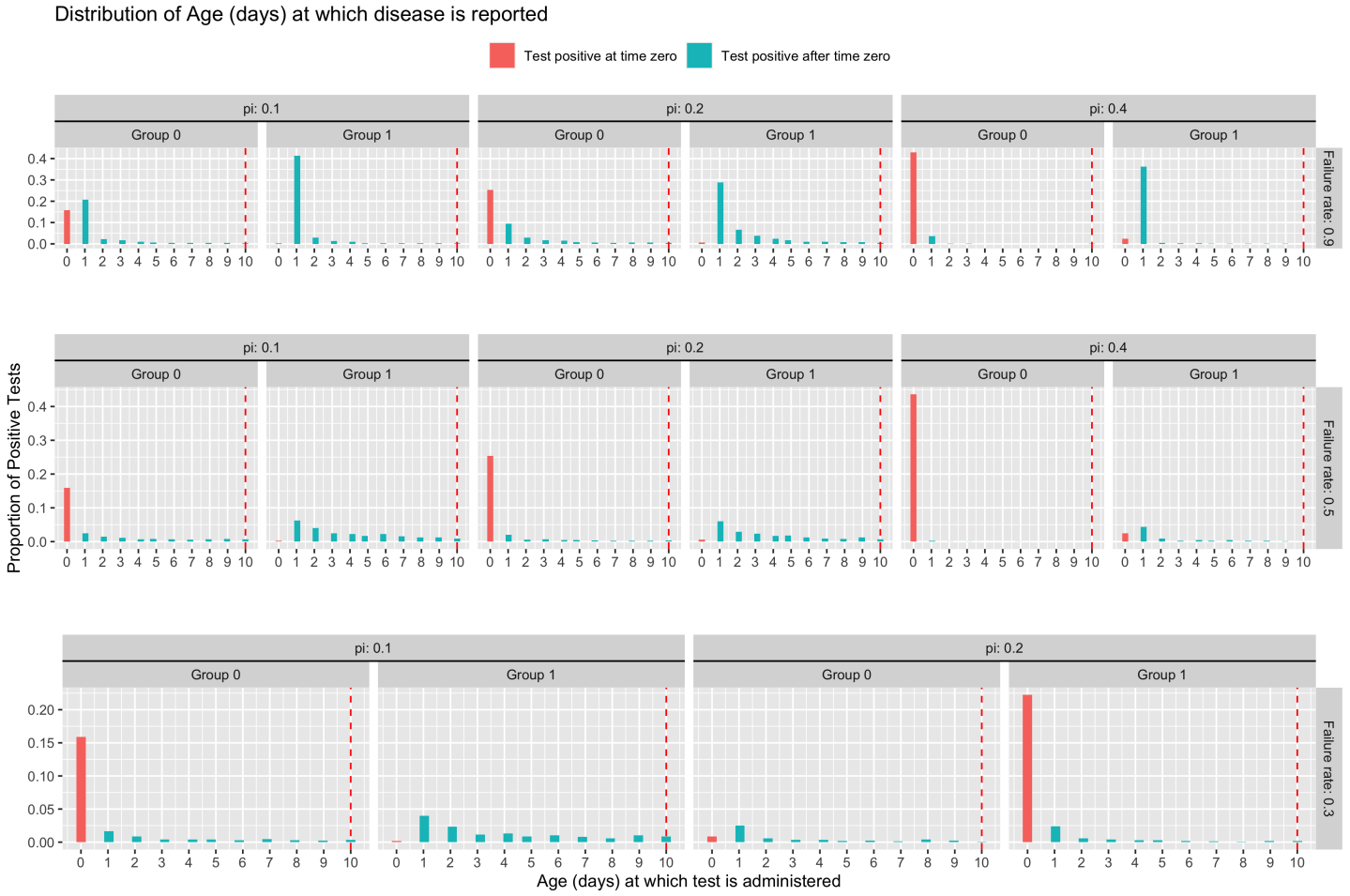
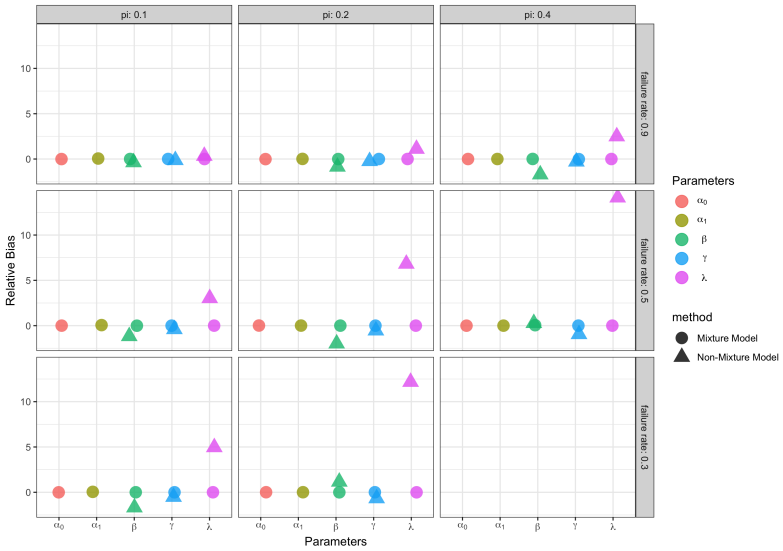


Figure 1.1. Distribution of event times T, when assuming proportional hazards: π is the probability of $T = 0$. Failure rate is defined as the $Pr(T < \tau)$, where τ denotes the end of the study (follow up). *cov* denotes a binary covariate that influences π and $T \mid T > 0$

Figure 1.2. Distribution of event times T, when relaxing proportional hazards: π is the probability of $T = 0$. Failure rate is defined as the $Pr(T < \tau)$, where τ denotes the end of the study (follow up). *cov* denotes a binary covariate that influences π and $T \mid T > 0$



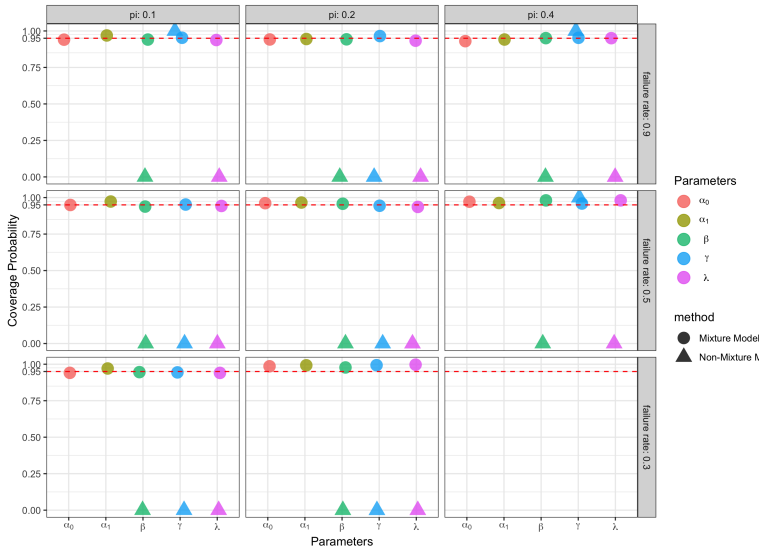
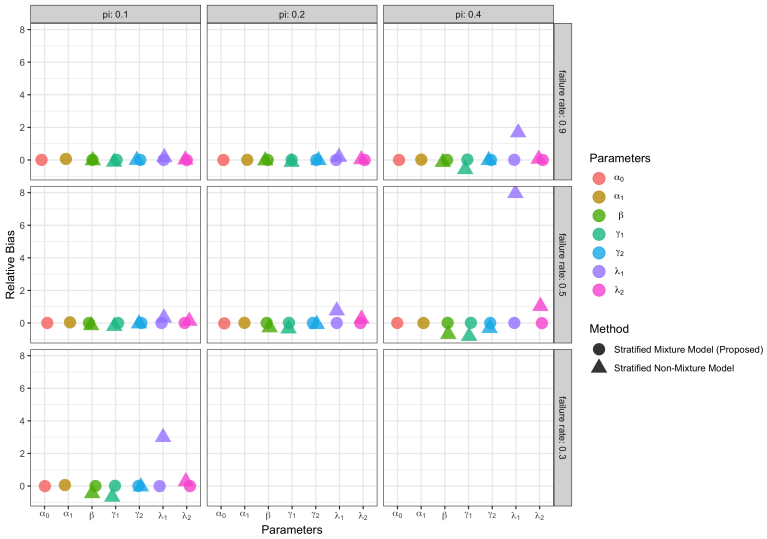


Figure 1.3. Assuming proportion hazards: Simulation Results



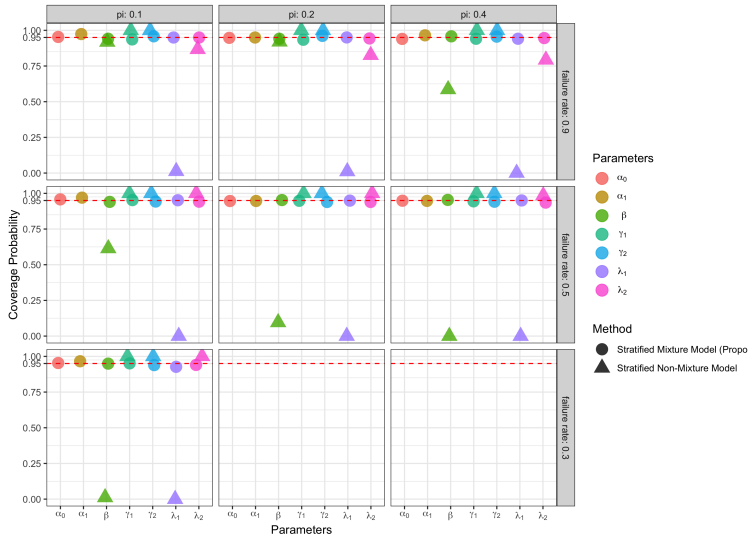


Figure 1.4. Relaxing the proportional hazards assumption: Simulation results

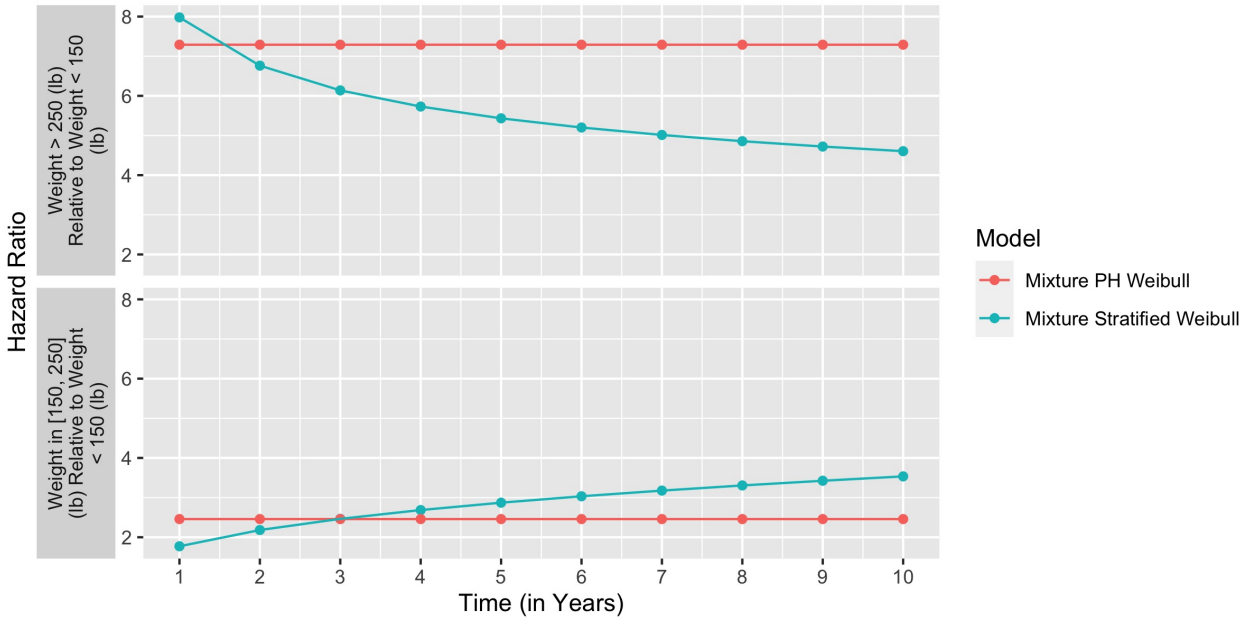


Figure 1.5. NHANES/NHEFS: Plot of hazard ratio by time (in years) with respect to different weight groups. Red dotted line made by mixture PH Weibull model. It's constant over time. Blue dotted line represent hazard ratio from mixture stratified Weibull model. It'll change with time.

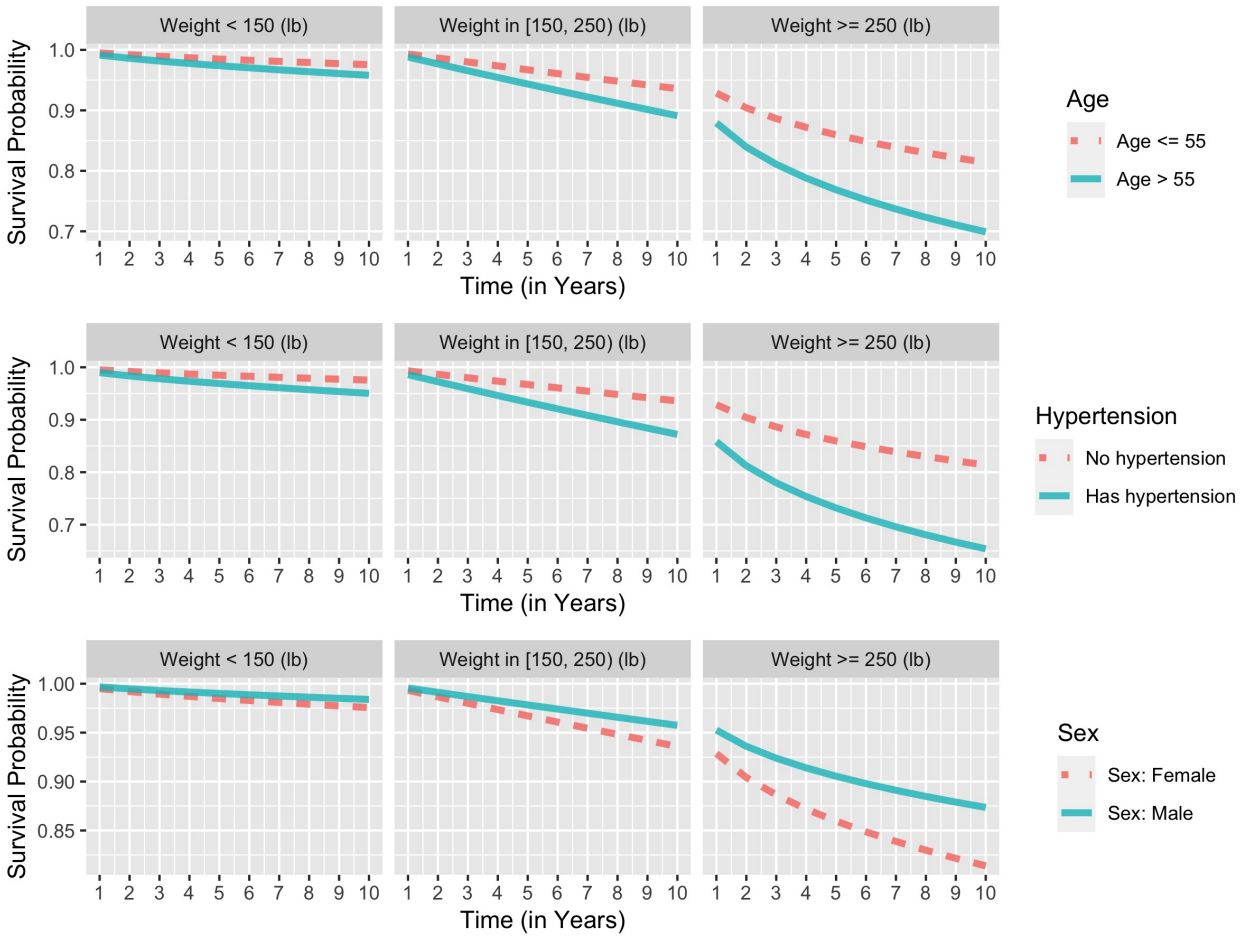


Figure 1.6. NHANES/NHEFS: Proportion of surviving free of type 2 diabetes for each categories. Red dotted line represent baseline for each variable, for example age ≤ 60, no hypertension and female group. Solid lines are for groups other than baseline. Each column show sdifferent weight groups. Three covariates age, hypertension and sex are located in rows.

	Failure Rate = 0.9			Failure Rate = 0.5			Failure Rate = 0.3	
	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$
$\hat{\beta}(sd(\hat{\beta}))$	0.70 (5.16E-2)	0.70 (5.73E-2)	0.71 (1.17E-1)	0.70 (7.75E-2)	0.70 (9.92E-2)	0.74 (3.63E-1)	0.7 (1.07E-1)	0.70 (2.26E-1)
True β	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Coverage Prob. of β (%)	0.94	0.94	0.95	0.94	0.96	0.98	0.95	0.98
$\hat{\gamma}(sd(\hat{\gamma}))$	0.29 (1.33E-2)	0.55 (1.74E-2)	0.03 (5.14E-3)	0.69 (2.71E-2)	0.57 (2.76E-2)	0.25 (2.91E-2)	0.53 (3.01E-2)	0.34 (6.42E-2)
True γ	0.29	0.55	0.03	0.69	0.57	0.25	0.53	0.33
Coverage Prob. of γ (%)	0.95	0.97	0.95	0.95	0.94	0.96	0.94	0.99
$\hat{\lambda}(sd(\hat{\lambda}))$	0.91 (3.94E-2)	0.45 (2.56E-2)	0.71 (7.87E-2)	0.07 (6.50E-3)	0.07 (7.84E-3)	0.05 (1.76E-2)	0.05 (6.00E-3)	0.05 (1.67E-2)
True λ	0.91	0.45	0.71	0.07	0.07	0.05	0.05	0.05
Coverage Prob. of λ (%)	0.94	0.93	0.95	0.94	0.94	0.98	0.94	1.00
$\hat{\alpha}_0(sd(\hat{\alpha}_0))$	-0.90 (8.19E-2)	-0.10 (6.65E-2)	1.60 (9.02E-2)	-0.90 (6.60E-2)	-0.1 (5.88E-2)	1.70 (0.40E-2)	-0.9 (6.50E-2)	-4.23 (8.03E-1)
True α_0	0.9	-0.1	1.6	-0.9	-0.1	1.7	-0.9	-4.2
Coverage Prob. of α_0 (%)	0.94	0.94	0.93	0.95	0.96	0.97	0.94	0.99
$\hat{\alpha}_1(sd(\hat{\alpha}_1))$	-4.22 (2.62)	-4.06 (3.37E-1)	-4.41 (1.95E-1)	-4.25 (3.11)	-4.03 (3.23E-1)	-4.51 (1.87E-1)	-4.19 (2.11)	4.03 (8.20E-1)
True α_1	-4	-4	-4.4	-4	-4	-4.5	-4	4
Coverage Prob. of α_1 (%)	0.97	0.95	0.94	0.97	0.97	0.96	0.97	0.99

Table 1.1. Assuming proportion hazards: Simulation Results obtained by fitting the mixture Weibull PH model

	Failure Rate = 0.9			Failure Rate = 0.5			Failure Rate = 0.3	
	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$
$\hat{\beta}(sd(\hat{\beta}))$	0.44(4.58E-2)	0.10(4.33E-2)	-0.65(4.86E-2)	-0.12(5.68E-2)	-0.68(5.75-2)	0.89(6.80E-2)	-0.48(6.88E-2)	1.51(6.32E-2)
True β	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.7
Cov. Prob.(%)	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00
$\hat{\gamma}(sd(\hat{\gamma}))$	0.25(4.58E-2)	0.42(3.11E-2)	0.23(5.09E-2)	0.41(3.67E-2)	0.26(4.46E-2)	0.07(2.63E-1)	0.26(5.27E-2)	0.11(9.81E-2)
True γ	0.29	0.55	0.31	0.69	0.57	0.25	0.53	0.33
Cov. Prob.(%)	1.00	0.00	99.7	0.00	0.00	1.00	0.00	0.00
$\hat{\lambda}(sd(\hat{\lambda}))$	1.22(3.44E-2)	0.95(3.51E-2)	2.57(3.80E-2)	0.28(5.03E-2)	0.55(4.23E-2)	0.76(4.85E-2)	0.30(5.22E-2)	0.66(5.07E-2)
True λ	0.91	0.45	0.61	0.07	0.07	0.05	0.05	0.05
Cov. Prob.(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Table 1.2. Simulation results when assuming proportional hazards: parameter estimates are obtained by fitting Weibull PH model which ignores prevalent cases at baseline.

	Failure Rate = 0.9			Failure Rate = 0.5			Failure Rate = 0.3
	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$
$\hat{\beta}(sd(\hat{\beta}))$	0.71 (5.76E-2)	0.70 (6.02E-2)	0.70 (6.30E-2)	0.70 (7.09E-2)	0.70 (8.09E-2)	0.71 (1.33E-1)	0.70 (1.10E-1)
True β	0.70	0.70	0.70	0.70	0.70	0.70	0.70
Cov. Prob.(%)	0.94	0.94	0.96	0.95	0.95	0.95	0.95
$\hat{\gamma}_1(sd(\hat{\gamma}_1))$	0.10 (1.73E-2)	0.10 (2.06E-2)	0.10 (3.48E-2)	0.10 (1.22E-2)	0.10 (1.37E-2)	0.30 (5.02E-2)	0.10 (1.98E-2)
True γ_1	0.10	0.10	0.10	0.10	0.10	0.30	0.10
Cov. Prob.(%)	0.94	0.93	0.94	0.95	0.95	0.94	0.95
$\hat{\gamma}_2(sd(\hat{\gamma}_2))$	0.20 (1.55E-2)	0.20 (1.51E-2)	0.30 (1.72E-2)	0.20 (2.10E-2)	0.20 (2.13E-2)	0.10 (1.66E-2)	0.20 (2.11E-2)
True γ_2	0.20	0.20	0.30	0.20	0.20	0.10	0.20
Cov. Prob.(%)	0.96	0.96	0.96	0.94	0.94	0.94	0.94
$\hat{\lambda}_1(sd(\hat{\lambda}_1))$	2.00 (1.07E-1)	2.31 (1.36E-1)	1.21 (1.58E-1)	1.00 (5.91E-2)	0.80 (5.96E-2)	0.20 (3.87E-2)	0.10 (1.37E-2)
True λ_1	2.00	2.30	1.20	1.00	0.80	0.20	0.10
Cov. Prob.(%)	0.95	0.95	0.94	0.95	0.95	0.95	0.93
$\hat{\lambda}_2(sd(\hat{\lambda}_2))$	1.2 (5.50E-2)	1.10 (5.25E-2)	0.90 (4.50E-2)	0.10 (9.00E-3)	0.10 (9.44E-3)	0.10 (1.21E-2)	0.10 (1.05E-2)
True λ_2	1.20	1.10	0.90	0.10	0.10	0.10	0.10
Cov. Prob.(%)	0.95	0.94	0.95	0.94	0.94	0.94	0.94
$\hat{\alpha}_0(sd(\hat{\alpha}_0))$	-0.91 (8.89E-2)	-0.60 (8.33E-2)	2.20 (1.27E-1)	-0.90 (8.64E-2)	-0.10 (7.61E-2)	1.50 (8.34E-2)	-0.90 (7.22E-2)
True α_0	-0.90	-0.60	2.20	-0.90	-0.10	1.50	-0.90
Cov. Prob.(%)	0.95	0.95	0.94	0.96	0.95	0.95	0.95
$\hat{\alpha}_1(sd(\hat{\alpha}_1))$	-4.25 (2.06)	-3.04 (2.66E-1)	-7.15 (9.97E-1)	-4.17 (1.85)	-4.04 (3.23E-1)	-4.00 (1.52E-1)	-4.24 (2.51)
True α_1	-4.00	-3.00	-7.00	-4.00	-4.00	-4.00	-4.00
Cov. Prob.(%)	0.95	0.95	0.97	0.97	0.95	0.95	0.97

Table 1.3. Relaxing the proportion hazards assumption: Simulation Results obtained by fitting a mixture stratified Weibull model.

	Failure Rate = 0.9			Failure Rate = 0.5		
	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.4$
$\hat{\beta}(sd(\hat{\beta}))$	0.68(5.60E-2)	0.68(5.85E-2)	0.61(5.77E-2)	0.60(6.14E-2)	0.50(6.09E-2)	0.41(7.64E-2)
True β	0.70	0.70	0.70	0.70	0.70	0.70
Cov. Prob.(%)	0.92	0.92	0.59	0.61	0.10	0.05
$\hat{\gamma}_1(sd(\hat{\gamma}_1))$	0.09(1.77E-1)	0.09(2.07E-1)	0.04(3.71E-1)	0.08(1.12E-1)	0.07(1.37E-1)	0.03(3.28E-1)
True γ_1	0.10	0.10	0.10	0.10	0.10	0.10
Cov. Prob.(%)	1.00	1.00	1.00	1.00	1.00	1.00
$\hat{\gamma}_2(sd(\hat{\gamma}_2))$	0.20(7.72E-2)	0.20(7.55E-2)	0.29(5.73E-2)	0.19(1.05E-1)	0.18(1.05E-1)	0.20(1.05E-1)
True γ_2	0.20	0.20	0.30	0.20	0.20	0.20
Cov. Prob.(%)	1.00	1.00	1.00	1.00	1.00	1.00
$\hat{\lambda}_1(sd(\hat{\lambda}_1))$	2.32(4.72E-2)	2.68(5.06E-2)	3.22(5.68E-2)	1.32(4.68E-2)	1.41(4.67E-2)	2.92(5.46E-2)
True λ_1	2.00	2.30	1.20	1.00	0.80	0.80
Cov. Prob.(%)	0.01	0.02	0.00	0.00	0.00	0.00
$\hat{\lambda}_2(sd(\hat{\lambda}_2))$	1.22(4.53E-2)	1.13(4.65E-2)	0.95(4.72E-2)	0.11(8.39E-2)	0.13(8.15E-2)	0.12(8.82E-2)
True λ_2	1.20	1.10	0.90	0.10	0.10	0.10
Cov. Prob.(%)	0.87	0.84	0.79	1.00	1.00	1.00

Table 1.4. Relaxing the proportion hazards assumption: Simulation Results obtained by fitting a Weibull stratified Cox model.

Subjects who report type 2 diabetes at their first interview		
Variable	**Odds Ratio (OR)**	**95% CI**
Baseline	1.00	N/A
Weight in [150,250) (lb)	1.6	(1.28, 2)
Weight >250 (lb)	3.36	(1.74, 6.5)
Hypertension	2.71	(2.23, 3.3)
Age >55	2.62	(1.98, 3.46)
Sex (Male)	0.9	(0.73, 1.11)
Subjects who report type 2 diabetes after first interview (for baseline ONLY)		
Variable	**Hazard Ratio (HR)**	**95% CI**
Hypertension	2.06	(1.73, 2.46)
Age >55	1.74	(1.41, 2.14)
Sex (Male)	0.66	(0.54, 0.79)

Table 1.5. NHANES/NHEFS: Covariate coefficients table from mixture Logistic-stratified Weibull model. The first 7 rows are the estimates of subjects who get diabetes at the entry of study. Last 4 rows represent subjects who develop type 2 diabetes after time zero.

CHAPTER 2

TIME TO FIRST POSITIVE DNA-PCR IN HIV-1 INFECTED, NON-BREASTFED INFANTS IN US COHORTS

2.1 Introduction

Prompt antiretroviral therapy (ART) can be lifesaving among HIV-infected infants, and accurate early diagnosis is essential to ensuring early effective treatment intervention [73, 11]. While traditional serologic antibody-detection tests are effective in detecting HIV in adults, they are not valid in infants, since infants can carry passively acquired antibodies from their mothers for more than 9 months after birth [55]. Instead, virologic tests including viral culture, viral antigen (p24), and polymerase chain reaction (PCR) are used as methods for early diagnosis of human immunodeficiency virus (HIV) infection among infants and children [62]. Among these, DNA/RNA PCR tests show a high concordance and can be used as the "gold standard" for the diagnosis of infant HIV infection. Critically, timing and type of ARV exposure of mothers during pregnancy and at the time of labor/delivery can affect the performance of these assays.

Prior to the 1990's, HIV-infected women did not generally receive ARV during pregnancy, and transmission rates to their infants were 15% to 45%. In the early 2000s, most HIV-infected pregnant women received single antiretroviral treatment (usually zidovudine), however the virus can be resistant to a single agent and transmission rates to infants were still around 8% [21, 22]. Combined antiretroviral therapy (cART) was introduced in the 2010s and can prevent viral resistance by combining more than three antiretroviral drugs. Transmission rate to infants can be reduced to

below 2% with cART during the periods of pregnancy, labor and delivery [10] . However, in addition to decreasing transmission rate, cART increases false negative and indeterminant rate of testing in infants, especially soon after birth when the infant may still have residual cART in their system [47, 2, 28, 48, 71].

Various groups have studied the performance of virologic tests for the early diagnosis of HIV-infected infants [36, 14, 15, 55, 39, 52]. The majority of the literature in this area is focused on pediatric populations infected with subtype B virus and who are exposed to at most single antiretroviral regimens given to the mother during pregnancy and/or to the infant as prophylaxis. These previous reports have consistently showed that DNA/RNA PCR tests had a high specificity, meaning a positive test result reliably indicated a true positive infection. However, the sensitivity of those tests was consistently very low at birth and increased after two to four weeks of age. This literature forms the basis for current CDC guidelines that states that exclusion of HIV infection in non-breastfed infants can be based on two or more negative virologic tests with one negative test obtained at age ¿= 1 month and one at age ¿= 4 months. However, these guidelines do not account for the effects of exposure to potent cART that could result in delaying the time to earliest positive DNA PCR test is delayed in infants [56].

In our previous work, we combined data from multiple cohorts to evaluate the association of type and timing of prophylactic maternal and infant antiretroviral regimen with time to first positive HIV-1 DNA PCR test [2]. In this work, our focus was on non-breastfed infants infected with non-B subtype HIV-1 virus. Our results showed that in the subset of infants testing negative at birth, infants exposed to combination ARV had a longer time to DNA PCR test positivity. However, these results were based on a limited sample size in the combination ARV group [2].

In this study, we present results on the sensitivity of DNA PCR tests given to HIV-infected, non-breastfed infants born to HIV-infected mothers from the prospec-

tive Women and Infants Transmission Study (WITS) and the Perinatal AIDS Collaborative Transmission Study (PACTS), where subtype B HIV infections are dominant. We estimated the sensitivity of DNA PCR tests as a function of age at testing, in infants exposed to specific maternal and infant ARV regimens, including infants exposed to cART. Our analyses adjust for potential confounders such as viral load, CD4 count, mode of delivery, and gestational age and birth weight.

2.2 Methods

Cohorts: We included HIV-infected women and their non-breastfed HIV-infected infants from WITS and PACTS cohorts. WITS was a prospective epidemiologic study of the natural history of HIV infection in pregnant women and their infants carried out at obstetric/gynecologic and pediatric clinics in Boston, Chicago, Manhattan, Brooklyn, San Juan, and Houston. There were 788 HIV-infected pregnant women and 657 infants born to them admitted into the study before June 1993 [68]. PACTS was a multicenter, prospective cohort study of HIV-infected pregnant women and their newborns conducted in 4 US cities (New York City, 1986; Baltimore, 1989; Atlanta, 1990; and Newark, 1990). The study monitored the incidence of mother-to-child HIV transmission and described the natural course of pediatric HIV disease progression. It was supported by the CDC from 1986 through 1999 [34].

Inclusion and exclusion: We included infants who were HIV positive or indeterminate and had at least one DNA PCR test before age of 3 months and excluded those whose infant/maternal antiretroviral regimen were not recorded (N = 39). The final dataset included 428 HIV-infected infants (WITS: 129; PACT: 299). 103 infants in WITS and 162 infants in PACTS had complete covariate data (See Figure 2.1) so this subset of 265 infants was used for the adjusted model.

Covariates: Covariates accounted from in the analyses included maternal CD4+ cell count and maternal viral load obtained closest to time of delivery, mode of delivery, infant's gestational age and infant birthweight.

Each covariate was included into statistical models as categorical variables as follows: maternal CD4+ cell count closest to delivery was categorized into 4 levels according to (1) less than 200 cells/ul, (2) 200-350 cells/ul, (3) 350-500 cells/ul and (4) greater than 500 cells/ul; mode of delivery included (1) vaginal, (2) C-section before onset of labor/membrane rupture and (3) C-section after onset of labor/membrane rupture; gestational age was categorized as (1) ¡ 37 weeks and (2) ¿= 37 weeks; birth weight as (1) ¡ 2500 grams and (2) ¿= 2500 grams; maternal viral load closest to delivery categorized as (1) ¡ 400 copies/ml, (2) between 400 and 999 copies/ml, (3) between 1,000 and 9999 copies/ml, (4) between 10,000 and 99,999 copies/ml and (5) greater than or equal to 100,000 copies/ml.

2.2.1 Maternal Antiretroviral Regimen

Infants were grouped according to their mother's most complex antiretroviral regimen during the trimester closest to delivery and at the time of labor/delivery. Maternal antiretroviral regimen was categorized as: no ARV (N=198); Single NRTI referring to Single nucleoside reverse transcriptase inhibitor (N=89); 2-3 NRTIs without sdNVP referring to combination ARV regimen of 2-3 nucleoside reverse transcriptase inhibitors without single dose Nevirapine (N=11); 2-3 NRTIs + sdNVP referring to combination ARV regimen that includes 2-3 NRTIs with single-dose nevirapine (N=8); 3+ ARV referring to combination ARV regimens that included three or more ARVs with non-nucleoside reverse transcriptase inhibitors and with or without protease inhibitors (N=106); sdNVP referring to single-dose nevirapine only (N=6); and ZDV + sdNVP referring to the combination of zidovudine and single-dose nevirapine (N=10). (See details in Table 2.4)

2.2.2 Infant Antiretroviral Regimen

We only considered infant prophylactic regimens initiated prior to 45 days after birth. Infant prophylactic antiretroviral regimen was categorized as: no antiretroviral regimen (No ARV) (N=355); zidovudine (ZDV) (N=70); and Other (N=3). (See details in Table 2.4)

2.2.3 Statistical Analysis

The goal was to estimate the distribution of time to first positive DNA PCR test among all non-breastfed HIV-infected infants exposed to different maternal/infant antiretroviral regimens. For each infant, the time to earliest DNA PCR test positivity is only known to be within the interval from the time of the last negative test and that of the first positive test; to accommodate this uncertainty, we fit models appropriate for interval censored time to event outcomes. To include infants who tested positive at birth, we set their interval of time to earliest test positivity to be between 0 (birth) and 1 day. We applied Weibull proportional hazards (PH) regression to evaluate the association of type of prophylactic maternal and infant antiretroviral regimen with time to first positive HIV DNA PCR test. Due to the concordance of maternal and infant ARV regimens, we considered the effects of maternal and infant ARV regimen in separate models. The goodness of fit of the Weibull assumption was checked by comparing estimates of cumulative test positivity at various ages from the model to estimates from a non-parametric Kaplan Meier procedure.

We verified the validity of the PH assumption by testing the interaction of time and treatment in an expanded model using a likelihood ratio test (LRT). The statistical significance of the effects of maternal and infant ARV regimen were based on LRT obtained by comparing models with and without treatment. Weibull regression models also adjusted for potential confounders including maternal CD4+ cell count, viral load, mode of delivery, infant's gestational age and birth weight.

In a supplemental analysis, we fit a parametric mixture model to account for the subset of infants who tested positive at birth. See Appendix B for details.

2.3 Results

Our analysis included 428 HIV-infected, non-breastfed infants, including 129 infants in WITS and 299 in PACTS. Mothers of 46% (N=198) of infants received no ARV, 21% received Single NRTI (N=89) and 25% received cART (N=106). All other categories of maternal ARV included fewer than 10 infants each. The majority of infants born to mothers who received No ARV or Single NRTI were from PACTS; whereas, the majority of infants whose mothers received cART were from WITS (Appendix Tables B.1 and B.2).

83% (N=355) of infants were not given any prophylactic regimen at birth and 16% (N=70) of infants received zidovudine (ZDV) (Table 2.4). 44% of infants had maternal CD4+ count exceeding 500 cells/ul. 37% of infants had maternal viral load lower than 400 copies/ml. 71% of infants were delivered vaginally. 33% of infants were pre-term (gestational age less than 37 weeks). For 65% of infants, birthweight was greater than 2500 grams. Baseline characteristics by maternal/infant antiretroviral regimen categories can be found separately for WITS and PACTS (Supplemental Tables B.1, B.2, B.3 and B.4) in the Appendix. Maternal and infant characteristics were each not associated with maternal ARV ($p > 0.2$).

In both WITS and PACTS, the majority of DNA PCR tests were given in the first week after birth (Supplemental Figure B.1). In PACTS, there were no tests done after 90 days, while in WITS a few tests were given after 90 days (Supplemental Figure B.1). All available DNA PCR test results in the maternal No ARV group were given prior to 90 days of age. A similar distribution of timing of tests was observed among infants whose mothers received Single NRTI. The majority of DNA PCR tests were given prior to 90 days of age among infants whose mothers received

cART (Supplemental Figure B.2, Supplemental Table B.5). All infants in PACTS had a single DNA PCR test result. The number of DNA PCR tests per infant ranged from 1 to 6 in WITS. The majority of infants whose mothers received No ARV or Single NRTI had only 1 DNA PCR test result available; however, among infants whose mothers received cART, the number of test results available per infant ranged from 1 to 6.

2.3.1 DNA PCR test positivity and maternal ARV exposure

Maternal antiretroviral regimen was significantly associated with time to first positive HIV DNA PCR in a Weibull PH model without confounders (LRT p value $< 10e{-}13$). The test of proportional hazards was not rejected (p=0.06), indicating a lack of evidence for a violation of the PH assumption. The estimated probabilities of a positive HIV DNA PCR test in HIV-infected non-breastfed infants when tested on the first day after birth are significantly lower in infants whose mothers were given cART (5%, 95% CI: 2% - 9%) than in infants whose mothers received no ARV (29%, 95% CI: 22% - 38%) or received Single NRTI (25%, 95% CI: 17% - 35%). The differences in test positivity remained when infants were tested at 90 days of age (Table 2.4; Figure 2.2), with estimated probabilities of a positive DNA PCR test of 20% (95% CI: 12%-32%), 81% (95% CI: 68%-91%) and 74% (95% CI: 56% - 89%) in the cART, No ARV and Single NRTI groups, respectively. For infants whose mothers received other antiretroviral regimens, the estimated probabilities of a positive HIV DNA PCR test by 90 days of age varied between 18% and 80%. However, since sample sizes in these groups were small ($n < 10$), the 95% CIs were wide and overlapping with that for the No ARV group (See Table 2.4 and Figure 2.2). The association between maternal ARV with time to DNA PCR test positivity remained after adjusting for potential confounders including maternal viral load, CD4 count, mode of delivery, infant gestational age and birthweight (LRT p value $< 10e{-}4$). The hazard ratios

of time to test positivity relative to the cART group (reference) were 5.78 (95% CI: 2.61 – 12.82) in the No ARV group and 4.65 (95% CI: 1.95 – 11.07) in the Single NRTI group (Table 2.4). These results indicate that infants whose mothers received no ARV or Single NRTI were much more likely to have positive tests at birth or at earlier times after birth than those whose mothers received cART. A secondary analysis based on a mixture model to account for the infants who test positive at birth (n=13) resulted in similar findings (see Supplement).

2.3.2 DNA PCR test positivity and infant ARV prophylaxis

As our dataset included only 3 infants who received a prophylactic ARV regimen other than ZDV, we did have the ability to evaluate the impact of infant combination ARV therapy on the time to DNA PCR test positivity. When comparing infants who received ZDV to those who received no ARV, we observed no evidence of a delay in the time to DNA PCR test positivity (LRT test p-values of 0.9 and 0.38 in unadjusted and adjusted models, respectively).

2.4 Discussion

Our results show that exposure to cART significantly delays time to the first DNA PCR test positivity in models that adjust for potential confounders.

Infants with HIV infection whose mothers received no treatment or single NRTI are more likely to test positive at birth, compared to those whose mothers received cART. Among infants exposed to No ARV or maternal single NRTI, test sensitivity is low at birth, but increases rapidly thereafter. By 3 months of age, most HIV-infected infants in "no ARV" or single NRTI groups test positive, while for those in the cART group the detectability is much lower. Lastly, there was no evidence of a delay in time to test positivity in infants who received ZDV as prophylaxis following birth when

compared to infants who received no ARV. However, our study was not sufficiently powered to evaluate the effects of infant cART with DNA PCR sensitivity.

Our estimates of DNA PCR sensitivity at various ages are concordant with prior research conducted in cohorts of HIV infected infants without ARV exposure or among those exposed to simpler single NRTI treatments [55, 14, 15, 39, 7, 6, 41, 63, 75]. We find that for those infants whose mothers received no ARV, DNA PCR sensitivity is low at birth and increases dramatically after two weeks. Similar findings are reported in the literature published in the late 1990s. In a study of 56 HIV infected infants born to ARV naive mothers in the Bahamas and Montreal, DNA PCR sensitivity in dried blood spots is only 27% within 4 days of life but rises to 88.9% in 2-weeks and 97.2% by 3 months of age [7]. Another study analyzed data from 271 HIV-infected infants born to mothers with no ARV exposure by aggregating data from 12 different cohorts. The sensitivity of HIV-1 DNA PCR was 40% at birth and rose rapidly to 93% in the second week. By the end of one month, sensitivity was 96% [14]. These findings are largely concordant with our estimates in infants born to mothers with no ARV exposure during pregnancy or at the time of labor/delivery.

Among infants born to mothers exposed to Single NRTI, we observed a similar pattern of low DNA PCR sensitivity during the first week of age followed by a rapid increase in sensitivity afterwards. This pattern was also seen in several other studies. Dunn et al. (2000) [15], included 422 infants infected with HIV subtype B, combining data from four prospective, multi-center studies of HIV positive pregnant women. HIV positive infants included in this study were either not exposed to ARV or exposed to at most to monotherapy with ZDV. This study found that DNA PCR sensitivity at birth was 36% (95% CI: 31% - 41%) and approximately 100% by 1 month of age. Moreover, test positivity was not affected by maternal/infant ZDV exposure. A subset of infants in the Dunn et al. (2000)[15] analysis from PACTS and WITS who were exposed to Single NRTI were also included in our analysis. An-

other smaller study of 24 HIV-infected infants in a pediatric AIDS clinical trial found that the sensitivity of DNA PCR at birth was 10% and exceeded 80% at 6 weeks of age and thereafter16. HIV positive mothers in this trial were randomized either to ZDV combined with HIV-1 hyperimmune globulin (HIVIG) or ZDV combined with immuno-globulin lacking antibody to HIV-1 (IVIG). In a French study that included 65 HIV-infected infants exposed to a variety of maternal ARV regimens including 15 infants exposed to maternal triple ARV, DNA PCR sensitivity during the first week was 55%, 89% at 1 month and 100% by 3-months of age. Although limited by the available sample size, the study found no evidence that the sensitivity of DNA PCR is associated with types of maternal and infant ARVs [6].

A study conducted in Thailand where HIV-1 subtype E is predominant included 98 HIV-infected infants who were exposed to maternal ZDV treatment24. This study found that the detectability of HIV infection by DNA PCR at birth was dependent on the duration of maternal ZDV treatment [63], with longer treatment duration resulting in delays in HIV detection. While our study did not consider duration of maternal ARV treatment, these findings may explain the heterogeneity in the sensitivity estimates between studies that included infants from varying maternal or infant treatment types and durations. Another randomized, placebo-controlled clinical trial of short-course ZDV in Bangkok that included 395 non-breastfed infants born to HIV-infected mothers (91.7% with subtype E, 8.3% with subtype B) found that DNA PCR sensitivity at birth was 38% and reached 100% after 2 months [75]. Among infants whose mothers did not receive any treatment, sensitivity was 35% at birth, while among infants whose mothers received ZDV, sensitivity was 50%. However, these differences in sensitivity by maternal ARV exposure were not statistically significant [75]. The higher sensitivity at birth in the ZDV group relative to the placebo group may be due to the differences in the timing of HIV transmission between the ZDV and placebo arms - a larger proportion of HIV transmissions in the placebo group

during labor and delivery could lead to a lower overall DNA PCR test positivity at birth as intrapartum infections are likely to be detected at later times when compared to in-utero infections. In a previous paper by our group, we evaluated the association of maternal and infant ARV treatment on time to DNA PCR test positivity with a focus on infants infected with HIV-1 non-B subtype virus [2]. In this analysis, 165 infants were exposed to maternal Single NRTI – in this group, the sensitivity of DNA PCR was 85% one day after birth and 91% by 2 weeks of age.

Our analysis found evidence that the time to DNA PCR test positivity is significantly delayed in infants exposed to combination ARV regimens that include 3 or more ARVs that include NNRTIs and/or PIs (cART). Our findings are concordant with reports of studies conducted during the more recent ART-dominated era, although these reports have been in limited sample sizes. A prospective study conducted in South Africa where HIV-1 subtype E is prevalent included 38 HIV-infected infants whose mothers received maternal AZT or highly potent ARVs. A majority of the infants also received prophylaxis such as sdNVP combined with various durations of AZT (92%) or daily-dose NVP (7.9%). The sensitivity of DNA PCR at birth and at 2 weeks were 68.4% and 62%, but rising to 100% by 6 weeks of age []23. Due to a limited sample size (n=38), this study did not assess the sensitivity of DNA PCR by type of maternal ARV exposure. In our prior work in populations infected with non-B subtype virus, we evaluated DNA PCR sensitivity by type of maternal and infant ARV exposure in a combined cohort analysis of 405 perinatally infected infants who were not breastfed [2]. In the subgroup of 143 HIV-infected infants who tested negative at birth including 6 who received combination ARV treatment, we observed a longer time to test positivity among infants who received combination ART when compared to infants who received either no ARV or single NRTI.

Although significantly limited by the sample size in the combination ARV exposure group, the findings in our prior work [2] consistent with our findings here in

PACTS and WITS. To our knowledge, no previous study has been sufficiently powered to evaluate DNA PCR sensitivity with cART exposure; however, prior reports in the literature have presented evidence of cART exposure in infants leading to a substantial reduction in HIV titers to levels below the limit of detection by PCR [61]. A recent report from South Africa also described a case series of 3 HIV-infected infants with varying ART exposures in whom DNA PCR tests given after treatment initiation resulted in repeated false negative and indeterminate results, highlighting the challenge of diagnosis of HIV infection in the backdrop of exposure to potent ARV regimens [47].

Our research provides evidence that the detectability of DNA PCR positivity could vary by type of maternal ARV exposure, so that the definitive exclusion of HIV infection should also take this important characteristic into account. Our study has several strengths. To our knowledge, this is one of the largest studies of DNA PCR test performance among HIV-infected, non-breastfed infants who were exposed to highly active combination ARV regimens. We restricted our analysis to infants who were not breast fed to limit the duration of HIV exposure to the in-utero and intrapartum periods. While we restricted our analysis to infants who did not breast feed, our results equally apply to breast fed populations. By combining data from well characterized prospective cohorts in the US, we were able to disentangle the effects of varying types of maternal ARV treatments on the time to test positivity and to adjust for potential confounders of this association.

Our study has several limitations. First, our dataset did not include a large enough sample size of infants who received combination ARV therapies as prophylaxis, thus limiting our evaluation of the effects of infant prophylaxis. Second, we did not have complete information on timing and duration of antiretroviral treatments to allow a more in-depth exploration of the duration and type of ARV exposure on DNA PCR sensitivity. Third, our analysis is likely to be impacted by varying sample

collection/processing and differences in the types of assays used across studies and sites.

Our work provides evidence of a significant delay in DNA PCR test positivity in HIV-infected infants who are exposed to potent combination ARV therapies. Our findings have implications on the recommended schedule for HIV testing for infants born to HIV-positive mothers, especially in the current era when cART regimens are common. The focus of our study was on HIV-infected infants in the US where subtype B HIV-1 infections are prevalent. Future work on non-B subtype populations and in breast-fed populations is needed to further inform infant HIV testing guidelines.

Table 2.1. Infants classified by maternal antiretroviral regimen and infant antiretroviral regimen (N=428).

Maternal antiretroviral regimen	**Infant antiretroviral regimen**		
	None	ZDV	Other
No ARV[1]	193	5	0
Single NRTI[2]	54	35	0
sdNVP[3]+ ZDV[4]	8	2	0
sdNVP[3]only	5	1	0
2-3 NRTIs[2]without sdNVP[3]	9	2	0
2-3 NRTIs[2]with sdNVP[3]	6	2	0
cART[5]	80	23	3

[1] ARV: antiretroviral
[2] NRTI: nucleoside reverse transcriptase inhibitors
[3] sdNVP: single dose Nevirapine
[4] ZDV: Zidovudine
[5] cART: combination anti-retroviral therapy, defined as including Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) and/or Protease Inhibitors (PI)

Table 2.2. Probabilities of a positive HIV DNA PCR test [95% confidence interval] among HIV-infected non-breastfed infants by 1,14,30, 42 and 90 days after birth, according to type of maternal antiretroviral regimen. Results from a unadjusted Weibull PH model.

Maternal ARV	Day 1	Day 14	Day 30	Day 42	Day 90
No ARV[1] (N=198)	0.29[0.22,0.38]	0.58[0.5,0.66]	0.68[0.58,0.77]	0.72[0.61,0.82]	0.81[0.68,0.91]
Single NRTI[2] (N=89)	0.25[0.17,0.35]	0.51[0.39,0.65]	0.6[0.46,0.76]	0.65[0.49,0.8]	0.74[0.56,0.89]
sdNVP[3]+ ZDV[4] (N=10)	0.11[0.03,0.32]	0.25[0.09,0.59]	0.31[0.11,0.69]	0.34[0.13,0.73]	0.42[0.16,0.82]
sdNVP[3]only (N=6)	0.28[0.1,0.66]	0.57[0.23,0.93]	0.66[0.29,0.97]	0.7[0.32,0.98]	0.8[0.39,0.99]
2-3 NRTIs[2]without sdNVP[3] (N=11)	0.04[0.01,0.26]	0.1[0.01,0.52]	0.13[0.02,0.62]	0.14[0.02,0.66]	0.18[0.03,0.76]
2-3 NRTIs[2]with sdNVP[3] (N=8)	0.09[0.02,0.34]	0.21[0.06,0.62]	0.27[0.08,0.72]	0.3[0.08,0.76]	0.37[0.11,0.84]
cART[5] (N=106)	0.05[0.02,0.09]	0.11[0.06,0.18]	0.14[0.08,0.23]	0.16[0.09,0.25]	0.2[0.12,0.32]

[1] ARV: antiretroviral
[2] NRTI: nucleoside reverse transcriptase inhibitors
[3] sdNVP: single dose Nevirapine
[4] ZDV: Zidovudine
[5] cART: combination anti-retroviral therapy, defined as including Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) and/or Protease Inhibitors (PI)

Table 2.3. Hazard ratios of time to DNA PCR test positivity by type of maternal antiretroviral regimen from unadjusted and adjusted Weibull PH models.

Maternal ARV	**Unadjusted Model (N=428) Hazard Ratio [95% CI]**	**Adjusted Model* (N=265) Hazard Ratio [95% CI]**
No ARV	7.49 [4.06-13.81]	5.78 [2.61-12.82]
Single NRTI	6.18 [3.09-12.38]	4.65 [1.95-11.07]
ZDV + sdNVP	2.47 [0.7-8.7]	2.49 [0.67-9.29]
sdNVP only	7.23 [2.04-25.64]	3.5 [0.84-14.52]
2-3 NRTIs without sdNVP	0.9 [0.12-6.89]	1.76 [0.21-14.74]
2-3 NRTIs with sdNVP	2.09 [0.47-9.3]	1.96 [0.37-10.49]
cART	1 (Reference)	1 (Reference)

[1] ARV: antiretroviral
[2] NRTI: nucleoside reverse transcriptase inhibitors
[3] sdNVP: single dose Nevirapine
[4] ZDV: Zidovudine
[5] cART: combination anti-retroviral therapy, defined as including Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTI) and/or Protease Inhibitors (PI)

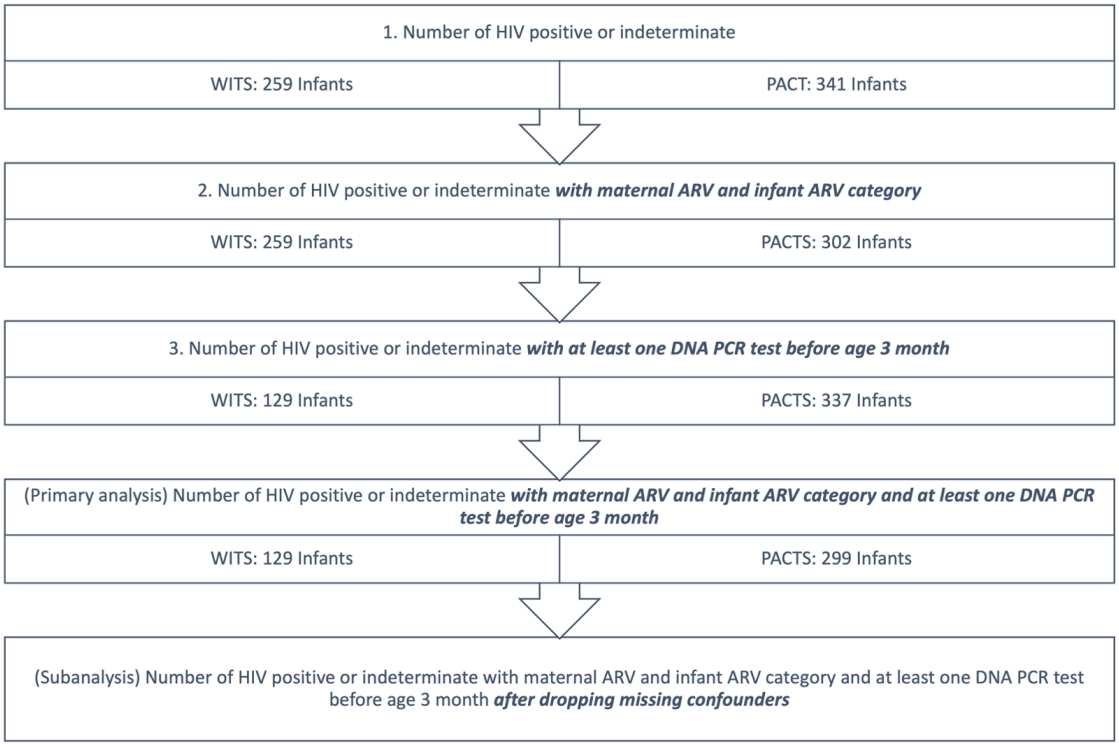


Figure 2.1. Inclusion/exclusion criteria for selecting participants from the PACTS and WITS cohorts into the analysis dataset.

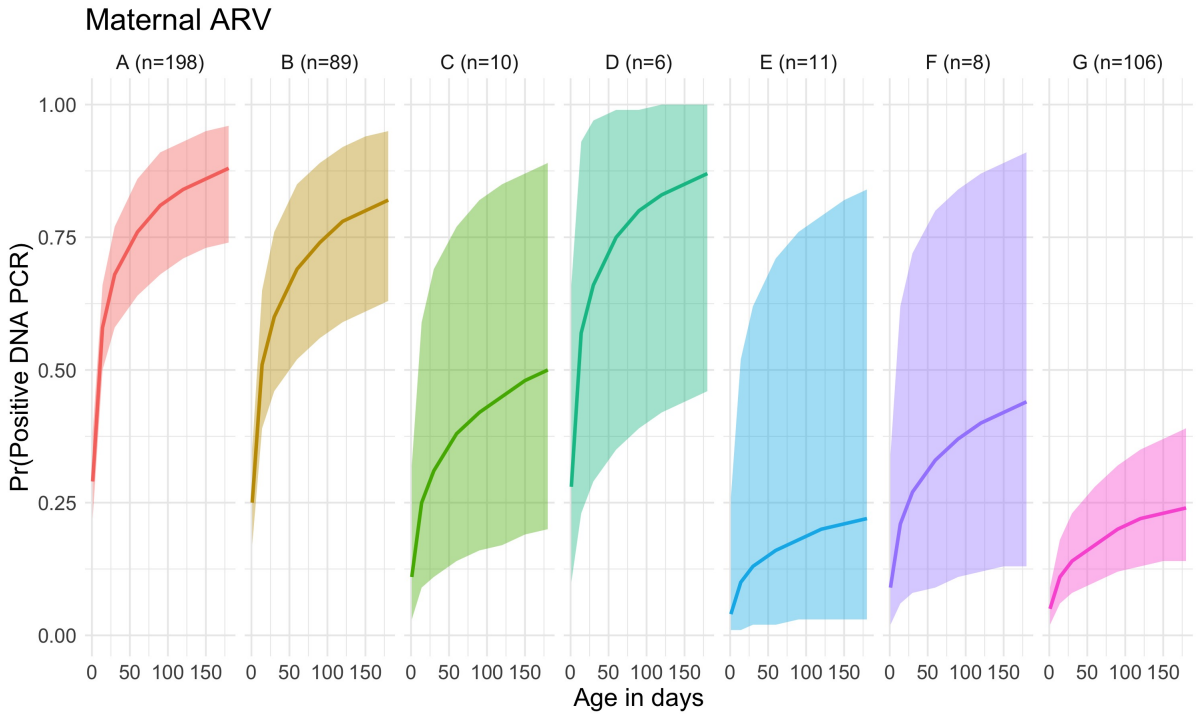


Figure 2.2. Probability of a positive HIV-1 DNA PCR test as a function of age (in days) from birth to 180 days among HIV-infected infants by type of maternal antiretroviral regimen. Shaded area indicate the 95% confidence interval. Results from a Weibull PH model without adjusting for confounders. A: no ARV; B: Single NRTI; C: sdNVP+ZDV; D: sdNVP only; E: 2-3 NRTIs without sdNVP; F: 2-3 NRTIs with sdNVP; G: cART.

CHAPTER 3

A MIXTURE MODEL FOR ESTIMATING THE RISK OF PROSTATE CANCER PROGRESSION AND THE FRACTION OF INDOLENT CANCER IN ACTIVE SURVEILLANCE

3.1 Introduction

Active surveillance (AS) has become a widely accepted management strategy for early-stage prostate cancer to reduce the adverse effects of unnecessary treatment and offer patients the opportunity to receive effective treatments [12, 29]. Patients with low-risk prostate cancer do not undergo active treatment immediately but are assigned to a schedule of regular biopsies and PSA measurements instead to detect cancer progression. Cancer progression is defined as an upgrade from low-risk to high-risk.

During the initial diagnosis, patients are classified into either low-risk, who may benefit from AS, or high-risk, who need to receive immediate treatment. However, the criteria for determining low-risk cancer in AS are not perfect. A study shows that the sensitivity and specificity of the diagnostic tests vary largely (8.5–97.9% and 24.7–97.8%, respectively) across 16 AS criteria, which suggests that there may be nonignorable misclassification at study entry [59]. Individuals already at high-risk may enter AS due to imperfect enrollment criteria at baseline, and be diagnosed with high-risk cancer during surveillance. Hence, prostate cancers monitored through AS are potentially heterogeneous consisting of (undetected) high-risk cancer at the time of diagnosis (i.e., prevalent cases) and truly low-risk at baseline.

Furthermore, among the non-prevalent cases, there are slow-growing cancers that remain indolent indefinitely. An autopsy study shows that some latent prostate tumors may not be progressive throughout the life span [74]. It is important to account for this cancer heterogeneity when studying the risk of prostate cancer progression in AS. Since the types of cancer – prevalent case, slow-growing but progressive cancer, and indolent cancer – are often unobserved, modeling time to progression cannot be done by applying conventional survival models.

A commonly used approach to account for unobserved heterogeneity in the population is the mixture model. In this paper, we consider extending the mixture cure model, which consists of two components – a survival model for susceptible subjects and a cure rate model for non-susceptible subjects. Several parametric mixture cure models have been proposed [5, 19, 58]. However, parametric mixture models are not robust when the underlying distribution is mis-specified. [37] proposed a generalized semiparametric model with a logistic regression model combined with the proportional hazards (PH) specification for survival time, which was developed on the basis of [19]. The semiparametric mixture cure models using the PH model [60, 70, 18, 30], the accelerated failure time model [77] or the proportional odds model [43] have been extensively studied for modeling the association between the event time and the risk factors. For estimating the risk of cancer progression in AS, the existing mixture cure models need to be further extended to incorporate the proportion of undetected prevalent cases (i.e., the misclassification rate at entry) as well as those who remain indolent.

Beyond its heterogeneity survival data nature, another challenging is that the onset of cancer progression cannot be directly observed for susceptible group. We can only know the occurrence of cancer progression happens between the last time subjects test to be low-risk and the first time subjects are in high-risk, which is characteristic as interval-censored. Mixture models have been investigated with respect to

interval censoring outcomes. A semiparametric model was proposed in the presence of non-susceptible group for case I interval censored (current status) data [38]. [45] studied mixture cure model under case II interval censored obtained from a sequence of examinations. [8] proposed semiparametric transformation mixture cure models for interval-censored data to describe distribution of non-susceptilbe rate and event time for susceptible group. However, none of these methods account for time-dependent covariates.

In addition to the heterogeneity of cancer, there are additional features of the data collected in AS that introduce challenges in modelling cancer progression. Individuals under AS follow predetermined protocols including regular clinical visits, prostate-specific antigen (PSA) measurements and repeated biopsies to detect prostate cancer progression. Prostate cancer progression is usually detected by an increase in tumor grade or volume on biopsy. Because cancer grades are evaluated through biopsy at scheduled visits, cancer progression time is not directly observed, but only known to lie between the previous visit (i.e., when the individual last tests negative) and the current visit at which the individual are classified as high-risk. Hence, the time to progression is interval-censored. In the literature, mixture models have been investigated for interval-censored outcomes. A semiparametric model was proposed in the presence of non-susceptible group for case I interval-censored (i.e., current status) data [38]. [45] studied the mixture cure model under case II interval-censored data obtained from a sequence of examinations. [8] proposed semiparametric transformation mixture cure models for interval-censored data to model the distribution of non-susceptible rate and the event time distribution for susceptible group. These approaches, however, cannot be applied to data collected in AS, even after incorporating the prevalent cases because the test results at each visit may not be accurate. In existing studies, the sensitivity and specificity of prostate biopsy ranged between 8.5–97.9% and 24.7–97.8%, respectively [59].

In this paper, we propose a semiparametric likelihood-based approach which is established relying on the work of [26] to handle interval-censored observations while incorporating the misclassification rates of biopsy. We construct the likelihood based on a mixture model with mixing parameters for the prevalence rate at baseline and the fraction of indolent cancer, and the survival model for cancer progression. The model allows us to estimate the risk of cancer progression accounting for prevalent and indolent cases, and the fraction of indolent cancer in AS.

The organization of this paper is as follows. In Section 2, we introduce the model, the likelihood function and the semiparametric maximum likelihood estimation for interval-censored data with misclassification. In Section 3, we simulate data and evaluate the performance of the proposed approach under various settings. We further assess how the proposed approach performs when the fraction of indolent cancer is ignored and when incorporating time-varying covariates. In Section 4, we apply the proposed method to data from the Canary Prostate Active Surveillance Study cohort[53] for estimating the risk of cancer progression among early-stage prostate cancer patients. We make conclusion and remarks in Section 5.

3.2 Methods

3.2.1 The mixture model

To account for potential cancer heterogeneity (i.e., prevalent, progressive and indolent cancers), we assume a mixture model that extends the two-component mixture cure model [19]. In our mixture model, we model (1) factors that are associated with the risk of cancer progression and (2) the fraction of indolent cancer adjusting for covariates, while incorporating the proportion of prevalent cases undetected at baseline. Let random variable T denote the time from prostate cancer diagnosis to cancer progression with a survival function $S(t \mid \boldsymbol{z})$, where $\boldsymbol{Z}$ is a $p\times 1$ vector of covariates for progressive cancer. Without loss of generality, we set $T^* = v\cdot 0+(1-v)\{(1-c)\cdot T+c\cdot\infty\}$,

where v and c are the indicators of prevalent case and indolent cancer, respectively. Then, $T^* = 0$ for prevalent cases ($v = 1$) and $T^* = \infty$ when cancer is indolent indefinitely ($c = 1$). We consider a mixture model for the marginal survival function in the form as follows:

$$\begin{aligned}\Pr(T^* > t; \boldsymbol{x}, \boldsymbol{z}) &= (1-\eta)\{(1-\pi(\boldsymbol{x}))\Pr(T^* > t \mid v = 0, c = 0) + \pi(\boldsymbol{x})\} \\ &= (1-\eta)\{(1-\pi(\boldsymbol{x}))S(t \mid \boldsymbol{z}) + \pi(\boldsymbol{x})\},\end{aligned}$$

where η is the probability of being prevalent at baseline and $\pi(\boldsymbol{x})$ is the probability of indolent cancer with $\boldsymbol{X}$ being a $d \times 1$ vector of covariates associated with the fraction of indolent cancer. Alternatively, the model can be rewritten as

$$\Pr(T^* \leq t; \boldsymbol{x}, \boldsymbol{z}) = \eta + (1-\eta)(1-\pi(\boldsymbol{x}))\{1 - S(t \mid \boldsymbol{z})\},$$

which is the cumulative risk. We note that $\Pr(T^* > t)$ and $\Pr(T^* \leq t)$ are improper mixture distributions with a mixing parameter η (i.e., the prevalence rate at baseline), a mixing proportion $\pi(\boldsymbol{x})$ (i.e., the indolent fraction), and the following component distributions: $\Pr(T^* > t \mid v = 1) = 0$, $\Pr(T^* \leq t \mid v = 1) = 1$, $\Pr(T^* > t \mid v = 0, c = 1) = 1$, and $\Pr(T^* \leq t \mid v = 0, c = 1) = 0$ as well as $\Pr(T^* > t \mid v = 0, c = 0) = S(t \mid \boldsymbol{z})$ and $\Pr(T^* \leq t \mid v = 0, c = 0) = 1 - S(t \mid \boldsymbol{z})$.

In the mixture model, we assume that the fraction of indolent cancer is associated with the covariates through a logit link as

$$\text{logit}(\pi(\boldsymbol{x})) = \widetilde{\boldsymbol{x}}^\top \boldsymbol{\alpha},$$

where $\widetilde{\boldsymbol{x}} = (1, \boldsymbol{x}^\top)^\top$ and $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_d\}$ is a $(d+1) \times 1$ vector of coefficients including the intercept. For modelling the risk of cancer progression, we assume a

Cox proportional hazards (PH) model. Given covariates $\boldsymbol{Z}$, the survival function of progressive cancer is

$$S(t \mid \boldsymbol{z}) = S_0(t)^{\exp(\boldsymbol{z}^\top \boldsymbol{\beta})}, \tag{3.1}$$

where $S_0(\cdot)$ is the baseline survival function and $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients.

3.2.2 Data, likelihood and estimation

In active surveillance, individuals, who were diagnosed with low-grade prostate cancer, including prevalent cases who were misclassified as low-grade at baseline, enter the program and undergo periodically scheduled biopsies to detect cancer progression. At each visit for biopsy, we observe whether each individual tested positive (i.e., high-grade cancer) or negative (i.e., low-grade). Therefore, time to cancer progression T is subject to interval-censoring and not observed directly.

Let N be the number of individuals in the study and n_i be the random variable denoting the number of each individual's visits, $i = 1, \ldots, N$. For the ith individual, we let $\boldsymbol{t}_i = \{t_{i1}, \ldots, t_{in_i}\}$ denote the sequence of visit times. Individuals continue to visit for scheduled biopsy until they test positive or the end of study. Hence, we observe a sequence of test results, $\boldsymbol{R}_i = \{r_{i1}, \ldots, r_{in_i}\}$, at corresponding visit times $\boldsymbol{t}_i$ where $r_{ik} = 0$ for $k = 1, \ldots, (n_i - 1)$ and r_{in_i} can either be 0 or 1 indicating negative and positive test results, respectively. The interval-censored time lies between $(t_{i(n_i-1)}, t_{in_i})$ when $r_{in_i} = 1$. When $r_{in_i} = 0$, time to progression is right-censored at t_{in_i}. We assume that $\boldsymbol{t}_i$ is independent of T_i^*, which implies that the individual follows the regular examination schedule regardless of cancer grade.

We denote the observed data as $\boldsymbol{O} = \{\boldsymbol{O}_i = (\boldsymbol{R}_i, \boldsymbol{t}_i, n_i, \boldsymbol{X}_i, \boldsymbol{Z}_i), i = 1, \ldots, N\}$. Let $\tau_1, \ldots, \tau_J$ be the distinct, ordered visit times among all $\{\boldsymbol{t}_i,\ i = 1, \ldots, N\}$, where $0 = \tau_0 < \tau_1 < \cdots < \tau_J < \tau_{J+1} = \infty$ and $[0, \tau_1), [\tau_1, \tau_2), \ldots, [\tau_J, \infty)$ become $J + 1$

disjoint intervals. Then, the conditional probability of the observed data given the covariates $(\boldsymbol{X}_i, \boldsymbol{Z}_i)$ for the ith individual is

$$g(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = \sum_{j=1}^{J+1} \Pr(\tau_{j-1} < T_i \leq \tau_j \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) \times \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid \tau_{j-1} < T_i \leq \tau_j, \boldsymbol{X}_i, \boldsymbol{Z}_i). \tag{3.2}$$

Following [26], we assume that

$$\Pr(\boldsymbol{R}_i \mid T_i, \boldsymbol{t}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \prod_{k=1}^{n_i} \Pr(r_{ik} \mid T_i, t_{ik}, \boldsymbol{X}_i, \boldsymbol{Z}_i).$$

This assumes that the sequence of biopsy tests are independent given individual's time to progression. Based on the derivation provided in the Appendix of [3], the probability in Equation (3.2) can be further simplified as:

$$g(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = \sum_{j=1}^{J+1} \Pr(\tau_{j-1} < T_i \leq \tau_j \mid \boldsymbol{X}_i, \boldsymbol{Z}_i)\omega_{ij},$$

where $\omega_{ij} = \prod_{k=1}^{n_i} \Pr(r_{ik} = 1 \mid \tau_{j-1} < T_i \leq \tau_j, t_{ik}, \boldsymbol{X}_i, \boldsymbol{Z}_i)$.

3.2.2.1 Biopsy misclassification during surveillance

Biopsy tests at regularly scheduled visits during surveillance are also imperfect as the initial diagnostic test. Let δ_1 and δ_0 denote the sensitivity and specificity of biopsy. We assume that the probability of testing positive given the progression time is independent of the covariates, which leads to

$$\Pr(r_{ik} = 1 | \tau_{j-1} < T_i \leq \tau_j, t_{ik}, \boldsymbol{X}_i, \boldsymbol{Z}_i) = \Pr(r_{ik} = 1 | \tau_{j-1} < T_i \leq \tau_j, t_{ik}) = \begin{cases} \delta_1, & t_{ik} \geq \tau_j \\ 1 - \delta_0, & t_{ik} \leq \tau_{j-1} \end{cases}.$$

In Equation (3.1), for the baseline survival function, we assume that

$$S_0(t) = S_0(\tau_{j-1}) \text{ for } \tau_{j-1} \leq t < \tau_j,$$

$j = 1, \ldots, J+1$, and let $S_j = S_0(\tau_{j-1})$. This leads to $1 = S_1 > S_2 > \cdots > S_{J+1} > 0$, among which $\{S_j, j = 2, \ldots, J+1\}$ are the unknown parameters. Then, for the ith individual,

$$\Pr(\tau_{j-1} < T_i \leq \tau_j \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = (S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})} - (S_{j+1})^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}.$$

It follows that the conditional probability in Equation (3.2) can be rewritten as

$$\begin{aligned} g(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) &= \sum_{j=1}^{J+1} \{(S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})} - (S_{j+1})^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}\} \omega_{ij} \\ &= \sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}, \end{aligned}$$

where D_{ij} is the (i,j)th element of the matrix $\boldsymbol{D} = \boldsymbol{\Omega} \boldsymbol{T}_{\boldsymbol{r}}$, in which $\boldsymbol{\Omega} = (\omega_{ij})_{N \times (J+1)}$ and

$$\boldsymbol{T}_{\boldsymbol{r}} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{(J+1)\times(J+1)},$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, J, J+1$. We note that D_{ij}'s incorporate the constant sensitivity and specificity rates of biopsy, δ_1 and δ_0, respectively, in their functions of the visit times and the corresponding test results.

Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \boldsymbol{S}^\top)^\top$, where $\boldsymbol{S} = (S_2, \ldots, S_J, S_{J+1})^\top$ be the set of unknown parameters. Then, the likelihood for the ith individuals is

$$L_i(\boldsymbol{\theta}) \propto g(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i) = \sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})}.$$

Note that the likelihood is constructed for susceptible individuals who have slow-growing but progressive cancer.

3.2.2.2 Prevalent cases and indolent cancer

Let b denote the diagnostic test result at baseline where $b = 0$ indicates low-grade cancer (i.e., negative). Since only individuals testing negative enter the surveillance program, everyone under surveillance has $b = 0$. Then, for η which represents the proportion of prevalent cases in the study in Section 3.2.1, we can define it as $\eta = \Pr(v = 1 \mid b = 0)$, where v indicates the true cancer grade at baseline. We assume that $\Pr(v_i = 1 \mid b_i = 0) = \Pr(v_i = 1 \mid b_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i)$, i.e., the probability of prevalent cases is independent of the covariates. For the ith individual, the likelihood is

$$\begin{aligned} L_i(\boldsymbol{\theta}) \propto & \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid b_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ = & \eta \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid v_i = 1, b_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ & + (1 - \eta) \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid v_i = 0, b_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i). \end{aligned}$$

We assume that individuals who are truly negative and test negative at baseline are random samples of the population who are truly negative at baseline. To further incorporate the fraction of indolent cancer among those who are truly negative (i.e., $v_i = 0$), we re-express $\pi(\boldsymbol{x})$ in Section 3.2.1 as $\pi(\boldsymbol{x}) = \Pr(c = 1 \mid v = 0, b = 0, \boldsymbol{x}) =$

$\Pr(c = 1 \mid v = 0, \boldsymbol{x})$ by the aforementioned assumption. For the ith individual, the likelihood becomes

$$\begin{aligned} L_i(\boldsymbol{\theta}) \propto & \eta \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid v_i = 1, b_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i) \\ & + (1-\eta)\left\{(1-\pi(\boldsymbol{x}_i)) \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid v_i = 0, c_i = 0, \boldsymbol{X}_i, \boldsymbol{Z}_i)\right. \\ & \left. + \pi(\boldsymbol{x}_i) \Pr(\boldsymbol{R}_i, \boldsymbol{t}_i, n_i \mid v_i = 0, c_i = 1, \boldsymbol{X}_i, \boldsymbol{Z}_i)\right\} \\ = & \eta D_{i1} S_1 + (1-\eta)\left\{(1-\pi(\boldsymbol{x}_i)) \sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})} + \pi(\boldsymbol{x}_i) D_{i(J+1)}\right\}. \end{aligned}$$

Then, the log likelihood for N individuals is

$$\ell(\boldsymbol{\theta}) \propto \sum_{i=1}^{N} \log \left[\eta D_{i1} S_1 + (1-\eta)\left\{(1-\pi(\boldsymbol{x}_i)) \sum_{j=1}^{J+1} D_{ij} (S_j)^{\exp(\boldsymbol{z}_i^\top \boldsymbol{\beta})} + \pi(\boldsymbol{x}_i) D_{i(J+1)}\right\}\right]. \tag{3.3}$$

The gradient function of log likelihood presented in Equation (3.3) is in Supplement (C.1.1).

3.2.2.3 Time-varying covariates

Suppose $\boldsymbol{\xi}_i(\tau_j)$ is the subject specific time-varying covariates measured at τ_j and assume it's constant during the interval $[\tau_j, \tau_{j+1})$. Let $\boldsymbol{P}_i(t) = (\boldsymbol{\xi}_i(t), \boldsymbol{Z}_i)$ be the design matrix including both time-varying and time-invariant covariates ($\boldsymbol{Z}_i$) that related to survival times. Let Λ_j and $\Lambda_j \exp(\boldsymbol{P}_i(\tau_j)^\top \boldsymbol{\beta})$ be the cumulative hazards during interval $[\tau_j, \tau_{j+1})$ for baseline group and the other group, respectively. The survival function for subject i at τ_{j-1} can be expressed as

$$S_j^{(i)} = \exp\left(-\sum_{l=0}^{j-2} \Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\right),\ j = 2, \ldots, J+1. \tag{3.4}$$

Let $\boldsymbol{\gamma} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top, \Lambda_0, \Lambda_1, \ldots, \Lambda_{J-1})^\top$ be the unknown parameters. The log likelihood function with time-varying covariates can be written as

$$\ell(\boldsymbol{\gamma}) \propto \sum_{i=1}^{N} \log \left[\eta D_{i1} + (1-\eta) \left\{ \left(1 - \pi_i(\boldsymbol{x}_i)\right) \sum_{j=1}^{J+1} D_{ij} \exp\left(- \sum_{l=0}^{j-2} \Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta}) \right) + \pi_i(\boldsymbol{\alpha}) D_{i(J+1)} \right\} \right]. \tag{3.5}$$

The gradient function of log likelihood presented in Equation (3.5) is in Supplement (C.1.2).

3.2.2.4 Estimation

To estimate parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, we directly maximize the log likelihood functions presented in Equation 3.3 and 3.5 with the constraints $1 = S_1 > S_2 > \cdots > S_{J+1} > 0$ and $\Lambda_j > 0$ for models with the time-invariant covariates and the time-varying covariates, respectively.

The algorithm is implemented using R package, `constrOptim` [64], by which we can set a feasible region for each parameter and search the interior of that region with constraints. The standard error of the point estimates can be obtained either asymptotically or by the bootstrap method. The asymptotic standard deviation is derived using the estimated covariance matrix, which is the inverse of the Hessian matrix, and the bootstrap standard error by resampling the dataset with replacement.

The odds ratio of indolent cancer and the hazard ratio of cancer progression, which are expressed by $\exp(\boldsymbol{\alpha})$ and $\exp(\boldsymbol{\beta})$, represent how likely individuals with low-risk cancer in a certain group have indolent cancer compared to the reference and how often individuals with low-risk cancer in a certain group experience cancer progression compared to the reference, respectively.

3.3 Simulation

We conduct simulation studies to assess the performance of the proposed approach accounting for (1) the proportion of prevalent cases (i.e., misclassification at study entry); (2) the fraction of indolent cancer; (3) biopsy misclassification (i.e., sensitivity and specificity of biopsy); and (4) the presence of time-varying covariates. Under various settings, we examine the bias and the standard deviation of the estimates, the standard errors, and the coverage probabilities.

We simulate 1000 data sets with $N = 1000$ subjects each. We assume that all subjects are assigned to the same fixed sequence of visit times $(\tau_1, \ldots, \tau_J)$, where $0 = \tau_0 < \tau_1, \ldots, \tau_J < \infty$. We set $J = 6$ and the visit times to be $\tau_1 = 0.5, \tau_2 = 1, \tau_3 = 2, \tau_4 = 4, \tau_5 = 6, \tau_6 = 10$ to mimic the protocol of the Canary Prostate Active Surveillance Study (PASS) [54].

3.3.0.0.1 Simulating event time with time-invariant covariate We simulate a covariate Z from a standard normal distribution. Under the setting where the covariate is time-invariant, we generate event times from an exponential distribution $T \sim Exp(\lambda)$, where $\lambda = \lambda_0 \times \exp(Z^\top \beta)$ assuming a Cox PH model for time to progression.

3.3.0.0.2 Simulating event time with time-varying covariate To consider time-varying covariates, we generate $\boldsymbol{\xi}_i(t)$ for subject i by simulating J random variables from $Uniform(0, 1)$, and sorting them in increasing orders to correspond to the covariate values at each visit $\boldsymbol{\xi}_i = \{\xi_i(\tau_1), \ldots, \xi_i(\tau_J)\}$. We then combine the time-varying covariate $\xi_i(\tau_j)$ and the time invariant covariate Z_i and denote it as $\boldsymbol{P}_i(\tau_j)$. For subject i, the hazard rate at τ_j is $\lambda_{ij} = \lambda_j^{(0)} \exp(\boldsymbol{P}_i(\tau_j)^\top \boldsymbol{\beta})$, where $\lambda_j^{(0)}$ is the baseline hazards at τ_j for $j = 1, \ldots, J$, and $\boldsymbol{\beta} = \{\beta^{(Z)}, \beta^{(\xi)}\}$ is a vector of the regression coefficients for time-invariant and time-varying covariates, respectively. The cumulative hazard function during interval $[\tau_j, \tau_{j+1})$ is derived as

$\Lambda^*_{ij} = \lambda^{(0)}_j \exp(\boldsymbol{P}_i(\tau_j)^\top \boldsymbol{\beta}) \times (\tau_{j+1} - \tau_j)$. Let $\Lambda^{(0)}_j = \lambda^{(0)}_j \times (\tau_{j+1} - \tau_j)$. The corresponding cumulative hazard function at τ_j can be written as $\Lambda_{ij} = \sum_{l=0}^{j-2} \Lambda^{(0)}_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})$. The event times T_i can be generated as follows

$$T_i = \begin{cases} \frac{-\log U_i}{\lambda^*_{i1}} & \text{if } 0 < -\log U_i < \Lambda_{i1} \\ \tau_{ik} + \frac{(-\log U_i - \Lambda_{ik})}{\lambda^*_{i(k+1)}} & \text{if } \Lambda_{ik} \leq -\log U_i < \Lambda_{ik+1}, 1 \leq k < J \\ \tau_{iJ} + \frac{(-\log U_i - \Lambda_{iJ})}{\lambda^*_{iJ}} & \text{if } -\log U_i \geq \Lambda_{iJ}, \end{cases}$$

where U_i is randomly simulated from $Uniform(0, 1)$.

3.3.0.0.3 Misclassification at study entry In Section 3.2.2.2, we define η to represent the proportion of prevalent cases. By definition, it follows that $1 - \eta$ is the baseline negative predictive value, which is the probability of being truly low-risk given that they were tested low-risk at baseline. If baseline diagnostic test perfectly screens those who are low-risk, $\eta = 0$. For each subject i, we simulate a binary random variable κ_{mi} from a binomial distribution with probability η. If $\kappa_{mi} = 1$, we replace event time T_i with 0, for $i = 1, \ldots, N$.

3.3.0.0.4 Fraction of indolent cancer For each subject i, we assume the probability of having indolent cancer is $\pi(x_i) = \frac{\exp(\boldsymbol{x}_i{}^\top \boldsymbol{\alpha})}{1+\exp(\boldsymbol{x}_i{}^\top \boldsymbol{\alpha})}$, where $\boldsymbol{x}_i = \{1, x_i\}$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$ and x_i is randomly sampled from a standard normal distribution. We prepecify *pcure* as the overall probability of being event-free indefinitely. We select $\boldsymbol{\alpha}$ that satisfies the overall probability $pcure = \frac{1}{N} \sum_i^N \pi_i(\pm x_i)$. We generate the indicator of indolent cancer κ_{ci} for each subject i from a binomial distribution with probability $\pi(\boldsymbol{x}_i)$). If $\kappa_{ci} = 1$, we replace event time T_i with ∞, for $i = 1, \ldots, N$.

3.3.0.0.5 Biopsy misclassification during surveillance We pre-specify the sensitivity and specificity of biopsy, δ_1 and δ_0. Let R^*_{ij} be the true cancer status for each subject i at visit time τ_j, which can be determined by the generated cancer

progression time T_i. We set $R^*_{ij} = 0$ if $T > \tau_j$ indicating a true low-grade cancer, and $R^*_{ij} = 1$ otherwise indicating a true high-grade cancer. We generate test results R_{ij} accounting for the sensitivity and specificity rates of biopsy given the true status. If $R^*_{ij} = 1$, then R_{ij} is simulated from a binomial distribution with probability δ_1. If $R^*_{ij} = 0$, then R_{ij} is randomly sampled from a binomial distribution with δ_0.

3.3.0.0.6 Scenario I First, we assess the performance of our proposed mixture model under varying levels of indolent cancer fraction. We set *pcure* to $1\%, 20\%$ and 50%, which correspond to small, median and large proportion of indolent cancer among those who are not prevalent, respectively. Additionally, we set $\beta = 0.7$, the misclassification rate at entry $\eta = 0.2$, the sensitivity rate $\delta_1 = 0.8$, and the specificity rate $\delta_0 = 1$. We compare our proposed mixture model with the model proposed in [26] which doesn't account for the fraction of indolent cancer. We summarize the point estimates and compute the coverage probabilities based on asymptotic standard errors and bootstrap standard errors.

3.3.0.0.7 Scenario II To test the robustness of the proposed approach, we consider a setting where there is no indolent cancer. Under this scenario, we set $pcure = 0$ and keep all other parameters the same as Scenario I.

3.3.0.0.8 Scenario III Lastly, we consider a time-varying covariate in the model and evaluate the performance of the proposed approach. We apply our semiparametric approach presented in Section 3.2.2.3. The asymptotic standard errors of $\boldsymbol{\beta}$ are derived by the delta method [16].

3.3.1 Results

In Table 3.1, we compare our semiparametric mixture cure model with a non-mixture model[26] which ignoring the fraction of indolent cancer under very small (1%), median (20%) and large (50%) proportion of indolent cancer in the simulate

data. We report both asymptotic standard deviation and bootstrap standard error and corresponding 95% coverage probability for our proposed method. In table of small proportion of indolent cancer, the covariate effect β for Cox PH model (Equation 3.1) is 0.7028 (95% CI: 96.24%, 97.1%-BS) and 0.6678 (95% CI: 90.1%) in proposed model and non-mixture model. The baseline survivals estimates at visit time 0.5, 1, and 2 from both models are quite accurate compared to the truth and 95% CI are all around 95%. While at visit time 4, 6 and 10, the survival estimates from non-mixture model are getting more and more biased with 95% CI decreasing to 91.51%, 88.41% and 76.1%, respectively. Table of median indolent cancer shows that the estimate of covariate effect β from non-mixture model is underestimated (0.41, Truth: 0.7) with a 0.21% coverage probability. The only accurate estimate is the survival at visit time 0.5, while the rest of survival estimates overestimate the truth with very low coverage probabilities (79.32%, 24.31%, 0%, 0%, 0%). As fraction of indolent cancer increasing to 50%, all of the estimates from non-mixture model overestimate the truth and coverage probability are all below 1%. This is because the non-mixture model treating all indolent cases into incident cases, so more subjects categorized in the risk set render an overestimate of baseline survivals.

Table 3.2 presents the results under no indolent cancer scenario. The Proposed model as a more general form can handle this scenario as well as the non-mixture model. The coverage probability of all estimates are around nominal 95% which indicates a goodness of fit.

We also report standard deviation and coverage probability based on both asymptotic and bootstrap method. The results are overall comparable, though bootstrap standard error is slightly larger than asymptotic one, resulting in a relative conservative coverage probability. We think this might due to local variation in the data [4].

In Table 3.3, we show the results from proposed semiparametric mixture cure model by incorperating time-varying covariates in Cox model. Point estimates are very close to the truth. Standard error from both methods are in consistency. The coverage probabilities from both methods are well aligned. The survival estimate at visit time 10 is small and the standard deviation are relative large, causing a slightly off coverage probability ($\tilde{9}0\%$), although still acceptable. This might be because there's very few subject fall into the last interval.

3.4 Application

The Canary Prostate Active Surveillance Study (PASS) [54] enrolled 1067 participants during 2008-2013. The eligibility criteria included confirmed prostate cancer with clinical stage T1-2, no previous treatment and either a $\geq$ 10-core biopsy within one year before enrollment or $\geq$ 2 biopsies one of which was within 2 year before enrollment. Participants measured PSA every 4 months and repeated biopsy 0.5, 1, 2, 4, 6 years after diagnosis. After exclusions, there were 652 patients with no progression at study entry, at least one follow-up biopsy and PSA after study begins included in the dataset. There were 428 (65.64%) of participants censored and the median follow-up time was 2 (IQR:1-3.45) years. The median of biopsy and PSA frequencies per patient were 1 (IQR: 1-2) and 8 (IQR: 4-14) respectively.

Demographics of cohorts are summarized in Table 3.4. Median age at diagnosis was 63 (IQR: 58-67). Median PSA during follow-up was 4.65 ng/mL (IQR:3.15-6.39). 91% of participants were white. Most participants were in clinical T1 stage (88%), didn't assess lymph nodes to cancer (NX=97%), didn't evaluate metastasis (MX=97%) at baseline. Figure 3.1 shows PSA changing overtime, where PSA from progression-free group (N=428) are more variant than progression group (N=224). For identifiability of the model parameters, we kept PSA records after the last biopsy visit because sufficient follow-ups can help yield unbiased results [17]. For those with

PSA missing at some visit times, we used the PSA level at previous visit. To improve convergence, we centered all continuous variables.

We are interested in the risk factor of prostate cancer progression from low-risk to high risk and the fraction of indolent cancer in the AS cohort. Figure 3.2 shows the nonparametric maximum likelihood estimates (NPMLE) [72, 23] of survivals for different categories ignoring indolent cancer. For example, survival curve of age greater than 63 year-old group levels off to a non-zero plateau after 5 years which counts for 24 (8.79%) participants. Since this long flat tail may indicate a presence of indolent cancer, we use the mixture model to account for the unobserved cancers that remain indolent.

We fit a mixture model with age and median PSA and adjust for one covariate each time to check the significance of covariates using likelihood ratio test (See Table 3.4). Race (pvalue=0.41), clinical T stage (pvalue=0.33), clinical N stage (pvalue=0.12), clinical M stage (pvalue=0.11) are all insignificant, hence they are not incorporated into the final model. We assume independent censoring, that is, patients' dropout is independent of their progression status. Our final model includes diagnostic age and median PSA into logistic regression model to account for the indolent cancer based on preliminary studies [69], and age and time-dependent variable PSA to estimate survivals. We use a step function to model time-varying covariate which assumes that the values of the time-varying covariate are constant in each time interval. To investigate the impact of imperfect biopsy test (i.e., misclassification rate during surveillance), we consider a range of biopsy sensitivity of 90%, 80%, 70%, and specificity of 90%, following the literature [31, 65]. The misclassification rate at entry η is set to be equal to 1 minus the negative predictive value (PPV). We set the prevalence (biopsy upgrade rate among patients eligible to active surveillance) to be 0.579 [1]. Then η can be obtained by the following formula,

$$\eta = 1 - PPV = 1 - \frac{Specificity \times (1 - Prevalence)}{Specificity \times (1 - Prevalence) + (1 - Sensitivity) \times Prevalence}$$

The standard errors and the 95% confidence intervals of the estimates are calculated through 100 bootstrap resampling.

The estimated coefficients, bootstrap standard errors, P-values, odds ratio (OR) or hazard ratio (HR), and the overall indolent cancer rate are reported in Table 3.5. In the logistic regression for indolent caner rate, there are no covariates with significant impact on the fraction of indolent cancer. For progressive cancer, *Age at Diagnosis* is significant with HR=9.08 (95% CI: 2.5-32.92) for sensitivity (δ_1)=0.9 and specificity (δ_0)=0.9, HR=6.79 (95% CI: 1.34-34.46) for $\delta_1 = 0.8$ and $\delta_0 = 0.9$, indicating that among those who will eventually progress to high risk, older subjects can have a higher risk than younger subjects, however, it becomes insignificant with HR=5.03 (95% CI: 0.79-32.22) when $\delta_1 = 0.7$ and $\delta_0 = 0.9$. PSA is not significantly associated with grade progression, which is consistent with findings in prior studies [66]. The predicted overall probability of indolent cancer ranges from a median of 50.04% to 57.51% with respect to different sensitivity and specificity assumptions, which is closed to a report of 55% in a 15-year follow-up Canadian active surveillance study [35]. As sensitivity decreases, the model tends to overestimate the fraction of indolent cancer.

3.5 Conclusion

Prior studies that estimate the risk of prostate cancer progression in active surveillance have not considered the fraction of indolent cancer. It has been shown that some low-risk participants in the AS program do not experience a progression to high-risk even after a long follow-up [32], which indicates the potential mixture of progressive and indolent cases in prostate cancer. Ignoring the presence of indolent cancer would lead to an underestimation of the risk of cancer progression.

This paper inherits the most commonly used framework for modeling interval censored survival outcome in a heterogeneous population. We extend the semiparametric mixture cure model to incorporate misclassification at entry, imperfect diagnostic tests and time-varying covariates. The proposed mixture model have three components, a mixing parameter for prevalent cases, a logistic regression model for the indolent cancer, and a semiparametric Cox PH model with a piecewise constant baseline hazard function to model time to progression. Our simulation results indicate that the proposed model provides satisfactory results under a large range of indolent cancer fraction given the sensitivity and specificity. However, the model tends to overestimate the indolent fraction when sensitivity is too low. The reliability of the estimates is conditional on reasonable assessment of the sensitivity and specificity.

Another limitation is that although our model can handle irregular visits, our NPMLE-type of estimation suffers from computational complexity as there are no closed form. Moreover, with too many distinct intervals, some Λ_j's in Equation 3.5 become zero and result in convergence issue. To circumvent these problems, we rounded the visit times to integers in the application. Further work is needed to refine the current proposed method, for example, by incorporating a penalized likelihood function which only retain the non-zero intervals [33]. It might also be interesting to modify our estimation approach to an efficient EM-type algorithm which estimates the Λ_j as weighted sums of Poisson rates [76].

A key assumption we made in this article is independent censoring. We assume that the time of censoring such as individual's dropout is independent of cancer progression or the underlying health condition. However, this assumption may be violated in the case where patients decide to receive treatment without testing positive due to an increasing PSA level. Ignoring dependent censoring may lead to biased estimates. The current framework can be extended by using a class of Archimedean cop-

ula models [51, 40, 44] or using an inverse probability of weighted censoring (IPWC) to derive unbiased estimator to account for dependent censoring [67, 49, 57].

Table 3.1. Results for Very small (1%), Median (20%) and Large (50%) indolent cancer. We compare estimates from the proposed semiparametric mixture cure model and a non-mixture model. β is the log(Hazard Ratio) of time-invariant covariate. $S_{0.5}, \ldots, S_{10}$ are survivals at visit time $0.5, \ldots, 10$.

Parameter	True	Estimate (Std, Std^{BS})		Coverage Probability		
		Proposed Model	Non-Mixture Model	Proposed	Proposed^{BS}	Non-Mixture Model
			1% of indolent cancer			
β	0.7	0.7028 (0.0583,0.0612)	0.6678 (0.0545)	0.9624	0.971	0.9006
$S_{0.5}$	0.9048	0.9009 (0.0173,0.0318)	0.9033 (0.0169)	0.9436	0.9679	0.9472
S_1	0.8187	0.8154 (0.0188,0.0324)	0.8183 (0.0182)	0.9426	0.9627	0.9462
S_2	0.6703	0.6667 (0.0216,0.0319)	0.6729 (0.0205)	0.9572	0.9772	0.9524
S_4	0.4493	0.4475 (0.0235,0.029)	0.4598 (0.0212)	0.9489	0.9752	0.9151
S_6	0.3012	0.2986 (0.0227,0.0266)	0.3162 (0.019)	0.953	0.9752	0.8841
S_{10}	0.1353	0.1332 (0.0201,0.0215)	0.1542 (0.015)	0.9426	0.942	0.7609
			20% of indolent cancer			
β	0.7	0.7025 (0.0676,0.0702)	0.4146 (0.0515)	0.9488	0.9606	0.0021
$S_{0.5}$	0.9048	0.9035 (0.0202,0.0297)	0.9095 (0.0182)	0.9606	0.9872	0.9446
S_1	0.8187	0.8171 (0.0218,0.0308)	0.8369 (0.0182)	0.9318	0.9616	0.7932
S_2	0.6703	0.6701 (0.0249,0.0318)	0.7233 (0.0195)	0.9478	0.968	0.2431
S_4	0.4493	0.4488 (0.0271,0.0308)	0.5671 (0.0203)	0.9499	0.9723	0
S_6	0.3012	0.3009 (0.0264,0.029)	0.4688 (0.0195)	0.9595	0.9755	0
S_{10}	0.1353	0.1366 (0.0241,0.0253)	0.3587 (0.0184)	0.9638	0.9638	0
			50% of indolent cancer			
β	0.7	0.7114 (0.0926,0.0984)	0.2557 (0.0647)	0.9676	0.9708	0
$S_{0.5}$	0.9048	0.9045 (0.0271,0.0298)	0.9437 (0.0182)	0.9468	0.9687	0.4311
S_1	0.8187	0.8202 (0.0292,0.0319)	0.8999 (0.0175)	0.9457	0.9697	0.0042
S_2	0.6703	0.6723 (0.0332,0.0356)	0.8313 (0.0184)	0.9384	0.9551	0
S_4	0.4493	0.4497 (0.0365,0.0381)	0.7404 (0.0194)	0.953	0.9656	0
S_6	0.3012	0.3002 (0.0362,0.0379)	0.6841 (0.0195)	0.9363	0.9363	0
S_{10}	0.1353	0.1335 (0.034,0.0352)	0.6212 (0.0196)	0.9301	0.9468	0

Table 3.2. Results from 1000 simulation with 500 bootstrap each, where there're 1000 subjects, *None* of which have indolent cancer. We compare estimates from the proposed semiparametric mixture cure model and a non-mixture model. β is the log(Hazard Ratio) of time-invariant covariate. $S_{0.5}, \ldots, S_{10}$ are survivals at visit time $0.5, \ldots, 10$.

Parameter	True	Estimate (Std, Std^{BS})		Coverage Probability		
		Proposed Model	Non-Mixture Model	Proposed	Proposed^{BS}	Non-Mixture Model
β	0.7	0.7097 (0.0561,0.0584)	0.7001 (0.0547)	0.9355	0.9369	0.9317
$S_{0.5}$	0.9048	0.9022 (0.0171,0.0326)	0.9041 (0.0168)	0.9602	0.9911	0.9546
S_1	0.8187	0.8163 (0.0187,0.0337)	0.8186 (0.0182)	0.9209	0.9734	0.9233
S_2	0.6703	0.6662 (0.0215,0.0328)	0.6692 (0.0206)	0.9177	0.9791	0.9144
S_4	0.4493	0.4423 (0.0229,0.029)	0.4473 (0.0213)	0.9245	0.9588	0.9311
S_6	0.3012	0.2937 (0.0213,0.0255)	0.3008 (0.019)	0.9528	0.9609	0.9343
S_{10}	0.1353	0.1271 (0.0175,0.019)	0.1349 (0.0143)	0.9132	0.9218	0.9358

Table 3.4. Baseline Characteristics for time-invariant covariates

Variable	N = 652	LRT P-value
Age at Diagnosis	63 (58, 67)	-
median PSA, ng/mL	4.65 (3.15, 6.39)	-
Race		0.41
White	596 (91%)	
Other	56 (8.6%)	
Clinical T Stage		0.33
cT1(cT1a, cT1b)	576 (88%)	
cT2(cT2a, cT2b)	76 (12%)	
Clinical N Stage		0.12
NX	634 (97%)	
N0	18 (2.8%)	
Clinical M Stage		0.11
MX	633 (97%)	
M0	19 (2.9%)	

[1] Median (IQR); n (%)

Table 3.3. Results from 1000 simulation with 500 bootstrap each, where there're 1000 subjects, *20%* of which have indolent cancer. $\{\alpha_0, \alpha_1, \alpha_2\}$ are intercept and slopes from logistic regression for indolent cancer population. β is the log(Hazard Ratio) of time-invariant covariate and β_t is the log(Hazard Ratio) of time-variant covariate. $S_{0.5}, \ldots, S_{10}$ are survivals at visit time $0.5, \ldots, 10$.

Parameter	True	Estimate	Std		Coverage Probability	
			Asymptotic	Bootstrap	Asymptotic	Bootstrap
α_0	-2.4488	-2.4788	0.2658	0.2887	0.9537	0.9623
α_1	-1.1857	-1.2074	0.1819	0.1931	0.9623	0.9763
α_2	2.4721	2.4964	0.2780	0.2992	0.9483	0.9591
β	0.7000	0.7046	0.0677	0.0696	0.9634	0.9644
β_t	1.2000	1.2116	0.3424	0.3508	0.9418	0.9418
$S_{0.5}$	0.9048	0.9042	0.0189	0.0191	0.9547	0.9537
S_1	0.8187	0.8193	0.0238	0.0240	0.9569	0.9591
S_2	0.6703	0.6689	0.0385	0.0392	0.9397	0.9450
S_4	0.4493	0.4490	0.0602	0.0609	0.9246	0.9203
S_6	0.3012	0.3004	0.0691	0.0690	0.9278	0.9343
S_{10}	0.1353	0.1414	0.0647	0.0646	0.8987	0.9041

Table 3.5. Analysis of 652 patients in PASS cohort. Estimates of covariates of interest are based on sensitivity (δ_1) and specificity (δ_0) pairs, including $(\delta_1, \delta_0) = (0.9, 0.9), (0.8, 0.9)$ and $(0.7, 0.9)$. Estimates for indolent cancer are from logistic regression, where exp(Est) represents Odds Ratio (OR). We fit a survival function to model event time in susceptible group, where exp(Est) represents Hazard Ratio (HR).

	Variable	Estimates	Std	P-value	exp(Est) [95% CI]	Median indolent cancer rate [IQR]
	Estimates from Logistic Model for Indolent Cancer					
	Intercept	0.08	0.49	0.88	1.08 [0.41,2.85]	
	Age at Diagnosis	-0.21	0.66	0.75	0.81 [0.22,2.95]	
Sensitivity = 0.9	Median PSA	0.58	0.99	0.56	1.79 [0.26,12.4]	50.04% [42.48%-58.44%]
Specificity = 0.9	*Estimates from Survival Model for Progressive Cancer*					
	Age at Diagnosis	2.21	0.66	<0.001	9.08 [2.5,32.92]	
	PSA(t)	-0.09	0.82	0.91	0.91 [0.18,4.59]	
	Estimates from Logistic Model for Indolent Cancer					
	Intercept	0.24	0.67	0.72	1.28 [0.34,4.78]	
	Age at Diagnosis	0.29	0.6	0.63	1.34 [0.41,4.36]	
Sensitivity = 0.8	Median PSA	0.47	1.01	0.64	1.6 [0.22,11.55]	54.63% [47.95%-62.94%]
Specificity = 0.9	*Estimates from Survival Model for Progressive Cancer*					
	Age at Diagnosis	1.92	0.83	0.02	6.79 [1.34,34.46]	
	PSA(t)	0.59	0.84	0.48	1.81 [0.35,9.5]	
	Estimates from Logistic Model for Indolent Cancer					
	Intercept	0.36	0.92	0.7	1.43 [0.23,8.72]	
	Age at Diagnosis	0.9	1	0.37	2.46 [0.35,17.43]	
Sensitivity = 0.7	Median PSA	0.92	1.19	0.44	2.52 [0.25,25.74]	57.51% [38.65%-75.34%]
Specificity = 0.9	*Estimates from Survival Model for Progressive Cancer*					
	Age at Diagnosis	1.62	0.95	0.09	5.03 [0.79,32.22]	
	PSA(t)	0.48	0.81	0.55	1.62 [0.33,7.99]	

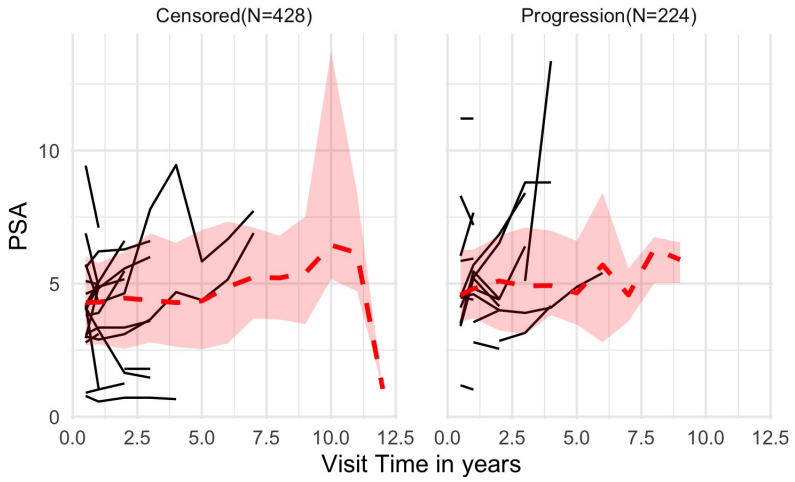


Figure 3.1. PSA trajectory in subgroups who were censored and whose cancers were detected to be progressive during surveillance. Black solid lines represent individual's PSA level across time for 40 randomly selected subjects. Red dashed line and shaded area represent median and interquartile range of PSA level over time among all subjects.

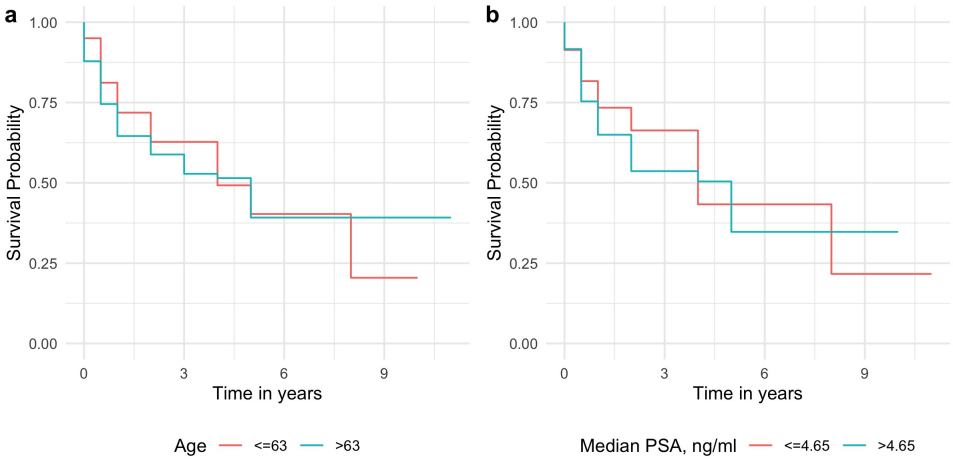


Figure 3.2. Turnbull's nonparametric estimates of survival probability for each subgroup.

APPENDIX A

SUPPLEMENT TO "FLEXIBLE, PARAMETRIC MIXTURE MODELS FOR TIME TO EVENT OUTCOMES, WITH INFLATION OF ZEROES AT BASELINE"

APPENDIX B

SUPPLEMENT TO "TIME TO FIRST POSITIVE DNA-PCR IN HIV-1 INFECTED, NON-BREASTFED INFANTS IN US COHORTS"

B.1 Parametric mixture model analysis to account for infants who test positive at birth

B.1.1 Methods

In a supplemental analysis, we fit a parametric mixture model to account for the subset of infants who tested positive at birth. In this analysis, we assumed that a proportion π of infants test positive at birth and a proportion $1 - \pi$ test positive after birth. In the latter group of infants, times of earliest DNA PCR positivity was modeled as a Weibull distribution. A similar approach has been applied in prior work on the sensitivity of DNA PCR assays in infants [47, 2]. The effects of maternal or infant ARV on π was modeled using a logistic function. In addition, the effects of maternal or infant ARV on the time to test positivity after birth was modeled using a Cox PH model. The statistical significance of maternal/infant ARV was estimated using LRT by comparing nested models with and without treatment as a predictor.

B.1.2 Results

To preserve statistical power, this analysis was limited to infants whose mothers received either No ARV, Single NRTI or cART (n=393). As in the previous analysis, there was no evidence of the violation of the PH assumption (LRT p value = 0.65). We observed a similar trend of delayed test positivity in the cART group (Supplemental Table B.5). Infants whose mothers were exposed to No ARV or to Single NRTI had a

significantly higher odds of a positive test at birth relative to infants whose mothers received cART, with an odds ratio (OR) of 4.29 (95% CI: 1.57 - 11.71) in the No ARV group and an OR of 5.53 (95% CI: 2.02 – 14.16) in the Single NRTI group. Among infants who did not test positive at birth, exposure to No ARV or to Single NRTI was associated with shorter times to test positivity relative to the group exposed to cART (Supplemental Table B.8).

Table B.1. Baseline characteristics by maternal antiretroviral regimen in WITS

Characteristic	Maternal Antiretroviral Regimen						
	no ARV, N = 3	Single NRTI, N = 0	ZDV+sdNVP, N = 10	sdNVP only, N = 6	2-3 NRTIs without sdNVP, N = 4	2-3 NRTIs with sdNVP, N = 8	cART, N = 98
Maternal CD4+ cell count, cells/ul							
0 to <200	1 (33%)	-	2 (20%)	2 (33%)	-	1 (14%)	11 (11%)
200 to <350	-	-	-	2 (33%)	1 (25%)	2 (29%)	18 (19%)
350 to <500	-	-	4 (40%)	-	-	2 (29%)	23 (24%)
>=500	2 (67%)	-	4 (40%)	2 (33%)	3 (75%)	2 (29%)	45 (46%)
Unknown	-	-	-	-	-	1	1
Mode of delivery							
Vaginal	-	-	5 (50%)	4 (67%)	2 (100%)	4 (67%)	34 (40%)
Cesarean Section before onset of labor and before membrane rupture	-	-	3 (30%)	2 (33%)	-	-	33 (39%)
Cesarean Section after onset of labor or membrane rupture	-	-	2 (20%)	-	-	2 (33%)	14 (17%)
Other	-	-	-	-	-	-	3 (3.6%)
Unknown	3	-	-	-	2	2	14
Gestational age, weeks							
0 to <37	1 (50%)	-	3 (30%)	3 (50%)	1 (25%)	3 (38%)	31 (32%)
>=37	1 (50%)	-	7 (70%)	3 (50%)	3 (75%)	5 (62%)	67 (68%)
Unknown	1	-	-	-	-	-	-
Birth weight, grams							
0 to <2,500	1 (50%)	-	3 (30%)	2 (33%)	1 (25%)	1 (12%)	19 (20%)
>=2,500	1 (50%)	-	7 (70%)	4 (67%)	3 (75%)	7 (88%)	75 (80%)
Unknown	1	-	-	-	-	-	4
Maternal viral load, copies/ml							
0 to <400	1 (33%)	-	4 (40%)	-	3 (75%)	2 (25%)	56 (57%)
400 to <1,000	1 (33%)	-	-	-	-	2 (25%)	9 (9.2%)
1,000 to <10,000	-	-	2 (20%)	2 (33%)	-	4 (50%)	18 (18%)
10,000 to 100,000	1 (33%)	-	3 (30%)	2 (33%)	1 (25%)	-	11 (11%)
>=100,000	-	-	1 (10%)	2 (33%)	-	-	4 (4.1%)

Table B.2. Baseline characteristics by maternal antiretroviral regimen in PACTS

Characteristic	Maternal Antiretroviral Regimen						
	no ARV, N = 195	Single NRTI, N = 89	ZDV+sdNVP, N = 0	sdNVP only, N = 0	2-3 NRTIs without sdNVP, N = 7	2-3 NRTIs with sdNVP, N = 0	cART, N = 8
Maternal CD4+ cell count, cells/ul							
0 to <200	24 (12%)	17 (19%)	-	-	3 (43%)	-	-
200 to <350	30 (15%)	22 (25%)	-	-	2 (29%)	-	4 (50%)
350 to <500	40 (21%)	23 (26%)	-	-	1 (14%)	-	2 (25%)
>=500	101 (52%)	27 (30%)	-	-	1 (14%)	-	2 (25%)
Mode of delivery							
Vaginal	149 (81%)	73 (82%)	-	-	5 (71%)	-	4 (67%)
Cesarean Section before onset of labor and before membrane rupture	6 (3.3%)	6 (6.7%)	-	-	1 (14%)	-	-
Cesarean Section after onset of labor or membrane rupture	29 (16%)	10 (11%)	-	-	1 (14%)	-	2 (33%)
Other	-	-	-	-	-	-	-
Unknown	11	-	-	-	-	-	2
Gestational age, weeks							
0 to <37	58 (31%)	34 (38%)	-	-	2 (29%)	-	1 (12%)
>=37	131 (69%)	55 (62%)	-	-	5 (71%)	-	7 (88%)
Unknown	6	-	-	-	-	-	-
Birth weight, grams							
0 to <2,500	78 (41%)	38 (43%)	-	-	2 (29%)	-	3 (38%)
>=2,500	111 (59%)	51 (57%)	-	-	5 (71%)	-	5 (62%)
Unknown	6	-	-	-	-	-	-
Maternal viral load, copies/ml							
0 to <400	22 (22%)	18 (30%)	-	-	2 (40%)	-	2 (40%)
400 to <1,000	1 (1.0%)	1 (1.6%)	-	-	-	-	-
1,000 to <10,000	27 (27%)	13 (21%)	-	-	-	-	1 (20%)
10,000 to 100,000	41 (41%)	20 (33%)	-	-	1 (20%)	-	2 (40%)
>=100,000	8 (8.1%)	9 (15%)	-	-	2 (40%)	-	-
Unknown	96	28	-	-	2	-	3

Table B.3. Baseline characteristics by infant antiretroviral regimen in WITS

Characteristic	Infant Antiretroviral Regimen		
	None, N = 98	ZDV, N = 28	Other, N = 3
Maternal CD4+ cell count, cells/ul			
0 to <200	14 (14%)	2 (7.4%)	1 (33%)
200 to <350	19 (20%)	3 (11%)	1 (33%)
350 to <500	23 (24%)	6 (22%)	-
>=500	41 (42%)	16 (59%)	1 (33%)
Unknown	1	1	-
Mode of delivery			
Vaginal	39 (50%)	9 (33%)	1 (33%)
Cesarean Section before onset of labor and before membrane rupture	21 (27%)	15 (56%)	2 (67%)
Cesarean Section after onset of labor or membrane rupture	15 (19%)	3 (11%)	-
Other	3 (3.8%)	-	-
Unknown	20	1	-
Gestational age, weeks			
0 to <37	28 (29%)	13 (46%)	1 (33%)
>=37	69 (71%)	15 (54%)	2 (67%)
Unknown	1	-	-
Birth weight, grams			
0 to <2,500	17 (18%)	9 (32%)	1 (33%)
>=2,500	76 (82%)	19 (68%)	2 (67%)
Unknown	5	-	-
Maternal viral load, copies/ml			
0 to <400	53 (54%)	12 (43%)	1 (33%)
400 to <1,000	8 (8.2%)	4 (14%)	-
1,000 to <10,000	14 (14%)	10 (36%)	2 (67%)
10,000 to 100,000	17 (17%)	1 (3.6%)	-
>=100,000	6 (6.1%)	1 (3.6%)	-

Table B.4. Baseline characteristics by infant antiretroviral regimen in PACTS

Characteristic	Infant Antiretroviral Regimen		
	None, N = 257	ZDV, N = 42	Other, N = 0
Maternal CD4+ cell count, cells/ul			
0 to <200	39 (15%)	5 (12%)	-
200 to <350	49 (19%)	9 (21%)	-
350 to <500	49 (19%)	17 (40%)	-
>=500	120 (47%)	11 (26%)	-
Mode of delivery			
Vaginal	200 (82%)	31 (74%)	-
Cesarean Section before onset of labor and before membrane rupture	8 (3.3%)	5 (12%)	-
Cesarean Section after onset of labor or membrane rupture	36 (15%)	6 (14%)	-
Other	-	-	-
Unknown	13	-	-
Gestational age, weeks			
0 to <37	78 (31%)	17 (40%)	-
>=37	173 (69%)	25 (60%)	-
Unknown	6	-	-
Birth weight, grams			
0 to <2,500	101 (40%)	20 (48%)	-
>=2,500	150 (60%)	22 (52%)	-
Unknown	6	-	-
Maternal viral load, copies/ml			
0 to <400	39 (28%)	5 (16%)	-
400 to <1,000	2 (1.4%)	-	-
1,000 to <10,000	33 (24%)	8 (25%)	-
10,000 to 100,000	52 (38%)	12 (38%)	-
>=100,000	12 (8.7%)	7 (22%)	-
Unknown	119	10	-

Table B.5. Number of infants who had at least one DNA PCR test administered by maternal ARV and age at the time of tests (days)

Maternal ARV	0 - 7 days	8 days - 14 days	15 days - 30 days	31 days - 90 days	91 days - 180 days	181 days - 365 days	366 days +
No ARV	113	25	36	26	0	0	0
Single NRTI	66	8	8	7	0	0	0
ZDV + sdNVP	6	3	2	7	2	2	0
sdNVP only	6	3	1	2	2	1	0
2-3 NRTIs without sdNVP	9	0	2	3	0	0	0
2-3 NRTIs with sdNVP	6	2	0	5	1	1	0
cART	92	20	13	50	8	2	1
Total number of DNA PCR tests	298	61	62	100	13	6	1

Table B.6. Number of infants who had at least one DNA PCR test administered by maternal ARV and age at the time of tests (days) in WITS/PACTS

Maternal ARV	0 - 7 days	8 days - 14 days	15 days - 30 days	31 days - 90 days	91 days - 180 days	181 days - 365 days	366 days +
No ARV	3/110	1/24	1/35	0/26	0/0	0/0	0/0
Single NRTI	0/66	0/8	0/8	0/7	0/0	0/0	0/0
ZDV + sdNVP + ZDV	6/0	3/0	2/0	7/0	2/0	2/0	0/0
sdNVP only	6/0	3/0	1/0	2/0	2/0	1/0	0/0
2-3 NRTIs without sdNVP	3/6	0/0	1/1	3/0	0/0	0/0	0/0
2-3 NRTIs with sdNVP	6/0	2/0	0/0	5/0	1/0	1/0	0/0
cART	85/7	20/0	13/0	49/1	8/0	2/0	1/0
Total number of DNA PCR tests	**109/189**	**29/32**	**18/44**	**66/34**	**13/0**	**6/0**	**1/0**

Table B.7. Number of DNA PCR tests per infant by cohort and by maternal ARV regimen

	Number of DNA PCR tests			
	1	2	3	4-6
Cohort				
PACTS	299	0	0	0
WITS	47	48	22	12
Maternal ARV regimen				
No ARV	196	2	0	0
Single NRTI	89	0	0	0
ZDV+sdNVP	3	2	2	3
sdNVP only	1	3	0	2
2-3 NRTIs without sdNVP	8	3	0	0
2-3 NRTIs with sdNVP	2	5	0	1
cART	47	33	20	6

Table B.8. Timing of earliest DNA PCR test positivity by type of maternal antiretroviral regimen from an unadjusted Logistic-Weibull PH mixture model. This analysis was restricted to infants whose mothers received one of the following: No ARV, Single NRTI or cART (n=393).

Maternal ARV	**Effect Estimate [95% CI]**
Association with positive DNA PCR test at birth	
No ARV	OR=4.29 [1.57-11.71]
Single NRTI	OR=5.53 [2.02-14.16]
cART	1
Association with time to the first positive DNA PCR test after birth	
No ARV	HR = 34.57 [11.72-101.99]
Single NRTI	HR = 12.71 [2.63-61.51]
cART	1

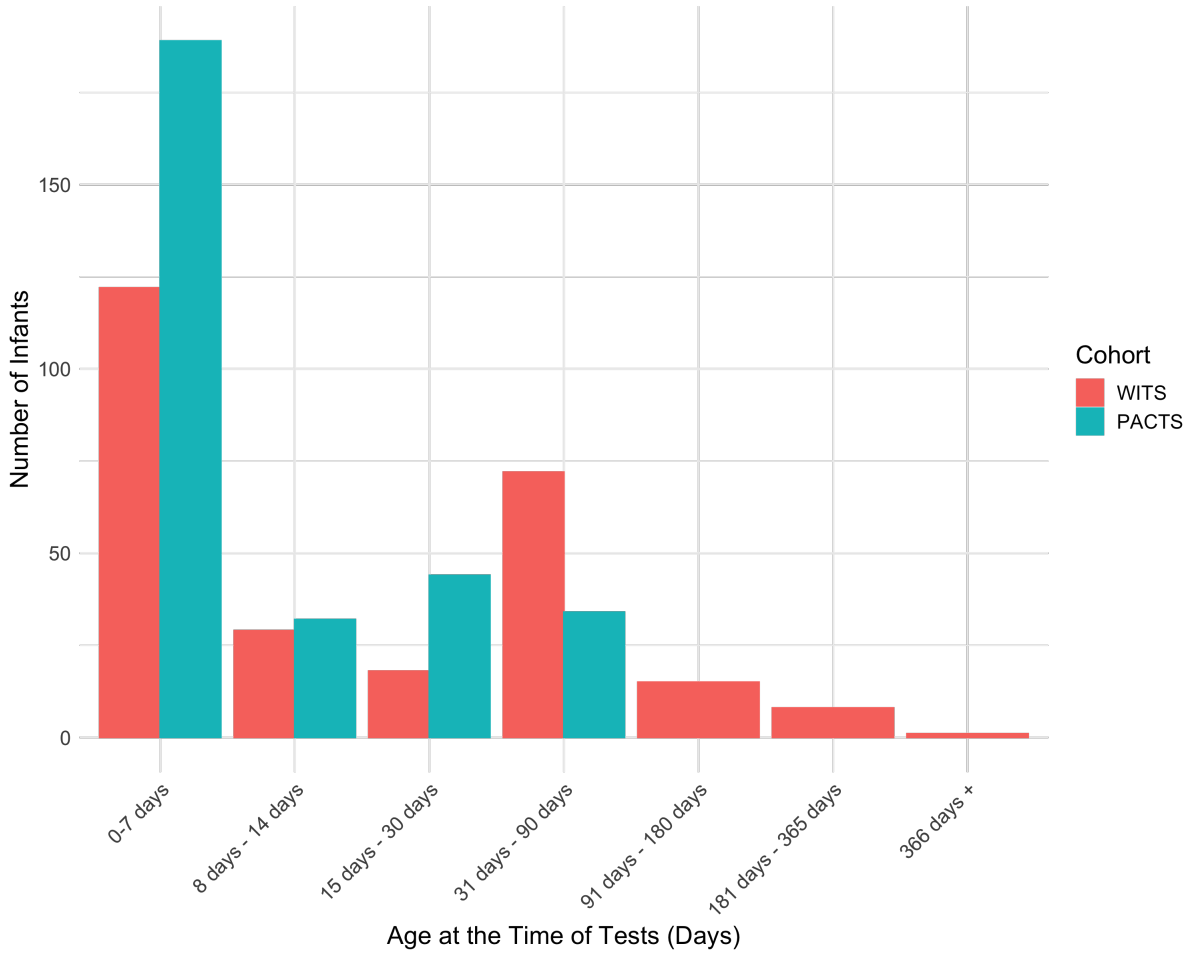


Figure B.1. Distribution of the number of infants who had at least one DNA PCR tests by age (days) and cohort.

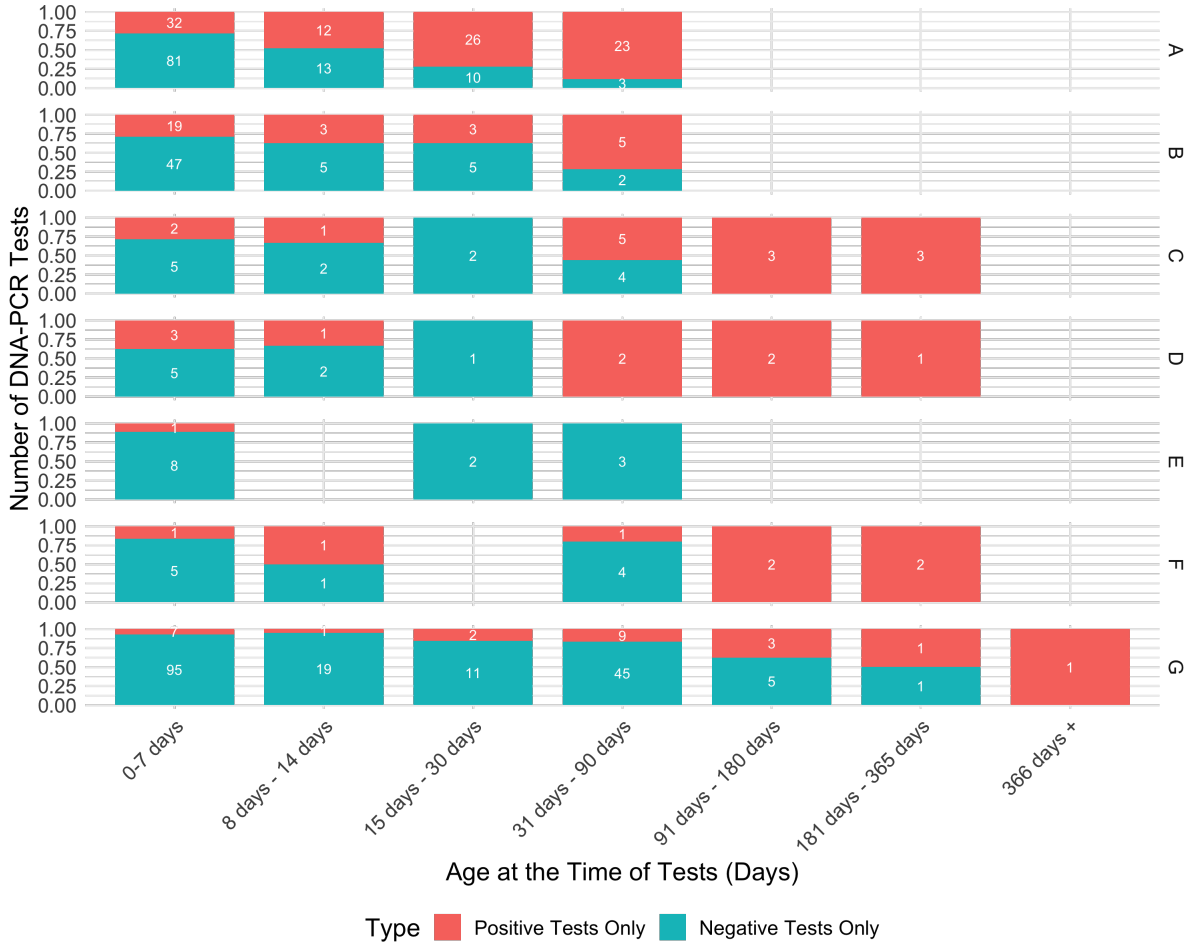


Figure B.2. Distribution of number of positive (red) and negative (blue) DNA PCR tests by age (days) and type of maternal antiretroviral regimen. A: no ARV; B: Single NRTI; C: ZDV + sdNVP; D: sdNVP only ; E: 2-3 NRTIs without sdNVP; F: 2-3 NRTIs with sdNVP; G: cART

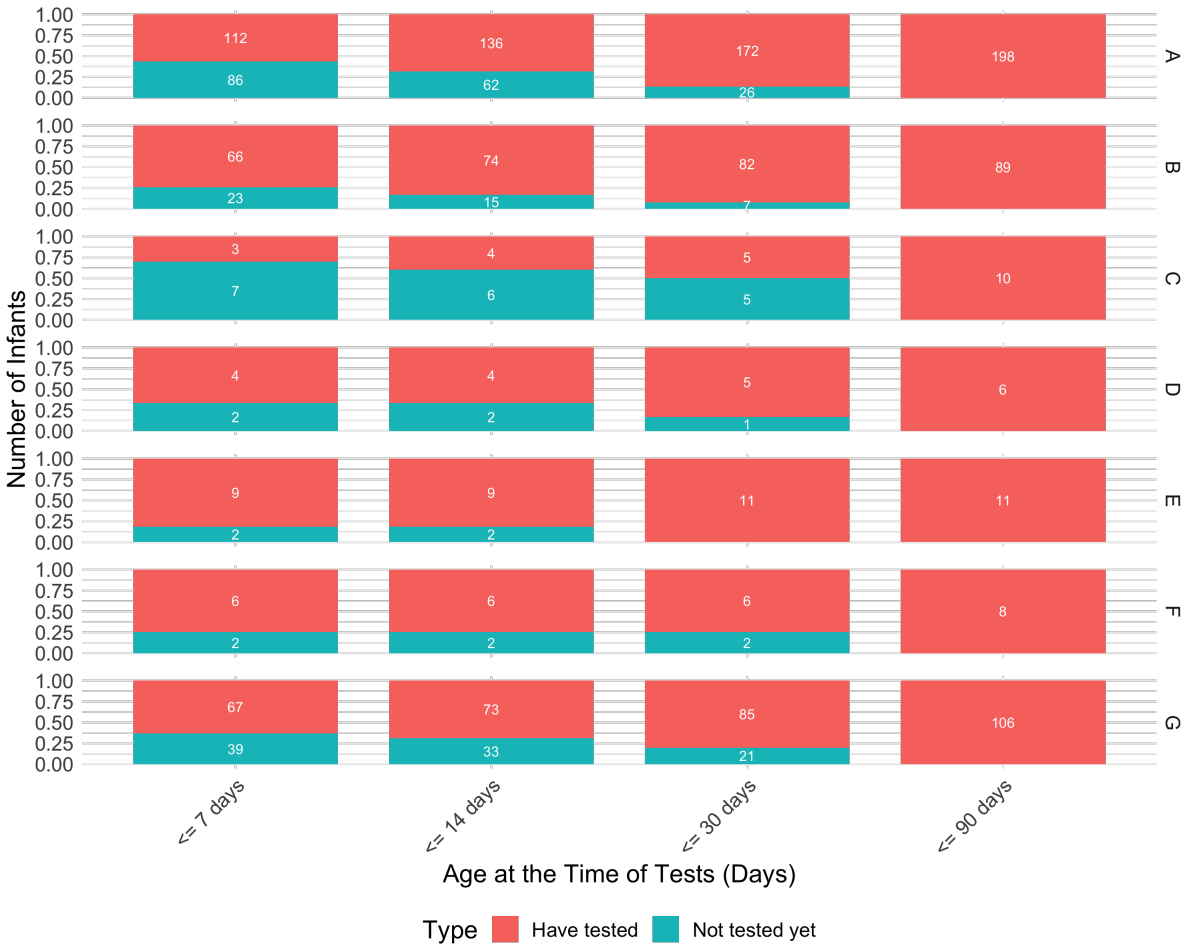


Figure B.3. Cumulative distribution of number of infants have had DNA PCR tests (red) and have tests (blue) by age (days) and type of maternal ARV.

A: no ARV; B: Single NRTI; C: ZDV + sdNVP; D: sdNVP only ; E: 2-3 NRTIs without sdNVP; F: 2-3 NRTIs with sdNVP; G: cART

APPENDIX C

SUPPLEMENT TO "A MIXTURE MODEL FOR ESTIMATING THE RISK OF PROSTATE CANCER PROGRESSION AND THE FRACTION OF INDOLENT CANCER IN ACTIVE SURVEILLANCE"

C.1 First Derivatives of the Log-Likelihood

C.1.1 Cox PH model with time-invariant covariates

The log-likelihood function is

$$\begin{aligned}\ell(\boldsymbol{y}^{obs};\boldsymbol{\alpha},\boldsymbol{\beta}) &= \sum_i \log\Bigg\{\eta D_{i1}S_1 + (1-\eta)\Bigg[\Big(1-\pi_i(\boldsymbol{\alpha})\Big)\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(\boldsymbol{Z}_i^\top\boldsymbol{\beta})} \\ &\quad +\pi_i(\boldsymbol{\alpha})D_{i(J+1)}\Bigg]\Bigg\}\end{aligned}$$

Let $\ell_i = log(L_i)$ be the log-likelihood function for subject i, where

$$L_i = \eta D_{i1}S_1 + (1-\eta)\Bigg[\Big(1-\pi_i(\boldsymbol{\alpha})\Big)\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(\boldsymbol{Z}_i^\top\boldsymbol{\beta})} + \pi_i(\boldsymbol{\alpha})D_{i(J+1)}\Bigg]$$

The first derivative of the log-likelihood function for subject i, ℓ_i, can be expressed as,

$$
\begin{aligned}
\frac{\partial \ell_i}{\partial \pi_i} &= \frac{1}{L_i}(1-\eta)\bigg(-\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta})} + D_{i(J+1)}\bigg) \\
\frac{\partial \pi_i}{\partial \alpha_d} &= \frac{\partial}{\partial \pi_i}\left(\frac{\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}{1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}\right) \\
&= \frac{1}{1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)} \exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d) X_{ij} - \exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d) \frac{\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}{(1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d))^2} X_{ij} \\
&= \pi_i X_{ij} - \pi_i^2 X_{ij} \\
&= X_{ij}(\pi_i - \pi_i^2) \\
\frac{\partial \ell_i}{\partial \alpha_d} &= \frac{\partial \ell_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial \alpha_d} \\
\frac{\partial \ell_i}{\partial S_j} &= \frac{1}{L_i}\frac{\partial L_i}{\partial S_j} \\
&= \frac{1}{L_i}(1-\eta)(1-\pi_i) D_{ij} \exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta})(S_j)^{\exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta})-1} \\
\frac{\partial \ell_i}{\partial \beta_k} &= \frac{1}{L_i}\frac{\partial}{\partial \beta_k}\bigg((1-\eta)(1-\pi_i)\sum_{j=1}^{J+1} D_{ij}(S_j)^{\exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta})}\bigg) \\
&= \frac{1}{L_i}(1-\eta)(1-\pi_i)\sum_{j=1}^{J+1} D_{ij}(\log S_j)(S_j)^{\exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta})} \exp(\boldsymbol{Z}_i^\top \boldsymbol{\beta}) Z_{ik}
\end{aligned}
$$

C.1.2 Cox PH model with time-varying covariates

The log-likelihood function is

$$
\begin{aligned}
\ell(\boldsymbol{y}^{obs}; \boldsymbol{\alpha}, \boldsymbol{\beta}) &= \sum_i \log\Bigg\{\eta D_{i1} + (1-\eta)\bigg[\Big(1-\pi_i(\boldsymbol{\alpha})\Big)\sum_{j=1}^{J+1} D_{ij}\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
&\quad + \pi_i(\boldsymbol{\alpha}) D_{i(J+1)}\bigg]\Bigg\}
\end{aligned}
$$

Let $\ell_i = log(L_i)$ be the log-likelihood function for subject i, where

$$
\begin{aligned}
L_i &= \eta D_{i1} + (1-\eta)\bigg[\Big(1-\pi_i(\boldsymbol{\alpha})\Big)\sum_{j=1}^{J+1} D_{ij}\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
&\quad + \pi_i(\boldsymbol{\alpha}) D_{i(J+1)}\bigg]
\end{aligned}
$$

The first derivative of the log-likelihood function for subject i, ℓ_i, can be expressed as,

$$
\begin{aligned}
\frac{\partial \ell_i}{\partial \pi_i} &= \frac{1}{L_i}(1-\eta)\bigg(-\sum_{j=1}^{J+1} D_{ij}\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) + D_{i(J+1)}\bigg) \\
\frac{\partial \pi_i}{\partial \alpha_d} &= \frac{\partial}{\partial \pi_i}\left(\frac{\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}{1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}\right) \\
&= \frac{1}{1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)X_{ij} - \exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)\frac{\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d)}{(1+\exp(\tilde{\boldsymbol{X}}^\top \boldsymbol{\alpha}_d))^2}X_{ij} \\
&= \pi_i X_{ij} - \pi_i^2 X_{ij} \\
&= X_{ij}(\pi_i - \pi_i^2) \\
\frac{\partial \ell_i}{\partial \alpha_d} &= \frac{\partial \ell_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial \alpha_d} \\
&\quad \text{Let } S_{ij} = \exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
\frac{\partial \ell_i}{\partial S_{ij}} &= \frac{1}{L_i}\frac{\partial L_i}{\partial S_{ij}} \\
&= \frac{1}{L_i}(1-\eta)(1-\pi_i)D_{ij} \\
\frac{\partial S_{ij}}{\partial \Lambda_l} &= S_{ij}\bigg(-\exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
&= -\exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
\frac{\partial S_{ij}}{\partial \beta_k} &= S_{ij}\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})P_{ik}(\tau_l)\bigg) \\
&= \exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg)\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})P_{ik}(\tau_l)\bigg) \\
\frac{\partial \ell_i}{\partial \Lambda_l} &= \sum_{j=1}^{J+1}\frac{\partial \ell_i}{\partial S_{ij}}\frac{\partial S_{ij}}{\partial \Lambda_l} \\
&= \frac{1}{L_i}(1-\eta)(1-\pi_i)\sum_{j=1}^{J+1} D_{ij}\bigg(-\exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg)\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg) \\
\frac{\partial \ell_i}{\partial \beta_k} &= \sum_{j=1}^{J+1}\frac{\partial \ell_i}{\partial S_{ij}}\frac{\partial S_{ij}}{\partial \beta_k} \\
&= \frac{1}{L_i}(1-\eta)(1-\pi_i)\sum_{j=1}^{J+1} D_{ij}\exp\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})\bigg)\bigg(-\sum_{l=0}^{j-2}\Lambda_l \exp(\boldsymbol{P}_i(\tau_l)^\top \boldsymbol{\beta})P_{ik}(\tau_l)\bigg)
\end{aligned}
$$

C.2 Data Exploration in Application

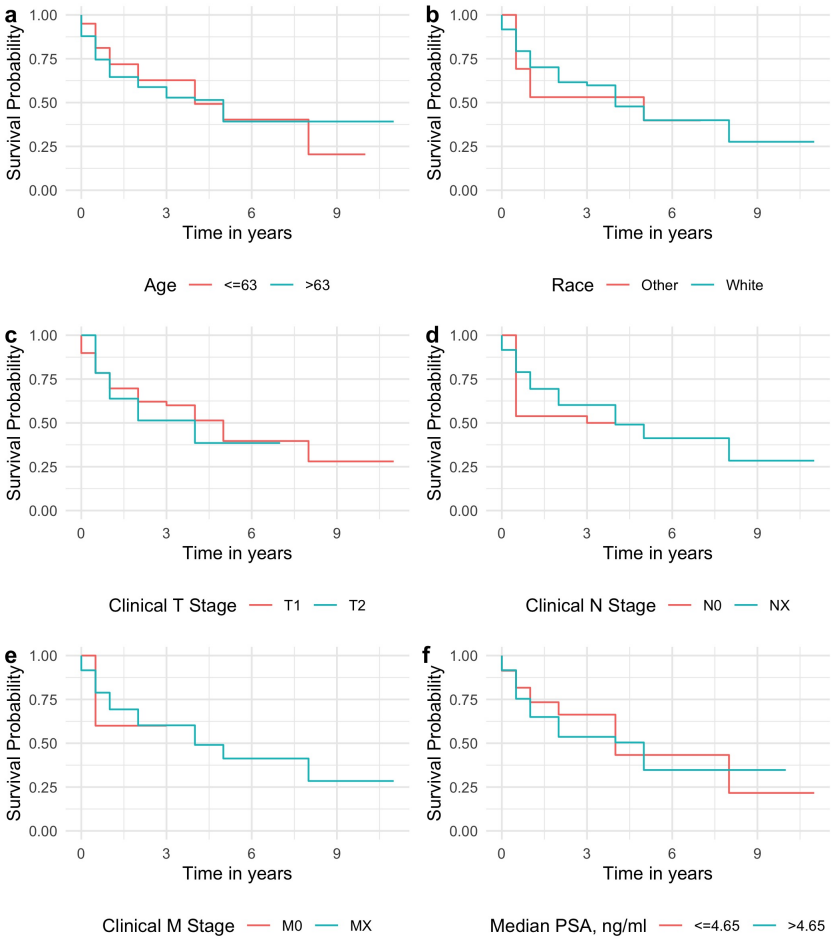


Figure C.1. Turnbull's nonparametric estimates of survival probability for each subgroup.

BIBLIOGRAPHY

[1] Altok, Muammer, Troncoso, Patricia, Achim, Mary F, Matin, Surena F, Gonzalez, Graciela N, and Davis, John W. Prostate cancer upgrading or downgrading of biopsy gleason scores at radical prostatectomy: prediction of "regression to the mean" using routine clinical features with correlating biochemical relapse rates. *Asian journal of andrology 21*, 6 (2019), 598.

[2] Balasubramanian, Raji, Fowler, Mary Glenn, Dominguez, Kenneth, Lockman, Shahin, Tookey, Pat A, Huong, Nicole Ngo Giang, Nesheim, Steven, Hughes, Michael D, Lallemant, Marc, Tosswill, Jennifer, et al. Time to first positive hiv-1 dna pcr may differ with antiretroviral regimen in infants infected with non-b subtype hiv-1. *AIDS (London, England) 31*, 18 (2017), 2465.

[3] Balasubramanian, Raji, and Lagakos, Stephen W. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika 90*, 1 (2003), 171–182.

[4] Banks, Harvey Thomas, Holm, Kathleen, and Robbins, Danielle. Standard error computations for uncertainty quantification in inverse problems: Asymptotic theory vs. bootstrapping. *Mathematical and computer modelling 52*, 9-10 (2010), 1610–1625.

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

[6] Burgard, Marianne, Blanche, Stéphane, Jasseron, Carine, Descamps, Philippe, Allemon, Marie-Christine, Ciraru-Vigneron, Nicole, Floch, Corinne, Heller-Roussin, Brigitte, Lachassinne, Eric, Mazy, Fabienne, et al. Performance of hiv-1 dna or hiv-1 rna tests for early diagnosis of perinatal hiv-1 infection during anti-retroviral prophylaxis. *The Journal of pediatrics 160*, 1 (2012), 60–66.

[7] Cassol, Sharon, Butcher, Ann, Kinard, Sharon, Spadoro, Joanne, Sy, Tidiane, Lapointe, Normand, Read, Stanley, Gomez, Perry, Fauvel, Micheline, and Major, Carol. Rapid screening for early detection of mother-to-child transmission of human immunodeficiency virus type 1. *Journal of clinical microbiology 32*, 11 (1994), 2641–2645.

[8] Chen, Chyong-Mei, Shen, Pao-sheng, and Huang, Wei-Lun. Semiparametric transformation models for interval-censored data in the presence of a cure fraction. *Biometrical Journal 61*, 1 (2019), 203–215.

[9] Cheung, Li C, Pan, Qing, Hyun, Noorie, Schiffman, Mark, Fetterman, Barbara, Castle, Philip E, Lorey, Thomas, and Katki, Hormuzd A. Mixture models for undiagnosed prevalent disease and interval-censored incident disease: applications to a cohort assembled from electronic health records. *Statistics in medicine 36*, 22 (2017), 3583–3595.

[10] Cooper, Ellen R, Charurat, Manhattan, Mofenson, Lynne, Hanson, I Celine, Pitt, Jane, Diaz, Clemente, Hayani, Karen, Handelsman, Edward, Smeriglio, Vincent, Hoff, Rodney, et al. Combination antiretroviral strategies for the treatment of pregnant hiv-1-infected women and prevention of perinatal hiv-1 transmission. *JAIDS-HAGERSTOWN MD- 29*, 5 (2002), 484–494.

[11] Cotton, Mark F, Violari, Avy, Otwombe, Kennedy, Panchia, Ravindre, Dobbels, Els, Rabie, Helena, Josipovic, Deirdre, Liberty, Afaaf, Lazarus, Erica, Innes, Steve, et al. Early time-limited antiretroviral therapy versus deferred therapy in south african infants infected with hiv: results from the children with hiv early antiretroviral (cher) randomised trial. *The Lancet 382*, 9904 (2013), 1555–1563.

[12] Dall'Era, Marc A, Albertsen, Peter C, Bangma, Christopher, Carroll, Peter R, Carter, H Ballentine, Cooperberg, Matthew R, Freedland, Stephen J, Klotz, Laurence H, Parker, Christopher, and Soloway, Mark S. Active surveillance for prostate cancer: a systematic review of the literature. *European urology 62*, 6 (2012), 976–983.

[13] De Angelis, R, Capocaccia, R, Hakulinen, T, Soderman, B, and Verdecchia, A. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in medicine 18*, 4 (1999), 441–454.

[14] Dunn, David T, Brandt, Carl D, Krivine, Anne, Cassol, Sharon A, Roques, Pierre, Borkowsky, William, De Rossi, Anita, Denamur, Erick, Ehrnst, Anneka, and Loveday, Clive. The sensitivity of hiv-1 dna polymerase chain reaction in the neonatal period and the relative contributions of intra-uterine and intra-partum transmission. *AIDS (London, England) 9*, 9 (1995), F7–11.

[15] Dunn, David T, Simonds, RJ, Bulterys, Marc, Kalish, Leslie A, Moye Jr, Jack, De Maria, Andrea, Kind, Christian, Rudin, Christoph, Denamur, Erick, Krivine, Anne, et al. Interventions to prevent vertical transmission of hiv-1: effect on viral detection rate in early infant samples. *Aids 14*, 10 (2000), 1421–1428.

[16] Efron, Bradley. Nonparametric standard errors and confidence intervals. *canadian Journal of Statistics 9*, 2 (1981), 139–158.

[17] Etzioni, Ruth, and Gulati, Roman. Recognizing the limitations of cancer overdiagnosis studies: a first step towards overcoming them. *JNCI: Journal of the National Cancer Institute 108*, 3 (2016).

[18] Fang, Hong-bin, Li, Gang, and Sun, Jianguo. Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics 32*, 1 (2005), 59–75.

[19] Farewell, Vernon T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* (1982), 1041–1046.

[20] Finkelstein, Dianne M. A proportional hazards models for interval-censored failure time data. *Biometrics 42* (1986), 845–854.

[21] for Disease Control, Centers, Prevention, et al. Enhanced perinatal surveillance-participating areas in the united states and dependent areas, 2000-2003. *HIV/AIDS surveillance supplemental report 13*, 4 (2008), 1–35.

[22] for Disease Control, Centers, Prevention, et al. Enhanced perinatal surveillance—15 areas, 2005-2008. *HIV Surveill Suppl Rep 16*, 2 (2011), 1–32.

[23] Frydman, Halina. A note on nonparametric estimation of the distribution function from interval-censored and truncated observations. *Journal of the Royal Statistical Society: Series B (Methodological) 56*, 1 (1994), 71–74.

[24] Gomez, Guadalupe, Calle, M Luz, Oller, Ramon, and Langohr, Klaus. Tutorial on methods for interval-censored data and their implementation in r.

[25] Gu, Xiangdong, and Balasubramanian, Raji. *straweib: Stratified Weibull Regression Model*, 2019. R package version 1.1.

[26] Gu, Xiangdong, Ma, Yunsheng, and Balasubramanian, Raji. Semiparametric time to event models in the presence of error-prone, self-reported outcomes—with application to the women's health initiative. *The annals of applied statistics 9*, 2 (2015), 714.

[27] Gu, Xiangdong, Shapiro, David, Hughes, Michael D, and Balasubramanian, Raji. Stratified weibull regression model for interval-censored data. *The R journal 6*, 1 (2014), 31.

[28] Haeri Mazanderani, Ahmad, Moyo, Faith, Kufa, Tendesayi, and Sherman, Gayle G. Declining baseline viremia and escalating discordant hiv-1 confirmatory results within south africa's early infant diagnosis program, 2010–2016. *J Acquir Immune Defic Syndr 77*, 2 (2018), 212–6.

[29] Hayes, Julia H, Ollendorf, Daniel A, Pearson, Steven D, Barry, Michael J, Kantoff, Philip W, Stewart, Susan T, Bhatnagar, Vibha, Sweeney, Christopher J, Stahl, James E, and McMahon, Pamela M. Active surveillance compared with initial treatment for men with low-risk prostate cancer: a decision analysis. *Jama 304*, 21 (2010), 2373–2380.

[30] He, Haijin, Han, Dongxiao, Song, Xinyuan, and Sun, Liuquan. Mixture proportional hazards cure model with latent variables. *Statistics in Medicine 40*, 29 (2021), 6590–6604.

[31] Inoue, Lurdes YT, Trock, Bruce J, Partin, Alan W, Carter, Herbert B, and Etzioni, Ruth. Modeling grade progression in an active surveillance study. *Statistics in medicine 33*, 6 (2014), 930–939.

[32] Irshad, Shazia, Bansal, Mukesh, Castillo-Martin, Mireia, Zheng, Tian, Aytes, Alvaro, Wenske, Sven, Le Magnen, Clémentine, Guarnieri, Paolo, Sumazin, Pavel, Benson, Mitchell C, et al. A molecular signature predictive of indolent prostate cancer. *Science translational medicine 5*, 202 (2013), 202ra122–202ra122.

[33] Jaspers, Stijn, Aerts, Marc, Verbeke, Geert, and Beloeil, Pierre-Alexandre. A new semi-parametric mixture model for interval censored data, with applications in the field of antimicrobial resistance. *Computational Statistics & Data Analysis 71* (2014), 30–42.

[34] Kapogiannis, Bill G, Soe, Minn M, Nesheim, Steven R, Abrams, Elaine J, Carter, Rosalind J, Farley, John, Palumbo, Paul, Koenig, Linda J, and Bulterys, Marc. Mortality trends in the us perinatal aids collaborative transmission study (1986–2004). *Clinical infectious diseases 53*, 10 (2011), 1024–1034.

[35] Klotz, Laurence, Vesprini, Danny, Sethukavalan, Perakaa, Jethava, Vibhuti, Zhang, Liying, Jain, Suneil, Yamamoto, Toshihiro, Mamedov, Alexandre, and Loblaw, Andrew. Long-term follow-up of a large active surveillance cohort of patients with prostate cancer. *J Clin oncol 33*, 3 (2015), 272–277.

[36] Krivine, Anne, Yakudima, Ahmed, Le May, Mireille, Pena-Cruz, Victor, Huang, Alice S, and McIntosh, Kenneth. A comparative study of virus isolation, polymerase chain reaction, and antigen detection in children of mothers infected with human immunodeficiency virus. *The Journal of pediatrics 116*, 3 (1990), 372–376.

[37] Kuk, Anthony YC, and Chen, Chen-Hsin. A mixture model combining logistic regression with proportional hazards regression. *Biometrika 79*, 3 (1992), 531–541.

[38] Lam, KF, and Xue, Hongqi. A semiparametric regression cure model with current status data. *Biometrika 92*, 3 (2005), 573–586.

[39] Lambert, John S, Harris, D Robert, Stiehm, E Richard, Moye, John, Fowler, Mary Glenn, Meyer III, William A, Bethel, James, Mofenson, Lynne M, et al. Performance characteristics of hiv-1 culture and hiv-1 dna and rna amplification assays for early diagnosis of perinatal hiv-1 infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes 34*, 5 (2003), 512–519.

[40] Li, Yi, Tiwari, Ram C, and Guha, Subharup. Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*, 3 (2007), 285–306.

[41] Lilian, Rivka R, Kalk, Emma, Bhowan, Kapila, Berrie, Leigh, Carmona, Sergio, Technau, Karl, and Sherman, Gayle G. Early diagnosis of in utero and intrapartum hiv infection in infants prior to 6 weeks of age. *Journal of clinical microbiology 50*, 7 (2012), 2373–2377.

[42] Lindsey, Jane C, and Ryan, Louise M. Methods for interval-censored data. *Statistics in medicine 17*, 2 (1998), 219–238.

[43] Lu, Wenbin, and Ying, Zhiliang. On semiparametric transformation cure models. *Biometrika 91*, 2 (2004), 331–343.

[44] Ma, Ling, Hu, Tao, and Sun, Jianguo. Cox regression analysis of dependent interval-censored failure time data. *Computational Statistics & Data Analysis 103* (2016), 79–90.

[45] Ma, Shuangge. Mixed case interval censored data with a cured subgroup. *Statistica Sinica* (2010), 1165–1181.

[46] Magliano, Dianna J, Islam, Rakibul M, Barr, Elizabeth L. M., Gregg, Edward W., Pavkov, Meda E., Harding, Jessica L., Tabesh, Maryam, Koye, Digsu N., and Shaw, Jonathan E. Trends in incidence of total or type 2 diabetes: systematic review.

[47] Mazanderani, AF Haeri, Du Plessis, Nicolette Marie, Thomas, Winifred Nancy, Venter, Elizabeth, and Avenant, Theunis. Loss of detectability and indeterminate results: Challenges facing hiv infant diagnosis in south africa's expanding art programme. *South African Medical Journal 104*, 8 (2014), 574–577.

[48] Mazanderani, Ahmad Haeri, and Sherman, Gayle G. Evolving complexities of infant hiv diagnosis within prevention of mother-to-child transmission programs. *F1000Research 8* (2019).

[49] Miloslavsky, Maja, Keleş, Sündüz, van der Laan, Mark J, and Butler, Steve. Recurrent events analysis in the presence of time-dependent covariates and dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*, 1 (2004), 239–257.

[50] Neelon, Brian, O'Malley, A James, and Smith, Valerie A. Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Statistics in Medicine 35*, 27 (2016), 5070–5093.

[51] Nelsen, Roger B. *An introduction to copulas*. Springer Science & Business Media, 2007.

[52] Nesheim, Steven, Palumbo, Paul, Sullivan, Kevin, Lee, Francis, Vink, Peter, Abrams, Elaine, and Bulterys, Marc. Quantitative rna testing for diagnosis of hiv-infected infants. *JAIDS Journal of Acquired Immune Deficiency Syndromes 32*, 2 (2003), 192–195.

[53] Newcomb, Lisa F, Brooks, James D, Carroll, Peter R, Feng, Ziding, Gleave, Martin E, Nelson, Peter S, Thompson, Ian M, and Lin, Daniel W. Canary prostate active surveillance study: design of a multi-institutional active surveillance cohort and biorepository. *Urology 75*, 2 (2010), 407–413.

[54] Newcomb, Lisa F, Thompson, Ian M, Boyer, Hilary D, Brooks, James D, Carroll, Peter R, Cooperberg, Matthew R, Dash, Atreya, Ellis, William J, Fazli, Ladan, Feng, Ziding, et al. Outcomes of active surveillance for clinically localized prostate cancer in the prospective, multi-institutional canary pass cohort. *The Journal of urology 195*, 2 (2016), 313–320.

[55] Newell, Marie-Louise, Loveday, Clive, Dunn, David, Kaye, Steve, Tedder, Richard, Peckham, Catherine, De Maria, Andrea, Giaquinto, Carlo, Omenaca, Felix, Canosa, Cipriano, et al. Use of polymerase chain reaction and quantitative antibody tests in children born to human immunodeficiency virus-1-infected mothers. *Journal of medical virology 47*, 4 (1995), 330–335.

[56] Nielsen-Saines, Karin, Watts, D Heather, Veloso, Valdilea G, Bryson, Yvonne J, Joao, Esau C, Pilotto, Jose Henrique, Gray, Glenda, Theron, Gerhard, Santos, Breno, Fonseca, Rosana, et al. Three postpartum antiretroviral regimens to prevent intrapartum hiv infection. *New england Journal of medicine 366*, 25 (2012), 2368–2379.

[57] Othus, Megan, Li, Yi, and Tiwari, Ram C. A class of semiparametric mixture cure survival models with dependent censoring. *Journal of the American Statistical Association 104*, 487 (2009), 1241–1250.

[58] Pack, Simon E, and Morgan, Byron JT. A mixture model for interval-censored time-to-response quantal assay data. *Biometrics* (1990), 749–757.

[59] Palisaar, Jüri R, Noldus, Joachim, Löppenberg, Björn, von Bodman, Christian, Sommerer, Florian, and Eggert, Thilo. Comprehensive report on prostate cancer misclassification by 16 currently used low-risk and active surveillance criteria. *BJU international 110*, 6b (2012), E172–E181.

[60] Peng, Yingwei, and Dear, Keith BG. A nonparametric mixture model for cure rate estimation. *Biometrics 56*, 1 (2000), 237–243.

[61] Persaud, Deborah, Ray, Stuart C, Kajdas, Joleen, Ahonkhai, Aima, Siberry, George K, Ferguson, Kimberly, Ziemniak, Carrie, Quinn, Thomas C, Casazza, Joseph P, Zeichner, Steven, et al. Slow human immunodeficiency virus type 1 evolution in viral reservoirs in infants treated with effective antiretroviral therapy. *AIDS research and human retroviruses 23*, 3 (2007), 381–390.

[62] PoToHDPaPoP, Transmission. Recommendations for the use of antiretroviral drugs during pregnancy and interventions to reduce perinatal hiv transmission in the united states.

[63] Prasitwattanaseree, Sukon, Lallemant, Marc, Costagliola, Dominique, Jourdain, Gonzague, and Mary, Jean-Yves. Influence of mother and infant zidovudine treatment duration on the age at which hiv infection can be detected by polymerase chain reaction in infants. *Antivir Ther 9*, 2 (2004), 179–185.

[64] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[65] Rocco, Bernardo, de Cobelli, Ottavio, Leon, Maria Elena, Ferruti, Mario, Mastropasqua, Mauro G, Matei, Deliu Victor, Gazzano, Giacomo, Verweij, Fabrizio, Scardino, Epifanio, Musi, Gennaro, et al. Sensitivity and detection rate of a 12-core trans-perineal prostate biopsy: preliminary report. *European urology 49*, 5 (2006), 827–833.

[66] Ross, AE, Loeb, S, Landis, P, et al. Re: Prostate-specific antigen kinetics during follow-up are an unreliable trigger for intervention in a prostate cancer surveillance program. *J Clin Oncol 28* (2010), 2810–6.

[67] Rotnitzky, Andrea, and Robins, James M. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika 82*, 4 (1995), 805–820.

[68] Sheon, Amy R, Fox, Harold E, Rich, Kenneth C, Stratton, Pamela, Diaz, Clemente, Tuomala, Ruth, Mendez, Hermann, Carrington, Jane, Alexander, Geraldine, Women, and Group, Infants Transmission Study. The women and infants transmission study (wits) of maternal-infant hiv transmission: study design, methods, and baseline data. *Journal of Women's Health 5*, 1 (1996), 69–78.

[69] Steyerberg, EW, Roobol, MJ, Kattan, MW, Van der Kwast, ThH, De Koning, HJ, and Schröder, FH. Prediction of indolent prostate cancer: validation and updating of a prognostic nomogram. *The Journal of urology 177*, 1 (2007), 107–112.

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

[71] Technau, Karl-Günter, Mazanderani, Ahmad Haeri, Kuhn, Louise, Hans, Lucia, Strehlau, Renate, Abrams, Elaine J, Conradie, Martie, Coovadia, Ashraf, Mbete, Ndileka, Murnane, Pamela M, et al. Prevalence and outcomes of hiv-1 diagnostic challenges during universal birth testing–an urban south african observational cohort. *Journal of the International AIDS Society 20* (2017), 21761.

[72] Turnbull, Bruce W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological) 38*, 3 (1976), 290–295.

[73] Violari, Avy, Cotton, Mark F, Gibb, Diana M, Babiker, Abdel G, Steyn, Jan, Madhi, Shabir A, Jean-Philippe, Patrick, and McIntyre, James A. Early antiretroviral therapy and mortality among hiv-infected infants. *New England Journal of Medicine 359*, 21 (2008), 2233–2244.

[74] Yatani, Ryuichi, Chigusa, Ichiro, Akazaki, Kaneyoshi, Stemmermann, Grant N, Welsh, Ronald A, and Correa, Pelayo. Geographic pathology of latent prostatic carcinoma. *International journal of cancer 29*, 6 (1982), 611–616.

[75] Young, Nancy L, Shaffer, Nathan, Chaowanachan, Thongpoon, Chotpitayasunondh, Tawee, Vanparapar, Nirun, Mock, Philip A, Waranawat, Naris, Chokephaibulkit, Kulkanya, Chuachoowong, Rutt, Wasinrapee, Punneeporn, et al. Early diagnosis of hiv-1-infected infants in thailand using rna and dna pcr assays sensitive to non-b subtypes. *Journal of acquired immune deficiency syndromes (1999) 24*, 5 (2000), 401–407.

[76] Zeng, Donglin, Mao, Lu, and Lin, DY. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika 103*, 2 (2016), 253–271.

[77] Zhang, Jiajia, and Peng, Yingwei. A new estimation method for the semiparametric accelerated failure time mixture cure model. *Statistics in medicine 26*, 16 (2007), 3157–3171.