INCORPORATING DIAGNOSTIC ACCURACY INTO THE ESTIMATION OF DISCRETE SURVIVAL FUNCTION

by

Abidemi K. Adeniji

M.S., University of Pittsburgh, 2008

B.S., University of Maryland, 2004

Submitted to the Graduate Faculty of

Graduate School of Public Health

in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Abidemi K. Adeniji

It was defended on

July 26, 2012

and approved by

Abdus S. Wahed, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Howard Rockette, Ph.D., Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Steven Belle, Ph.D., Professor, Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh

Jong-Hyeon Jeong, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Abdus S. Wahed, Ph.D., Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

INCORPORATING DIAGNOSTIC ACCURACY INTO THE ESTIMATION OF DISCRETE SURVIVAL FUNCTION

Abidemi K. Adeniji, PhD

University of Pittsburgh, 2012

The Empirical distribution function (EDF) is a commonly used estimator of the population cumulative distribution function. The Survival function is estimated as the complement of the EDF. However, the clinical diagnosis of an event is often subject to misclassification, by which the event is assessed with some uncertainty. In the presence of such errors, the true distribution of the time to first event is unknown. We develop a method to estimate the true survival distribution by incorporating negative predictive values (NPV) and positive predictive values (PPV), which are assumed to be known, into a product-limit style construction of a survival function. This allows us to quantify the bias of the EDF that do not account for misclassification due to the presence of misclassified events in the observed data. We present an unbiased estimator of the true survival function and its variance. In addition to dealing with misclassified clinical outcomes, this dissertation addresses survival function estimates in the presence of misclassified and incomplete data. The product limit (KM) estimator is commonly used to estimate the survival function when follow-up time is incomplete due to drop-outs. Typically this method is employed assuming that the outcome is known with certainty. We develop a method to estimate the true survival distribution by incorporating the NPV and PPV into a Kaplan-Meier-like construction. This allows us to quantify the bias in the KM survival estimates due to the presence of misclassified events in the observed data. We present an unbiased estimator of the true survival function and its variance. Asymptotic properties of the proposed estimators are provided and these properties are examined through simulations. We demonstrate our methods using data from the VIRAHEP-C study.

Estimating the true distribution of time to an event such as time to symptom resolution among subgroups of population with certain characteristics is of public health importance. When the event is measured with error, the actual distribution cannot be estimated without bias, providing an inaccurate picture of the population. The new methods provide clinical investigators with a tool to accurately estimate the survival probabilities in the presence of misclassified events.

Keywords: Misclassification, Measurement error, Diagnostic testing, Product limit estimation, Generalized estimating equations, Binary classification.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I wish to express sincere appreciation to Dr. Abdus S. Wahed for his support, vision, instruction, and for his vast reserve of knowledge and patience. The completion of this dissertation would have been impossible without his support, thank you.

It is a pleasure to thank my committee members, Dr. Howard Rockette, Dr. Steven Belle and Dr. Jong-Hyeon Jeong for their insightful critique of my work. Special gratitude to Dr. Howard Rockette for his belief in my capabilities as a graduate student, I thank you for your financial support at the very beginning of my doctoral studies.

I am grateful to my colleagues and friends for their encouragement and for their support. Most notably, I wish to offer a sincere gratitude to Jesse Y. Hsu for always being available to listen to my frustrations, thank you.

1.0 INTRODUCTION

In this chapter, we review some important concepts that will be repeatedly used in this dissertation. The topics that will be described briefly here are:

- Discrete survival analysis,
- Generalized estimating equations (GEE),
- Measures of diagnostic accuracy and
- VIRAHEP-C.

1.1 DISCRETE SURVIVAL ANALYSIS

Survival analysis is used to study the time until occurence of some event in a population. This time is called the survival time or failure time. The basic quantity employed to describe time-to-event is the survival function, the probability of an individual surviving beyond time x, that is, to experience the event after time x. This function is defined as

$$S(x) = Pr(X > x),$$

that is, the probability of not experiencing the event up to and including time x.

When X is a continuous random variable, the survival function is the complement of the cumulative distribution function, that is, $S(x) = 1 - F(x)$, where $F(x) = Pr(X \leq x)$. In addition, the survival function is the integral of the probability density function, $f(x)$, that is,

$$S(x) = Pr(X > x) = \int_x^\infty f(t)dt,$$

therefore,

$$f(x) = -\frac{dS(x)}{dx}.$$

Note that $f(x)d(x)$ may be thought of as the "approximate" probability that the event will occur at time x and that $f(x)$ is a nonnegative function with the area under $f(x)$ being equal to one [Klein and Moeschberger, 2003]. Many types of survival curves can be shown but the important point to note is that they all have the same basic properties. They are monotone, nonincreasing functions equal to one at zero and zero as time approaches infinity [Klein and Moeschberger, 2003].

Discrete time survival analysis is used when time is divided into discrete units or groups. Discrete time arises due to grouping of survival times into intervals, rounding off time measurements, or when lifetimes refer to an integral number of units. Often, survival times are grouped into discrete intervals of time (e.g. months). In this case, the length of time can be summarized using a set of positive integers (1,2,. . . .), that is, although the underlying process occurs in continuous time, the data are not observed in such form, hence, the data are summarized discretely rather than continously. Another reason for discrete time data is when the underlying process is intrinsically discrete. An example of an intrinsically discrete time process is that of fertility given by Jenkins (2008)- if one were interested in the duration from puberty to first birth, it might make sense to measure time in terms of the number of menstrual cycles rather than in terms of the number of months or days. Suppose that X, the number of menstrual cycles, can take values x_j, $j = 1, 2, \ldots$ with probability mass function (p.m.f) $p(x_j) = P(X = x_j)$, $j = 1, 2, \ldots$, where $x_1 < x_2 \ldots$. The survival function for a discrete random variable X is given by

$$S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j),$$

when X is discrete, the survival function is a nonincreasing step function.

1.2 KAPLAN-MEIER ESTIMATOR OF SURVIVAL FUNCTION

A common estimator of the survival function, particularly in the presence of censored data, proposed by Kaplan and Meier Kaplan and Meier [1958], is called the Product-Limit estimator. Censoring is an incomplete observance of an individual's survival time. This estimator is defined as follows for all values of x in the range where there are data:

$$\hat{S}(x) = \prod_{x_i \leq x} (1 - \frac{d_i}{Y_i})$$

$\hat{S}(x)$ is equal to 1 if $x < x_1$(the time to the first failure), with, at time x_i, Y_i observations, d_i failures, and probability of failure x_i, $\frac{d_i}{Y_i}$. The Product-Limit estimator is a step function with jumps at the observed event times. The size of these jumps depends not only on the number of events observed prior to and at each event time x_i, but also on the number of censored observations prior to x_i [Klein and Moeschberger, 2003].

1.3 GENERALIZED ESTIMATING EQUATIONS (GEE)

Generalized estimating equations (GEE) is a statistical method to estimate the marginal mean from longitudinal data [Liang and Zeger, 1986, Zeger and Liang, 1986]. The marginal mean is a mean response, it depends only on the covariates of interest, and not on any random effects or previous responses. Let $\mathbf{Y}_i = [Y_{i1}, \cdots, Y_{in_i}]^T$ be a $n_i \times 1$ vector of the outcome measurement for subject i and $E(\mathbf{Y}_i|\mathbf{X}_i) = \mathbf{X}_i\boldsymbol{\beta}$ be the marginal mean, where $\mathbf{X}_i^T = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}]$ is a $n_i \times p$ matrix of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, the generalized estimating equations is given by

$$\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \{\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}\} = \mathbf{0},$$

where $\mathbf{V}_i = \phi \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$; $\mathbf{R}_i(\boldsymbol{\alpha})$ is an $n_i \times n_i$ "working" correlation matrix specified by a $s \times 1$ vector $\boldsymbol{\alpha}$; $\mathbf{A}_i$ is an $n_i \times n_i$ diagonal matrix with $v_{im}(\mathbf{X}_i; \boldsymbol{\beta})$ as the m^{th} element, where $v_{im}(\mathbf{X}_i, \boldsymbol{\beta})$ is the assumed working variance function of Y_{im} and ϕ is the dispersion parameter for $m \in \{1, \cdots, n_i\}$. The solution of the generalized estimating equations, $\hat{\boldsymbol{\beta}}$, can be obtained through the iterative Gauss-Newton algorithm:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \left(\sum_{i=1}^{n} \mathbf{X}_i^T \tilde{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \tilde{\mathbf{V}}_i^{-1} \left\{ \mathbf{Y}_i - \mathbf{X}_i \hat{\boldsymbol{\beta}}^{(r)} \right\},$$

where $\tilde{\mathbf{V}}_i = \mathbf{V}_i[\hat{\boldsymbol{\beta}}^{(r)}, \hat{\boldsymbol{\alpha}}\{\hat{\boldsymbol{\beta}}^{(r)}, \hat{\phi}(\hat{\boldsymbol{\beta}}^{(r)})\}]$.

1.4 MEASURES OF DIAGNOSTIC ACCURACY

Two basic measures of diagnostic accuracy are sensitivity and specificity. Their definitions can be illustrated by a table with 2 rows and 2 columns, or decision matrix, where the rows summarize the data according to the true condition status of the patients and the columns summarize the test results. We denote the true condition status by the indicator variable D, where $D = 1$ if the condition is present and 0 if the condition is absent. Test results indicating that the condition is present are called positive; those indicating that the condition is absent are called negative. We denote positive test results as $T = 1$, negative test results as $T = 0$. Table 1 has such characteristics; it is called a count table because it indicates the numbers of participants in various categories. The numbers of participants with and without the condition, respectively, are denoted by, n_1 and n_0; the numbers of participants with the condition who test positive and negative are, respectively, s_1 and s_0; and the numbers of participants without the condition who test positive and negative, are respectively, r_1 and r_0. The total number of participants in the study group, N, is equal to $N = s_1 + s_0 + r_1 + r_0 = n_1 + n_0$.

The sensitivity (Se) of a test is its ability to detect the condition when it is present, so $Se = P(T = 1|D = 1)$, the probability that the test result is positive ($T = 1$), given that the condition is present ($D = 1$). Sensitivity is estimated by, $\hat{Se} = \frac{s_1}{n_1}$. The specificity (Sp) of a test is its ability to exclude the condition when it is not present, so $Sp = P(T = 0|D = 0)$, the probability that the test result is negative ($T = 0$), given that the condition is absent ($D = 0$). Specificity is estimated by, $\hat{Sp} = \frac{r_0}{n_0}$.

Table 1: Count Table

Basic 2x2 Count Table

	Test Result:		
True Condition Status:	Pos (T=1)	Neg (T=0)	total
Present(D=1)	s_1	s_0	n_1
Absent (D=0)	r_1	r_0	n_0
total	m_1	m_0	N

We now address two questions important to the main result of this dissertation: For a participant with a positive test result, what is the probability that the participant has the condition (positive predictive value, PPV)? For a participant with a negative test result, what is the probability that the patient does not have the condition (negative predictive value, NPV)? Note PPV and NPV depend not only on the sensitivity and specificity of the test respectively, but also on the probability of the condition [Xiao-Hua Zhou, 2002].

1.5 VIRAHEP-C STUDY

Chronic hepatitis C virus (HCV) is a cause of chronic liver disease and the most common indication for liver transplantation in the United States [CDC, 2012]. Population-based surveys indicate that 1.3% of the US population, approximately 3.2 million Americans, have chronic HCV infection, as shown by detection of antibodies to HCV (anti-HCV) and HCV RNA in serum[G.L. Armstrong, 2006]. Major advances have been made over the last decade in the field of antiviral therapy for chronic hepatitis C. Combination therapy with interferon and ribavirin has improved virological sustained response rates to nearly 40% in people with genotype 1 compared to only 10-15% for patients treated with interferon alone. Sustained response rates in African American patients appear to be significantly less than in Caucasian patients treated with the same regimens[John G. Mchutchison, 2000]. However, recently, the most exciting research aims to minimize side effects and increase virological sustained response rates by using combinations of Direct-acting antiviral agents (DAA) without interferon or ribavirin. The benefits of telaprevir plus peg-interferon and ribavirin over peg-interferon and ribavirin have have been demonstrated in the PROVE 1 + 2 trials as well as the ADVANCE trials [L. Y. Lee, 2012]. Telaprevir increased SVR up to 75% compared to about 44% with peg-interferon and ribavirin therapy [L. Y. Lee, 2012].

VIRAHEP-C was a multicenter, collaborative clinical trial, sponsored by NIDDK-NIH, designed to test the hypothesis that African Americans respond less well to antiviral therapy than Caucasian patients. A total of 196 African American and 205 Caucasian American treatment-naive patients with HCV genotype 1 infection were treated with peginterferon alfa-2a (180 microg/wk) and ribavirin (1000-1200 mg/day) for up to 48 weeks. The primary end point was sustained virologic response (SVR). SVR is defined as undetectable viral load 24 weeks after completion of antiviral therapy for chronic hepatitis C virus (HCV) infection.Clinical and virological data from these treatment cohorts were used to evaluate factors associated with resistance to antiviral therapy in African Americans and Caucasians with chronic hepatitis C. Pretreatment variables such as history of alcohol use, HCV RNA levels, hepatic histology, among others, were also investigated to determine which factors were associated with sustained virological response (SVR).

1.6 MOTIVATION AND AIMS

The primary objective of this dissertation is to incorporate correct classification probabilities (NPV and PPV) into a product limit estimator of the survival distribution and to examine its large-sample properties. Two specific aims are to:

1.6.1 Aim 1:

Derive an unbiased estimator of the true survival distribution in the presence of misclassified events and establish its large-sample properties, including an estimator for the asymptotic variance and to prove consistency and asymptotic normality of the derived estimator.

1.6.2 Aim 2:

Derive an unbiased estimator of the true survival distribution in the presence of misclassified events and incomplete data and establish its large-sample properties, including an estimator for the asymptotic variance.

1.6.2.1 Modify the KM estimator to account for misclassification.

1.6.2.2 Prove consistency and asymptotic normality of the derived estimator in Aim 1.6.2.1.

Standard methods in the analysis of survival data assume no error in classifying the event of interest. To estimate the time-to-event distribution, one widely used estimator is the Kaplan-Meier (KM) product limit estimator [Kaplan and Meier, 1958], Greenwood's formula can be used to estimate its variance [Greenwood, 1926]. However, when there is uncertainty in classifying the outcome of interest, the KM estimator may be a biased estimate of the true survival distribution.

One of the challenges in constructing an estimator of the time-to-event in the presence of misclassification is that the true survival distribution is latent (unseen). If a diagnostic tool is prone to error, then the true distribution of time to event is unobservable, all that we are provided is an error prone observance of the truth.

This dissertation is organized into two self-contained manuscripts. Each manuscript addresses one specifc aim, and are presented in Chapters 2 and 3 respectively. Chapter 4 offers some concluding thoughts and future directions.

2.0 DISCRETE SURVIVAL ANALYSIS WITH MISCLASSIFIED EVENTS

2.1 INTRODUCTION

Diagnoses of many clinical outcomes are given with ambiguity. Maladies such as the early stages of acute lymphocytic leukemia and pancreatic cancer often have little to no physical manifestations, so they may not be diagnosed in their earliest stages. On the other hand, inaccurate test results or symptoms that occur in multiple conditions may lead to a false diagnosis of an event.

Time-to-event data are common in epidemiologic studies. Standard techniques in survival analysis can handle fatal events as such events can always be classified correctly when observed. However, for nonfatal events, classification can be inaccurate. For example, the clinical diagnosis of Alzheimers disease is a complex process which includes, but is not limited to, eliminating all other causes of dementia. The clinical criteria for the diagnosis of Alzheimers disease involves the progressive deterioration of memory, cognitive skills and behavior (McKhann et al. 1984) . Thus, in the absence of an error-proof test, diagnoses may be made with error, in both directions. As a consequence, the time to such a misclassified event will not reflect the true time to event, Figure 1 provides an illustration.

The negative predictive value and the positive predictive value are the rates of correct classification of the diagnostic tool. The goals of this dissertation are to incorporate rates of classification and develop valid methods of inference for time-to-event data in the presence of misclassified events, to establish its large-sample properties, and to provide simulations that verify the large-sample properties in moderate samples.

An estimator of the population cumulative distribution function is the empirical distribution function (EDF). Studies have employed this method with several underlying assumptions including that the time to the outcome of interest is known without error. However, such is not always the case. Diagnostic tools that are perfectly accurate may be too costly with respect to time or money to routinely conduct, or may not even be available. So the question arises as to the accuracy of EDF estimates for estimating the survival function when the diagnostic test misclassifies a subject as having the outcome, when in truth he or she does not, or vice-versa. The work presented in this dissertation deals with the setting in which the observed data are prone to misclassification while the true status of the individual is unobservable. We will show that when the event is not accurately determined the EDF method leads to incorrect inferences. We will use the positive predictive value (PPV) and negative predictive value (NPV) of the diagnostic tool to construct a bridge between the observed and the unobserved distributions of outcomes. We will show that when the diagnostic tool used to measure failure is not perfect, it may lead to incorrect inferences. The bridging will lead to a product-limit estimator of the true survival distribution that can be recursively calculated using the NPV and PPV of the diagnostic tool and the observed survival distribution (events measured with error). The methods developed in this dissertation are applicable to studies investigating the incidence and timing of an event when the event is determined with uncertainty.

There have been prior investigations into the aforementioned problem. Racine-Poon (1984) offered a nonparametric estimation of the survival function that is analogous to the Kaplan-Meier approach for which the cause of death was uncertain. Their work assumed that the estimated probability of the risk of interest being the cause of death can be estimated without bias. Their endpoint, death, could be determined without error. The method does not consider the case of multiple endpoints nor does it consider the estimation of covariate effects. Snapinn (1998) offered methodology that dealt with scenarios for which a subject may experience a number of potential nonfatal endpoints. For example, a patient might experience several episodes of chest pain resembling myocardial infarctions, only some of which represent true myocardial infarctions. In the complex process of the diagnosis of myocardial infarction, there may be cases for which experts disagree on the diagnosis. To deal with uncertain endpoints, an adjudication (endpoint) committee may declare potential endpoints to have occured or not. A shortcoming is that the boundary for declaring true endpoints from false endpoints is somewhat arbitrary so different committees' final conclusions could differ. The other issue here is that when endpoints are classified as true or false, there is a loss of information in the level of certainty. To bypass the uncertainty in the declaration of a first true event, they proposed modifying the Cox proportional hazards regression model to incorporate information from all potential endpoints as well as the level of uncertainty. Although multiple endpoints were included in the model, the focus was on the time to the first true endpoint. Their method is similar to the ordinary Cox regression model in the sense that the focus is with the estimation and inference regarding the first true event only. A weight was given to each potential endpoint and incorporated into the modified Cox model; this weight represented the estimated likelihood that the corresponding potential endpoint is in fact the first true endpoint for a specific patient.

Richardson and Hughes (2000) expressed that low sensitivity or specificity of a diagnostic test results in biased estimates of the time to first event using product limit estimation. Working within the context of infectious diseases, they constructed two special cases. The first was a treatable self-limiting infectious disease. In such a case the disease can be resolved with treatment, for example, a sexually transmitted disease like gonorrhea. The important point from this situation is that no additional follow-up testing is needed since the disease is cured after the first positive test. The second case is an infectious disease that remains for the duration of the persons life; an example of this is HIV-1 infection. In such a scenario, the symptoms of the disease may be treatable but the disease itself is unresolved. Additional follow-ups after initial detection may be needed to verify disease status. They developed statistical methods to obtain unbiased estimates of the distribution of event times when the diagnostic test for the event has less than perfect sensitivity or specificity. Their method applies to cases in which all subjects are followed at discrete time points until their first positive test. The EM algorithm was used to obtain unbiased estimates of the conditional probability of disease for each specified time point. They introduced two EM algorithms: one for the case without follow-up after detection (treatable self-limiting infectious disease) and the other for the case with follow-up after detection (lifelong infectious disease). The methods produced ways to obtain less biased estimates of the cumulative distribution function of the time to first event when the outcome may be misclassified; they derived an EM algorithm for the product limit estimate of the survivor function.

Magder and Hughes (1997) incorporated information on the values of sensitivity and specificity into the estimation of the parameters in a logistic regression model. The regression coefficients and their standard errors were estimated using the Expectation Maximization (EM) algorithm. Neuhaus (1999) showed that ignoring errors in responses can lead to biased estimates of the associations of covariates with response. They derived general expressions for the magnitude of the bias in estimating the covariate effects due to errors in the response. They assumed that the true (unobserved) binary responses follow a binary regression model in the class of generalized linear models as described by McCullagh and Nelder (1989). The relationship between the unobserved truth and the observed error-prone responses are the response classification probabilities, namely sensitivity and specificity of the measurement. Even when the error probabilities were less than 0.1, the losses in efficiency were substantial. The paper derived the expressions for the magnitude of the bias in regression coefficients and efficiency loss due to errors in binary responses. Neuhaus showed that unless sensitivity and specificity are very high, the ignorance of errors in the response will yield highly biased covariate effect estimates. They quantified the magnitude of the bias due to errors in the response in terms of misclassification probabilities of the diagnostic test. They also showed that when the true error free response follows a generalized linear model with known misclassification probabilities, the observed responses also follow such a model with a modified link function. They showed that errors in the response lead to an increased standard error and a smaller test statistic, hence loss in estimation efficiency. From investigating the errors in response in settings with a single observation per subject, Neuhaus (2002) extended his work to investigate the effects of response misclassification on inference with clustered and longitudinal binary responses. Neuhaus (1999) investigated the effects of misclassification in settings with a single binary response per subject, Neuhaus (2002), some of the results were extended to the population-averaged model. In this work, the within-cluster covariance structure of the response is specified; it is assumed that the misclassification probabilities do not depend on the random effects and the observed responses follow a generalized linear model with a modified link function. In the cluster specific case, the association of the predictors to the response also depends on a modified link function. It was shown that ignoring the errors in response leads to substantially biased estimates of the associations of covariates with response. Expressions for the bias due to error in the response for both approaches were derived.

Meier et al. (2003) proposed an adjusted proportional hazards model (APH) that accurately estimates both cumulative survival and hazards ratios in the presence of misclassified outcomes. The performance of the APH method depends on the accuracy of the diagnostic test, namely sensitivity and specificity. Unlike the proportional hazards (PH) model which assumes perfect sensitivity and specificity, the APH model incorporates sensitivity and specificity of the diagnostic test in the estimation process. This model estimates the baseline cumulative survival and covariate effects by numerically maximizing the likelihood. Given accurate estimates of the tests' sensitivity and specificity, along with the caveat that a "reasonable" amount of data is available, the APH method provides unbiased estimates of cumulative survival as well as hazard ratios.

Balasubramanian and Lagakos (2001) developed regression methods for the distribution of the timing of perinatal HIV transmission. The gold standards for determining whether an infant is infected with HIV are the ELISA and Western Blot antibody tests; however imperfect diagnostic tests are often used because the ELISA and Western Blot antibody tests are only reliable when administered to infants beyond 18 months of age since infants can carry maternal antibodies for more than a year after birth. With no information on the true infection status of infants, but by assuming perfect specificity and time-dependent sensitivity, their method provided an estimator of the cumulative probability of perinatal transmission. All testing occurred after birth, thus, the period of exposure had ended. In their regression methods, data from different types of diagnostic tests can be utilized within the same analysis.

Retaining the context of sequentially-administered and error-prone diagnostic tests, Balasubramanian and Lagakos (2003) presented statistical methods for estimating the distribution of the time until an event in settings in which individuals can have different periods of exposure to the elements that place them at risk for HIV infection. The individuals could be infected with HIV in utero, at birth or from being breast fed. They developed a likelihood function and estimated the cumulative distribution function of the timing of vertical transmission of HIV through maximizing their proposed log likelihood. Also provided are approximate 95% pointwise confidence intervals based on a normal approximation using a bootstrap variance estimator.

Current status observation is a form of interval censoring; it refers to the situations in which the only available information on a survival random variable T is whether or not T exceeds a random independent monitoring time C (Jewell and van der Laan, 2002). McKeown and Jewell (2010) extended the nonparametric maximum likelihood estimator (NPMLE) of the distribution function underlying current status data when there is no misclassification to allow for time-dependent misclassification rates. They also extended their model to allow for misclassification rates that varied over time. Pointwise confidence intervals for the NPMLE were obtained through the use of a bootstrap method. Banerjee and Wellner (2005) provide further information.

McKeown and Jewell (2010) extended their ideas to the regression context. To deal with outcome misclassification, they adapted the techniques of binary generalized linear models (Neuhaus 1999). To adjust for errors in classification, they proposed that the observed outcome follows a generalized linear model with a modified link function. Rosas and Hughes (2010) extended these ideas by proposing nonparametric maximum likelihood estimator (NPMLE) of the distribution function of the failure time when sensitivity and specificity may vary among individuals or subgroups. Since the log likelihood function is concave with respect to the parameters of interest the modified iterative convex minorant (MICM) algorithm (Jongbloed 1998) was utilized to obtain an estimator of the distribution of failure time. Two sample hypothesis testing was discussed and a statistic to test for a difference in distribution functions was proposed. Since the data are subject to outcome misclassification the baseline hazard function was adjusted to obtain accurate estimation of the regression coefficients. The Cox proportional hazards model was adjusted to account for the errors in outcome classification. The estimation of regression parameters was achieved using an expectation maximization (EM) algorithm.

Event misclassification has been studied in frameworks slightly different from ours. In many randomized clinical trials, the primary outcome is the time to the first of a number of possible clinical events. Clinical endpoints such as disease progression in oncology trials are often determined with uncertainty. The conventional approach is to process uncertain cases through an endpoint adjudication committee. These cases are classified as true or false via a voting scheme and only the first confirmed endpoint for each patient is included in analysis, for instance, a Cox regression analysis. However, when interim analyses are performed on such trials, the final classifications for many of the reported events are unknown. Ignoring unconfirmed events may lead to incorrect statistical inference and analyses making use of all reported events are far more up-to-date than using only confirmed events (Cook and Kosorok, 2004) .

Cook and Kosorok (2004) studied the problem of event misclassification in the analysis of time-to-event data. They addressed the issue of incomplete adjudication in interim analysis. In many randomized clinical trials, the primary endpoint is the time to the first of a number of possible events. It is common in such studies for a selected set of study events, initially reported and classified by clinical investigators, to be reviewed by an event classification committee, whose role is to determine whether an event reported by an investigator is actually true. The use of an endpoint classification committee guarantees that criteria are uniformly applied to all reported events; however, it introduces additional delay between the time that an event is reported and the time that the final classification is known. This delay has implications for the timeliness of interim analysis of accumulating study data (Cook and Kosorok, 2004) .

They discussed that analyses that ignore events with incomplete adjudication or treat them as if they were confirmed events result in bias in the Kaplan-Meier estimates. They introduced methods to correctly address the statistical issues in the proper interpretation of interim data in the aforementioned setting; we explain the fundamental idea underlining their proposed methods. Suppose there is a dataset that contains a mixture of adjudicated and unadjudicated events. For a subject with a series of unadjudicated events, a particular event of interest is the first confirmed event provided that all earlier events are refuted. They considered the outcome of adjudication to be a binary random variable and assumed the outcomes are independent. Due to the fact that a subset of reported events underwent the adjudication process, estimates of these probabilities were obtained and applied to unadjudicated events. Thus events that were confirmed primary endpoint constituents were considered to have confirmation probability 1. In addition, subjects with no confirmed events were censored at the end of follow-up with probability equal to the probability that all reported events were refuted. They augmented the original dataset by randomly adjudicating all unadjudicated events according to the estimated probabilities. They then derived the asymptotic properties of the generalized Kaplan-Meier estimate of survival, parameter estimates under the Cox proportional hazards model, and a weighted generalized log-rank test. However variance estimates were more challenging because there were multiple and likely correlated events within each subject; that is, one individual may be likely to experience events of a given type than another individual; this phenomenon was captured using a frailty model. Also, the weights assigned to each event were estimated from the data. Therefore, standard martingale techniques were not applicable, and empirical process methods were required. An important result of their study is that complete adjudication may be unnecessary.

In this chapter, we propose a method to estimate the distribution of the true (latent) time to event by incorporating the NPV and PPV of the diagnostic test into the observed distribution of events in a product-limit-type construction. This estimator is a function of the NPV and the PPV of the diagnostic tool, which are assumed to be known. We conducted an extensive literature review, to the best of our knowledge, no research study has taken this approach in handling the issue of outcome misclassification. The other studies we reviewed have incorporated the sensitivity and specificity of the diagnostic tool into their estimation techniques. Our estimator differs from other research in that it uses NPV and PPV instead of sensitivity and specificity which are dependent on the distribution of events. For instance, sensitivity is the fraction of subjects that tested positive for an outcome out of all subjects that actually have the outcome. We under-take the task of estimating the true (latent) distribution of outcomes. This estimator is a function of the NPV and the PPV of the diagnostic tool, which are assumed to be known. An estimator of the variance of this estimator is also proposed. We assume that all participants are followed for the prespecified time period, and hence there is no dropout.

This chapter is organized as follows. We introduce notation, data, and assumptions in Section 2.2. In Section 2.3 we propose an estimate of the true survival function and derive its formulation. We also estimate the variance of our true survival rate estimator using the methods of M-estimators. We evaluate the large-sample properties of the proposed methods through simulations in Section 2.4. In Section 2.5 we appply our proposed methods to analyze data from the VIRAHEP-C study. We conclude our analysis with a discussion in Section 2.6. Section 2.7 provides a derivation of the estimating equations used in Section 2.3.

2.2 NOTATION, ASSUMPTIONS AND DATA

Define E_j as the occurence of an event at evaluation time $t_j, j = \{1, 2, \ldots K\}$. To delineate true (latent) events from potentially misclassified events we use the symbol '*' for true event; hence E_j^* is the true occurence of an event at time t_j and E_j is the potentially misclassified event at time t_j. Both E_j^* and E_j can take values 1 or 0, indicating the occurence or non-occurence of the event, respectively. Let θ_j and ϕ_j be respectively, the NPV and PPV of the evaluation process at time t_j. More specifically, let $\theta_j = P(E_j^* = 0 \mid E_j = 0)$ and $\phi_j = P(E_j^* = 1 \mid E_j = 1)$, where $P(.)$ denotes the probability of a specified event. Let T^* represent the time to the true event $\{E_j^* = 1\}$.

Let P_{jm}^* be the unconditional probability of the true event $\{E_j^* = m\}$ at time t_j, and P_{jm} be the unconditional probability of observing an event $\{E_j = m\}$ at time t_j, $m \in \{0, 1\}$; $j \in \{1, 2, \ldots, K\}$. The conditional probability of having the true event $\{E_j^* = n_j\}$ at time t_j given $\{E_{j-1}^* = n_{j-1}\}, \{E_{j-2}^* = n_{j-2}\}, \ldots, \{E_1^* = n_1\}$, that is $P(E_j^* = n_j \mid E_{j-1}^* = n_{j-1}, ..., E_1^* = n_1)$, is abbreviated as $P_{jn_j|(j-1)n_{j-1},..,1n_1}^*, n_k \in \{0, 1\}, k \in$

$\{1, 2, \ldots, j\}; j \in \{1, 2, \ldots, K\}$. The same shorthand notation follows for the observed events, namely, $P(E_j = n_j \mid E_{j-1} = n_{j-1}, \ldots, E_1 = n_1) = P_{jn_j|(j-1)n_{j-1},\ldots,1n_1}$.

Our development is based on several assumptions. First, θ_j and ϕ_j do not change over time, hence the probability of missclassification is constant across time points and individuals. Thus, we write

$$\theta_j = \theta, \text{ and } \phi_j = \phi, \ \forall j = 1, 2, \ldots, K. \tag{2.1}$$

In addition, given an observed event E_j at time t_j; the probability of the true outcome E_j^* can be ascertained without the knowledge of the previous true events. That is, $P(E_j^* \mid E_j, E_k^*, k = 1, 2, \ldots, j-1) = P(E_j^* \mid E_j)$. Thus, for example,

$$P(E_j^* = 0 \mid E_j = 0, E_{j-1}^* = 0, E_{j-2}^* = 0, \ldots, E_1^* = 0) = P(E_j^* = 0 \mid E_j = 0) = \theta, \tag{2.2}$$

$$P(E_j^* = 0 \mid E_j = 1, E_{j-1}^* = 0, E_{j-2}^* = 0, \ldots, E_1^* = 0) = P(E_j^* = 0 \mid E_j = 1) = 1 - \phi. \tag{2.3}$$

Furthermore, given observed and true outcomes at all previous $(j-1)$ time points, the probability of an observed outcome E_j at time t_j does not depend on the previous $(j-1)$ true outcomes. This assumption basically states that the estimation of the observed probabilities does not depend on the true status at previous time points. More specifically,

$$P(E_j \mid E_k, E_k^*, k = 1, 2, \ldots, j-1) = P(E_j \mid E_k, k = 1, 2, \ldots, j-1). \tag{2.4}$$

We consider only the first occurences of the event. Once the event occurs, follow-up ends. Thus, $P(E_j^* < E_{j-1}^*) = P(E_j < E_{j-1}) = 0 \ \forall j = 1, 2, \ldots, K$. In other words, $P(E_j = 0 \mid E_{j-1} = 1) = 0$, and $P(E_j^* = 0 \mid E_{j-1}^* = 1) = 0$. As mentioned, our goal is to estimate the survival distribution of the time to the first event (T^*) in the latent population in which the subjects are evaluated for the event at K fixed time points $t_1, t_2, \ldots, t_K$. Thus, the goal is to estimate $S(t_K) = P(T^* > t_K) = 1 - P(T^* \leq t_K)$. The observed data in this setting consist of a set of n identically distributed random vectors $\widetilde{E}_i = (E_{i1}, E_{i2}, \ldots, E_{iK}), i = 1, 2, \ldots, n$, where $E_{ij} = 1$ if a potentially misclassified event is observed at time t_j for subject i, 0, otherwise. Note that E_{ij}'s satisfy the following: $0 \leq E_{ij} \leq E_{i(j+1)} \leq 1, j \in \{1, 2, \ldots, K-1\}$ and $0 \leq \sum_{j=1}^{K} E_{ij} \leq 1$. The empirical distribution estimate of S(.) is defined as $\hat{S}(t_k) = 1 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} E_{ij}, k = 1, 2, \ldots K$.

2.3 INFERENCE FROM COMPLETE DATA: NO DROP-OUT

We first prove the following lemma necessary to develop our estimator for the survival distribution of T^*.

Lemma 2.3.1. *Under assumptions (2.1)-(2.4), and the notation described in section 2.2,*

$$P^*_{j0|(j-1)0,\ldots,10} = \theta\zeta^{(j-1)} + (1-\phi)(1-\zeta^{(j-1)}), j = 1, \ldots, K, \tag{2.5}$$

where,

$$\zeta^{(j)} = \theta \prod_{k=1}^{j} \frac{P_{k0|(k-1)0,\ldots,10}}{P^*_{k0|(k-1)0,\ldots,10}} \times P_{(j+1)0|j0,\ldots,10}, j = 1, \ldots, K-1.$$

and $\zeta^{(0)} = P_{10}$.

We outline the proof of Lemma 2.3.1 below.

Suppose follow-up starts at time t_0. At this time, no subjects will have an event, thus $P^*_{01} = P(T^* \leq t_0) = P(E^*_0 = 1) = 0$. The probability that an event does not occur by time t_1 is

$$\begin{aligned} P^*_{10} = P(E^*_1 = 0) &= P(E^*_1 = 0 \mid E_1 = 0)P(E_1 = 0) + P(E^*_1 = 0 \mid E_1 = 1)P(E_1 = 1) \\ &= \theta P(E_1 = 0) + (1-\phi)P(E_1 = 1) \\ &= \theta P_{10} + (1-\phi)P_{11}. \end{aligned} \tag{2.6}$$

The probability that a true event does not occur at time t_2 given that the event did not happen at the first time t_1,

$$
\begin{aligned}
P^*_{20|10} &= P(E^*_2 = 0 \mid E^*_1 = 0) \\
&= P(E^*_2 = 0 \mid E_2 = 0, E^*_1 = 0)P(E_2 = 0 \mid E^*_1 = 0) \\
&\quad + P(E^*_2 = 0 \mid E_2 = 1, E^*_1 = 0)P(E_2 = 1 \mid E^*_1 = 0).
\end{aligned}
$$

By assumptions (2) and (3),

$$
\begin{aligned}
P^*_{20|10} &= P(E^*_2 = 0 \mid E_2 = 0)P(E_2 = 0 \mid E^*_1 = 0) \\
&\quad + P(E^*_2 = 0 \mid E_2 = 1)P(E_2 = 1 \mid E^*_1 = 0) \\
&= \theta P(E_2 = 0 \mid E^*_1 = 0) \\
&\quad + (1 - \phi)P(E_2 = 1 \mid E^*_1 = 0) \qquad (2.7)
\end{aligned}
$$

Our main goal is to express each of the terms on the right hand side of equation (2.7) in terms of θ, ϕ, and observed marginal and transitional probabilities so that they can be estimated from the data given θ and ϕ. We first work with the term $P(E_2 = 0 \mid E^*_1 = 0)$. Conditioning on $\{E^*_1 = 0\}$,

$$
\begin{aligned}
P(E_2 = 0 \mid E^*_1 = 0) &= P(E_2 = 0 \mid E_1 = 0, E^*_1 = 0) \times P(E_1 = 0 \mid E^*_1 = 0) \\
&\quad + P(E_2 = 0 \mid E_1 = 1, E^*_1 = 0) \times P(E_1 = 1 \mid E^*_1 = 0).
\end{aligned}
$$

By assumption(4),

$$
\begin{aligned}
P(E_2 = 0 \mid E^*_1 = 0) &= P(E_2 = 0 \mid E_1 = 0) \times \frac{P(E^*_1 = 0 \mid E_1 = 0) \times P(E_1 = 0)}{P(E^*_1 = 0)} \\
&\quad + P(E_2 = 0 \mid E_1 = 1) \times \frac{P(E^*_1 = 0 \mid E_1 = 1) \times P(E_1 = 1)}{P(E^*_1 = 0)}, \\
&= \frac{P_{20|10}\theta P_{10}}{P^*_{10}}, \qquad (2.8)
\end{aligned}
$$

where the second term in the second-to-last line in equation (2.8) is zero since, $P(E_2 = 0 \mid E_1 = 1) = 0$. Thus, from equation (2.7),

$$
P^*_{20|10} = P(E^*_2 = 0 \mid E^*_1 = 0) = \theta\zeta^{(1)} + (1 - \phi)(1 - \zeta^{(1)}),
$$

where $\zeta^{(1)} = \theta P_{20|10}(\frac{P_{10}}{P^*_{10}})$.

Now, the probability that a true event does not occur at time t_3 given that the event did not occur at the earlier times,

$$\begin{aligned} P^*_{30|20,10} &= P(E^*_3 = 0 \mid E_3 = 0, E^*_2 = 0, E^*_1 = 0)P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0) \\ &\quad + P(E^*_3 = 0 \mid E_3 = 1, E^*_2 = 0, E^*_1 = 0)P(E_3 = 1 \mid E^*_2 = 0, E^*_1 = 0). \end{aligned}$$

By assumptions (2) and (3),

$$\begin{aligned} P^*_{30|20,10} &= P(E^*_3 = 0 \mid E_3 = 0)P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0) \\ &\quad + P(E^*_3 = 0 \mid E_3 = 1)P(E_3 = 1 \mid E^*_2 = 0, E^*_1 = 0) \\ &= \theta P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0) \\ &\quad + (1 - \phi)(E_3 = 1 \mid E^*_2 = 0, E^*_1 = 0). \end{aligned} \tag{2.9}$$

Again, our main goal is to express each of the terms on the right hand side of equation (2.9) in terms of θ, ϕ, and observed marginal and transitional probabilities. Let us consider the term $P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0)$. Conditioning on $\{E^*_2 = 0\}$ and $\{E^*_1 = 0\}$,

$$\begin{aligned} P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0) &= P(E_3 = 0, E_2 = 0, E_1 = 0 \mid E^*_2 = 0, E^*_1 = 0) \\ &= P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0, E_2 = 0, E_1 = 0) \\ &\quad \times P(E_2 = 0 \mid E^*_2 = 0, E^*_1 = 0, E_1 = 0) \\ &\quad \times P(E_1 = 0 \mid E^*_2 = 0, E^*_1 = 0) \end{aligned}$$

By assumption (4),

$$\begin{aligned} P(E_3 = 0 \mid E^*_2 = 0, E^*_1 = 0) &= P(E_3 = 0 \mid E_2 = 0, E_1 = 0) \\ &\quad \times \frac{P(E^*_2 = 0, E^*_1 = 0 \mid E_2 = 0, E_1 = 0)P(E_2 = 0, E_1 = 0)}{P(E^*_2 = 0, E^*_1 = 0, E_1 = 0)} \\ &\quad \times \frac{P(E^*_2 = 0, E^*_1 = 0, E_1 = 0)}{P(E^*_2 = 0, E^*_1 = 0)} \\ &= \frac{P_{30|20,10}\theta P_{20|10}P_{10}}{P^*_{20|10}P^*_{10}} \end{aligned}$$

Thus, from equation (2.9),

$$P^*_{30|20,10} = \theta\zeta^{(2)} + (1-\phi)(1-\zeta^{(2)}).$$

where $\zeta^{(2)} = \frac{\theta P_{30|20,10}P_{20|10}P_{10}}{P^*_{20|10}P^*_{10}} = \theta P_{30|20,10}(\frac{P_{20|10}P_{10}}{P^*_{20|10}P^*_{10}})$.
Continuing the same way we can write,

$$P^*_{j0|(j-1)0,\ldots,10} = \theta\zeta^{(j-1)} + (1-\phi)(1-\zeta^{(j-1)}),$$

where

$$\zeta^{(j)} = \theta\prod_{k=1}^{j}\frac{P_{k0|(k-1)0,\ldots,10}}{P^*_{k0|(k-1)0,\ldots,10}} \times P_{(j+1)0|j0,\ldots,10}, j = 1, \ldots, K-1.$$

and $\zeta^{(0)} = P_{10}$.
With the true conditional probabilities expressed as a function of θ, ϕ, and observed probabilities, we are ready to state the main result.

Theorem 1. *Under assumptions (1)-(4), the probability of having a true event by time t_j can be expressed as:*

$$P^*_{j1} = P^*_{(j-1)1} + P^*_{j1|(j-1)0,\ldots,10} \times \prod_{l=1}^{j} P^*_{(l-1)0|(l-2)0,\ldots,10}, j = 1, 2, \ldots, K. \tag{2.10}$$

Proof. (By Mathematical induction)
First, we note that Equation (2.10) holds for $j = 1$, since $P^*_{11} = P^*_{01} + P^*_{11|00}P^*_{00}$. This follows from the fact that at the start of follow-up everyone is event free, and hence $P^*_{01} = 0 = 1 - P^*_{00}$. Assume that equation (2.10) is true for j=m. Now, the probability that there will be a true event by $t_{(m+1)}$ is

$$\begin{aligned} P^*_{(m+1)1} &= P(T^* \leq t_{m+1}) \\ &= P(T^* \leq t_m) + P(t_m < T^* \leq t_{m+1}) \\ &= P^*_{m1} + (1 - P^*_{m1})P(E^*_{m+1} = 1 \mid E^*_j = 0, j = 1, 2, \ldots, m) \\ &= P^*_{m1} + P^*_{m0}P^*_{(m+1)1|m0,\ldots,00}. \end{aligned}$$

$$\begin{aligned}
P^*_{m0} &= P(E^*_j = 0 \ \forall j = 0, 1, 2, \ldots, m) \\
&= P(E^*_0 = 0)P(E^*_1 = 0 \mid E^*_0 = 0)P(E^*_2 = 0 \mid E^*_1 = 0, E^*_0 = 0) \times \cdots \times \\
&\quad \times P(E^*_m = 0 \mid E^*_{m-1} = 0, \ldots, E^*_0 = 0) \\
&= P^*_{00}P^*_{10|00}P^*_{20|10,00} \cdot \ldots \cdot P^*_{m0|(m-1)0,\ldots,00} \\
&= \prod_{l=1}^{m} P^*_{l0|(l-1)0,\ldots,00} \\
\therefore P^*_{(m+1)1} &= P^*_{m1} + \left\{ \prod_{l=1}^{m} P^*_{l0|(l-1)0,\ldots,00} \right\} P^*_{(m+1)1|m0,\ldots,00}.
\end{aligned}$$

This completes the proof.

□

Theorem 1 along with Lemma 2.3.1, provides the necessary tools to estimate the survival distribution of T^* at the evaluation times $t_1, t_2, t_3, \ldots, t_K$. Note from equation (2.6) that $P^*_{11} = 1 - \{\theta P_{10} + (1-\phi)P_{11}\}$. P_{11} is the probability that a potentially misclassified event is observed at evaluation time t_1. Therefore, $\hat{P}_{11} = \frac{\sum_{i=1}^n E_{i1}}{n}$. Hence,

$$\hat{P}^*_{11} = \frac{\sum_{i=1}^n \{1 - [\theta(1-E_{i1}) + (1-\phi)E_{i1}]\}}{n}.$$

To estimate P^*_{21}, we start with the recursive formula $P^*_{21} = P^*_{11} + P^*_{10}P^*_{21|10}$. By equation (2.7), $P^*_{21|10} = 1 - P^*_{20|10} = 1 - \theta\zeta^{(1)} - (1-\phi)(1-\zeta^{(1)})$, where $\zeta^{(1)} = \frac{\theta P_{20|10}P_{10}}{P^*_{10}}$. Now, a simple estimate of $P_{20|10}$ is given by $\hat{P}_{20|10} = \frac{\sum_{i=1}^n (1-E_{i2})}{\sum_{i=1}^n (1-E_{i1})}$. Therefore,

$$\hat{P}^*_{21} = \hat{P}^*_{11} + \hat{P}^*_{10}\hat{P}^*_{21|10},$$

where $\hat{P}^*_{21|10} = 1 - \theta\hat{\zeta}^{(1)} - (1-\phi)(1-\hat{\zeta}^{(1)})$ with $\hat{\zeta}^{(1)} = \frac{\theta\hat{P}_{20|10}\hat{P}_{10}}{\hat{P}^*_{10}}$. Continuing this way, we can alternate between Lemma 2.3.1 and Thereom 1 to obtain the estimates of P^*_{j1}, $j = 1, 2, \ldots, K$. Let $\hat{P}^*_1 = (\hat{P}^*_{11}, \hat{P}^*_{21}, \ldots, \hat{P}^*_{K1})^T$ denote the vector of the parameter estimates. Variance of this estimator can be obtained using the methods of M-estimator as detailed by Stefanski and Boos (2002). $\hat{P}_1^*$ can be written as a solution to the estimating equation $\sum_{i=1}^{n} \psi(\widetilde{E}_i; \hat{P}_1^*) = 0$, where

$$\psi(\widetilde{E}_i; P_1^*) = \left(\begin{array}{c} \phi + (1-\theta-\phi)(1-E_{i1}) - P_{11}^* \\ P_{11}^* + \phi(1-P_{11}^*) + \theta(1-\theta-\phi)(1-E_{i2})(1-E_{i1}) - P_{21}^* \\ \vdots \\ P_{11}^* + \phi[(K-1) - \sum_{m=1}^{K-1} P_{m1}^*] + \theta(1-\theta-\phi)[\sum_{g=1}^{K-1} \prod_{k=1}^{g+1} (1-E_{ik})] - P_{K1}^* \end{array} \right)$$

Then the variance of the estimator $\hat{P}_1^*$ can be estimated using the sandwich estimator. Since M-estimators are consistent and asymptotically normally distributed, $\hat{P}_1^*$ will be consistent and asymptotically normal. Point-wise confidence intervals for survival estimates can be constructed using Wald's method.

2.4 SIMULATION STUDY

In this section we evaluate the large sample properties of the proposed method in small to moderately large samples. We simulated data from a population with a design similar to the VIRAHEP-C study. Each individual is followed and evaluated a maximum of K = 8 times or until an event is observed, at which point that individual is no longer followed. We specify the population of interest as follows: The true survival distribution of T^*, the time to first event in the absence of classification error is specified by the true survival probabilities of the event at the 8 evaluation times, namely, $\mathbf{P}_0^* = (1 - \mathbf{P}_1^*) = (0.550, 0.400, 0.350, 0.325, 0.300, 0.275, 0.250, 0.200)^T$. For given values of $\mathbf{P}_0^*$, θ and ϕ, one can obtain the observed conditional probabilities through the results given in Lemma 2.3.1. At the first evaluation time t_j, $P_{10} = \frac{(1-\phi)-P_{10}^*}{(1-\theta-\phi)}$. For time points $t_2, t_3, \ldots, t_K$

$$P_{j0|(j-1)0,\ldots,10} = \frac{\sum_{m=1}^{j-1} \prod_{l=0}^{m-1} P_{l0|(l-1)0,..,10}^* (1 - P_{m0|(m-1)0,..,10}^*) + \phi \prod_{m=1}^{j-1} P_{m0|(m-1)0,..,10}^* - P_{j1}^*}{\theta(\theta+\phi-1) \prod_{m=1}^{j-1} P_{m0|(m-1)0,..,10}}.$$

Note that in this data generation process, the parameters (θ, ϕ and $\mathbf{P}_0^*$) need to be chosen carefully so that the probabilities lie between 0 and 1.

We use these conditional probabilities to generate 5000 Monte Carlo samples; n error-prone observations were drawn from the true population described above with θ and ϕ ranging from $(1.00, 1.00)$ to $(0.90, 0.80)$. Table 2 presents the results for $n = 250$. First consider the case of no misclassification (θ and ϕ equal to 1.0). Here, the proposed estimator is identical to the EDF estimator, as expected. The proposed estimator of the true survival probabilities are unbiased. The standard errors of the estimators are close to the Monte-Carlo standard errors, showing that the estimated variance is consistent. The coverage probabilities at all time points closely matched the nominal confidence of 95%.

When θ=1 and ϕ = 0.9, the EDF estimator is biased; at timepoint t_1 it is 5.0% compared to that at timepoint t_8 (0.5%). The proposed estimator of the true survival probabilities were unbiased. The standard errors of the estimators are close to the Monte-Carlo standard errors, again, showing that the estimated variance of the proposed estimator is consistent. The coverage probabilities at all time points closely matched the nominal confidence of 95%. In the case where $\theta = 1$ and $\phi = 0.8$ the EDF estimator shows even larger bias compared to the previous scenario where ϕ was set to 0.90. The bias at timepoint t_1 is 11.2% whereas at timepoint t_{10} it is 1.3%. On the other hand, even with such decrease in PPV the proposed estimators of the true survival probabilities and their standard errors remained unbiased. The coverage probabilities of 95% confidence intervals ranges between 94.5% and 94.9%.

When θ was reduced to 0.95 with ϕ fixed at 1.0, the bias of the EDF estimator ranged from 2.8% at timepoint t_1 to 2.1% at timepoint t_8. The proposed estimator of the true survival probabilities were unbiased, its standard errors matched the Monte-Carlo standard errors, and the coverage probabilities were between 94.6% and 95.3%. Table 2 shows that in the presence of error-prone events the estimates from the EDF method are biased. This bias increases as the PPV and NPV of the diagnostic tool decreases. Table 3 presents the results for $n = 500$. The results are similar to Table 2, except that (expectedly) the standard errors were smaller compared to Table 2 and the coverage probabilities at all time points were even closer to the nominal confidence of 95%. The results from the two tables described above show that if the NPV and PPV of the diagnostic tool are known, the true survival rates can be estimated with no bias using the proposed estimator.

2.5 ANALYSIS OF THE VIRAHEP-C DATA

All participants in the VIRAHEP-C study (Conjeevaram et al, 2006) were chronically infected with Hepatitis C virus (HCV) of genotype 1. The study was designed to test the hypothesis that African Americans respond less well to anti-viral therapy than Caucasians. Viral levels were measured at Days 1, 2, 3, 7, and weeks 2, 3, 4, 8, 12, 24 and 48. One of the aims of the study was to investigate the time to viral negativity. True viral negativity is defined as HCV RNA in serum at or below limit of detection (50 IU/ml) by a qualitative assay. Therefore, in the notation of this paper, $E^* = \{\text{viral levels} \leq 50$ IU/ml by qualitative assay$\}$. The potentially misclassified events are obtained from the quantitative assay, hence $E = \{\text{viral levels} \leq 600$ IU/ml by quantitative assay$\}$. Viral level measurements at Day 3 and Week 3 were discontinued due to limited resources after the study recruited about one-third of the participants. We therefore excluded those time points from our analysis. In addition, we limited our analysis to the evaluation points up to 24 weeks. Thus, the final data consist of the following visits: Days 1, 2, 7, and weeks 2, 4, 8, 12 and 24.

The definition of the true and observed event above implies that if $E = 0$ then $E^* = 0$. This is because if viral levels were greater than 600 IU/ml by the quantitative PCR, they will be detected by the qualitative PCR. Therefore, $\theta = P(E^* = 0 \mid E = 0) = 1$. However, if the quantitative PCR detect viral levels of less than 600 IU/ml, only a fraction of the results will be in agreement with the qualitative PCR. Therefore, $\phi = P(E^* = 1 \mid E = 1) < 1$. To calculate ϕ, we used the data from the timepoints at which both qualitative and quantitative assay were performed. Only 31% of the negative results from the quantitative assay (i.e. viral level < 600 IU/ml) were also negative by the qualitative assay (i.e. viral level < 50 IU/ml). Therefore ϕ was estimated as 0.31. Thus, for our data analysis, θ is set to 1 and ϕ to 0.31.

If an evaluation was missing but was followed by an occurrence or nonoccurrence of an event, then the missing evaluation result was set to a nonoccurrence, this is because we are interested in the time to the first event. Our methodology thus far only deals with individuals who are not lost to follow-up, as a consequence, participants with incomplete data were excluded from the analysis resulting in a total of 355 participants [171 African Americans (AA); 184 Caucasians (CA)]. The results of the data analysis are shown in Table 4. If the less sensitive quantitative assay results are used and an EDF estimator is used to estimate the survival function of time to virus negativity, then at Weeks 2 and 24, 87.9% and 23.7% of the patients are estimated to remain viral positive. These numbers underestimate the true survival function as seen by the corresponding estimates of 94.5% and 49.0% respectively using our proposed method; the EDF estimator and the proposed estimator deviates substantially as time progresses and the event rates increase.

We further investigate results of both methods by testing for differences between African Americans and Caucasians; tests of statistical significance were calculated by Wald's method. For example, the EDF estimate of the probability of survival at Day 1 for AA and CA are 99.4% and 96.7% respectively whereas our proposed estimates of the probability of survival at the same time of interest are 99.8% for AA and 98.9% for CA (Table 4). Figure 2 shows a graphical representation of the proportion remaining viral positive at different time points by race. Although the racial difference in estimated survival was relatively larger for the EDF estimates as time progressed, compared to the proposed estimates, the differences were not statistically significant by either method.

2.6 DISCUSSION

Analysis of time-to-event data is common in biomedical research. The timing and cumulative incidence of a disease are essential factors in physician and patient decision process in all phases of an illness. Accurate rates of survival in advanced chronic diseases such as Alzheimer's disease or cancer may have increased importance as patients approach the end of life, since this presents a time to reconsider the goals of treatment which for example may change from prolongation of life to alleviating pain. In time to event analysis with a binary outcome, such as the one analyzed in this article, event misclassification may occur. We have shown that when there is misclassification, frequently used techniques of estimating the survival probabilities e.g. EDF estimator, are biased. We have proposed a new method to estimate the survival probabilities with no bias by incorporating the NPV and PPV of the diagnostic tool into a product-limit-type estimator. An estimator of the variance of the proposed estimator is also given and shown to be consistent through simulation studies. The proposed method may not provide valid estimates when the PPVs and NPVs are very small, however this may not be a problem since diagnostic tools with such high rates of misclassification are rarely used in practice. Our method provides statisticians with a tool to accurately estimate the survival probabilities in the presence of misclassified events.

While the methods proposed here are useful, they cannot be used when individuals are lost to follow up. Censoring complicates the formulas derived in this paper, this problem will be considered in future work.

2.7 DERIVATION OF ESTIMATING EQUATIONS

The estimated probability of truly having an event at time t_1 is

$$\begin{aligned}\hat{P}^*_{11} &= 1 - \{\theta \hat{P}_{10} + (1-\phi)\hat{P}_{11}\} \qquad (2.11)\\ &= \phi + (1-\theta-\phi)\hat{P}_{10}\\ &= \phi + (1-\theta-\phi)(1-E_{i1}).\end{aligned}$$

The estimated probability of truly having an event at time t_2 is

$$\begin{aligned}\hat{P}^*_{21} &= \hat{P}^*_{11} + \hat{P}^*_{10} * \hat{P}^*_{21|10} \qquad (2.12)\\ &= \hat{P}^*_{11} + \hat{P}^*_{10} * \{1 - \theta\zeta^{(1)} - (1-\phi)(1-\zeta^{(1)})\}\\ &= \hat{P}^*_{11} + \hat{P}^*_{10} * \{1 - (\frac{\theta^2 \hat{P}_{20|10}\hat{P}_{10}}{\hat{P}^*_{10}}) - (1-\phi)(1 - \frac{\theta\hat{P}_{20|10}\hat{P}_{10}}{\hat{P}^*_{10}})\}\\ &= \hat{P}^*_{11} + \phi(1-\hat{P}^*_{11}) + \theta(1-\theta-\phi)\hat{P}_{20|10}\hat{P}_{10}\\ &= \hat{P}^*_{11} + \phi(1-\hat{P}^*_{11}) + \theta(1-\theta-\phi)(1-E_{i2})(1-E_{i1}).\end{aligned}$$

The estimated probability of truly having an event at time t_3 is

$$\begin{aligned}\hat{P}^*_{31} &= \hat{P}^*_{11} + \hat{P}^*_{10} * \hat{P}^*_{21|10} + \hat{P}^*_{10} * \hat{P}^*_{20|10}\hat{P}^*_{31|20|10} \qquad (2.13)\\ &= \hat{P}^*_{11} + \hat{P}^*_{10} * \{1 - \theta\zeta^{(2)} - (1-\phi)(1-\zeta^{(2)})\}\\ &= \hat{P}^*_{11} + \hat{P}^*_{10} * \{1 - (\frac{\theta^2 \hat{P}_{30|20|10}\hat{P}_{20|10}\hat{P}_{10}}{\hat{P}^*_{20|10} * \hat{P}^*_{10}}) - (1-\phi)(1 - \frac{\theta\hat{P}_{30|20|10}\hat{P}_{20|10}\hat{P}_{10}}{\hat{P}^*_{20|10} * \hat{P}^*_{10}})\}\\ &= \hat{P}^*_{11} + (2 - \hat{P}^*_{11} - \hat{P}^*_{21}) + \theta(1-\theta-\phi)(\hat{P}_{30|20|10}\hat{P}_{20|10}\hat{P}_{10} + \hat{P}_{20|10}\hat{P}_{10})\\ &= \hat{P}^*_{11} + (2 - \hat{P}^*_{11} - \hat{P}^*_{21}) + \theta(1-\theta-\phi)((1-E_{i3})(1-E_{i2})(1-E_{i1}) + (1-E_{i2})(1-E_{i1})).\end{aligned}$$

The above process is repeated to attain subsequent time points. The final formula for the estimated probability of truly having an event at time t_j is

$$\hat{P}^*_{j1} = \hat{P}^*_{11} + \phi[(K-1) - \sum_{m=1}^{K-1}\hat{P}^*_{m1}] + \theta(1-\theta-\phi)[\sum_{g=1}^{K-1}\prod_{k=1}^{g+1}(1-E_{ik})]. \qquad (2.14)$$

2.8 TABLES AND FIGURES

Table 2: Simulation results for estimating $P(T^* > t_k, k = 1, 4, 8)$ based on 5000 Monte Carlo (MC) samples of size 250. EST is the MC mean of the proposed estimate assuming θ and ϕ known, SE is the MC mean of the estimated standard errors, MCSE is the standard error of MC estimates, CP is the empirical coverage probablity, EDF is the Empirical distribution function estimate ignoring misclassification.

			$(\theta = 1.00)$					$(\theta = 0.95)$					$(\theta = 0.90)$				
	Parameter	Truth	EDF	EST	MCSE	SE	CP%	EDF	EST	MCSE	SE	CP%	EDF	EST	MCSE	SE	CP%
$(\phi = 1)$	$P(T^* > t_1)$	0.550	0.549	0.549	0.031	0.031	94.8	0.578	0.549	0.029	0.029	94.6	0.611	0.550	0.028	0.028	94.9
	$P(T^* > t_4)$	0.325	0.324	0.324	0.029	0.030	95.4	0.359	0.324	0.028	0.027	95.3	0.400	0.324	0.025	0.025	94.8
	$P(T^* > t_8)$	0.200	0.199	0.199	0.024	0.025	95.2	0.221	0.199	0.024	0.024	95.2	0.247	0.199	0.021	0.022	94.4
$(\phi = 0.9)$	$P(T^* > t_1)$	0.550	0.500	0.549	0.029	0.028	94.7	0.528	0.549	0.027	0.026	94.7	0.562	0.549	0.025	0.025	94.6
	$P(T^* > t_4)$	0.325	0.322	0.325	0.030	0.029	94.2	0.358	0.324	0.028	0.027	94.5	0.402	0.324	0.025	0.025	94.6
	$P(T^* > t_8)$	0.200	0.195	0.199	0.025	0.024	94.4	0.216	0.199	0.023	0.023	94.6	0.243	0.199	0.022	0.021	94.9
$(\phi = 0.8)$	$P(T^* > t_1)$	0.550	0.438	0.550	0.024	0.025	94.5	0.466	0.550	0.023	0.023	95.5	0.499	0.549	0.022	0.022	94.8
	$P(T^* > t_4)$	0.325	0.318	0.324	0.028	0.029	94.9	0.358	0.325	0.026	0.027	95.0	0.404	0.324	0.024	0.024	94.5
	$P(T^* > t_8)$	0.200	0.187	0.199	0.024	0.024	94.5	0.211	0.200	0.023	0.023	95.3	0.238	0.199	0.021	0.020	94.8

Table 3: Simulation results for estimating $P(T^* > t_k, k = 1, 4, 8)$ based on 5000 Monte Carlo (MC) samples of size 500. EST is the MC mean of the proposed estimate assuming θ and ϕ known, SE is the MC mean of the estimated standard errors, MCSE is the standard error of MC estimates, CP is the empirical coverage probablity, EDF is the Empirical distribution function estimate ignoring misclassification.

			$(\theta = 1.00)$					$(\theta = 0.95)$					$(\theta = 0.90)$				
	Parameter	Truth	EDF	EST	MCSE	SE	CP%	EDF	EST	MCSE	SE	CP%	EDF	EST	MCSE	SE	CP%
$(\phi = 1)$	$P(T^* > t_1)$	0.550	0.549	0.549	0.022	0.022	94.2	0.578	0.549	0.021	0.020	95.1	0.611	0.550	0.019	0.019	94.9
	$P(T^* > t_4)$	0.325	0.324	0.324	0.021	0.020	94.8	0.359	0.324	0.019	0.019	94.6	0.401	0.325	0.017	0.018	95.8
	$P(T^* > t_8)$	0.200	0.199	0.199	0.018	0.018	94.7	0.221	0.199	0.016	0.016	95.3	0.247	0.199	0.015	0.016	95.6
$(\phi = 0.9)$	$P(T^* > t_1)$	0.550	0.499	0.549	0.020	0.020	94.7	0.529	0.549	0.019	0.018	95.1	0.562	0.549	0.018	0.017	94.9
	$P(T^* > t_4)$	0.325	0.321	0.324	0.020	0.021	95.2	0.359	0.324	0.020	0.019	94.7	0.403	0.324	0.018	0.017	94.5
	$P(T^* > t_8)$	0.200	0.194	0.199	0.018	0.018	95.1	0.216	0.199	0.016	0.016	94.8	0.243	0.199	0.015	0.015	94.8
$(\phi = 0.8)$	$P(T^* > t_1)$	0.550	0.438	0.549	0.018	0.017	94.9	0.467	0.549	0.016	0.017	95.8	0.499	0.549	0.016	0.016	94.3
	$P(T^* > t_4)$	0.325	0.319	0.325	0.020	0.020	94.5	0.358	0.324	0.019	0.019	95.2	0.404	0.324	0.017	0.017	94.6
	$P(T^* > t_8)$	0.200	0.188	0.200	0.017	0.017	95.2	0.211	0.200	0.015	0.016	95.1	0.238	0.199	0.015	0.015	94.9

Table 4: Analysis results of the estimation of survival probabilities for time to viral negativity at selected time points. All is the overall estimated survival estimates by both methods. AA (n=171) and CA (n=184) stands for African Americans and Caucasians, respectively. EDF is the Empirical distribution function estimated survival ignoring misclassification. Proposed is our proposed estimate of the true survival. Right below the estimated survival probabilities are the 95% confidence intervals of the estimates; p-value compares survival rates between groups at the designated time point.

	EDF				Proposed			
Parameter	All	AA	CA	p-value	All	AA	CA	p-value
$P(T^* > 1)$	0.980	0.994	0.967	0.24	0.993	0.998	0.989	0.23
	(0.966,0.995)	(0.983,1.00)	(0.942,0.993)		(0.989,0.998)	(0.994, 1.000)	(0.981, 0.997)	
$P(T^* > 14)$	0.879	0.912	0.848	0.25	0.945	0.962	0.930	0.24
	(0.845,0.913)	(0.869,0.954)	(0.796,0.899)		(0.923,0.962)	(0.942, 0.982)	(0.904, 0.956)	
$P(T^* > 168)$	0.237	0.333	0.147	0.08	0.490	0.575	0.412	0.06
	(0.192,0.281)	(0.263,0.404)	(0.096,0.198)		(0.458,0.523)	(0.525, 0.623)	(0.372, 0.456)	

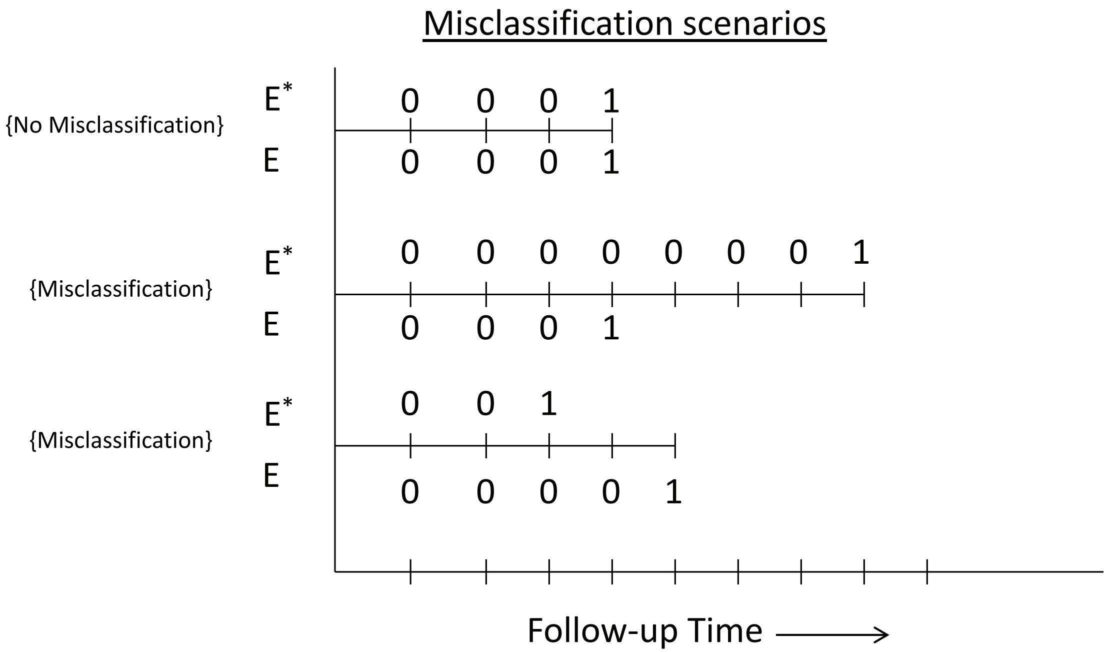

Figure 1: Examples of missclassification. E^* is the true occurence of an event and E is the potentially misclassified event. Both E^* and E can take values 1 or 0, indicating the occurence and non-occurence of the event, respectively.

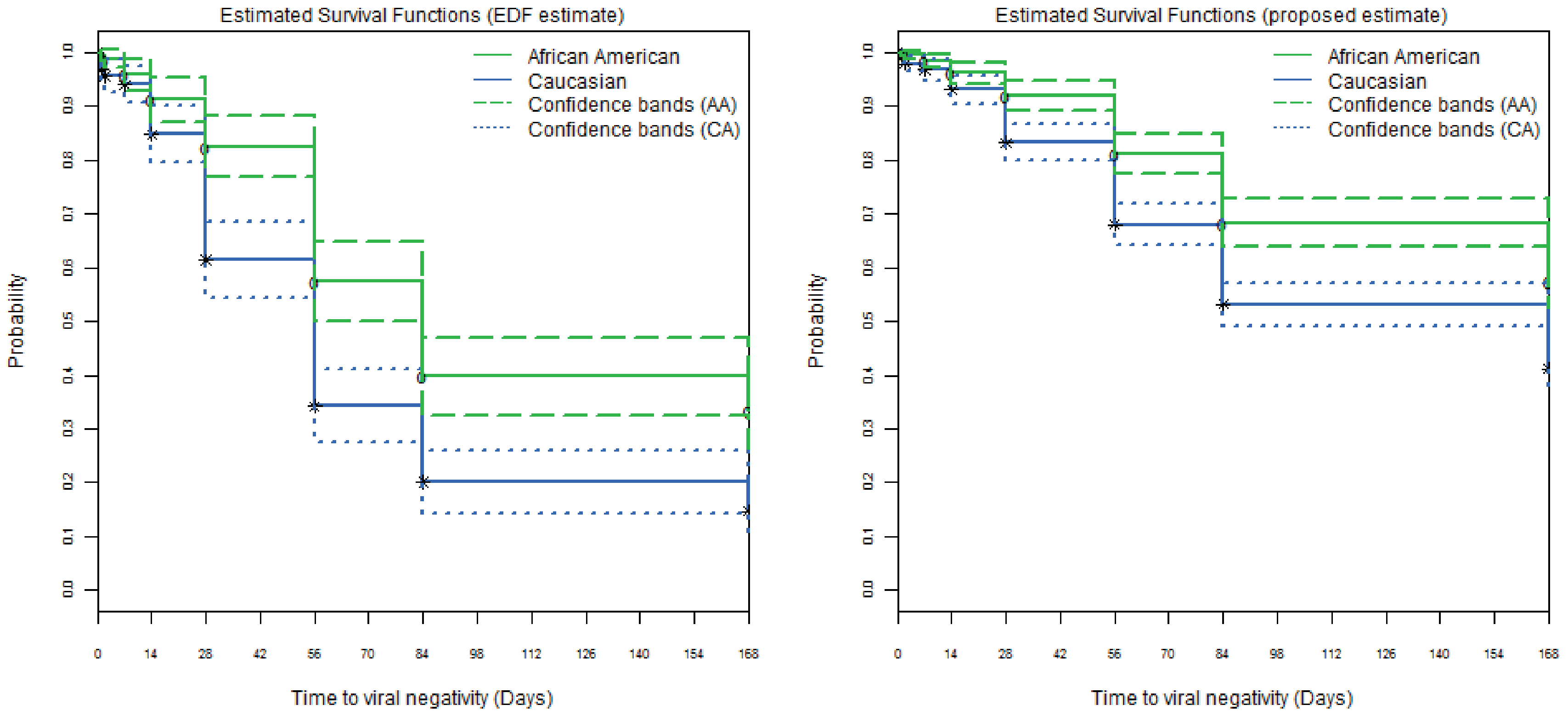

Figure 2: Estimated survival curves for time to viral negativity at selected time points for the VIRAHEP-C study. (Left panel: EDF estimates; right panel: proposed method).

3.0 DISCRETE SURVIVAL ANALYSIS WITH MISCLASSIFIED EVENTS AND WITH LOST-TO-FOLLOW-UP

3.1 INTRODUCTION

Incomplete follow-up is common in longitudinal studies for which time-to-event data are of primary interest. The structure of our framework is that event free individuals participate in a study and are followed by clinical visits until the occurrence of the event of interest or the individual is lost to follow-up or untill the end of the study. In such cases, some individuals drop out of the study without adequate follow-up. Also, since the number of clinical visits is not unlimited, there are some individuals who will be event free at the conclusion of the study. These scenarios bring about the important issue of right censored data.

Data from epidemiological studies are frequently used to estimate the survival distribution of the time to event of interest. Methods developed in Section 2 will not work for this purpose as they are not equipped to handle censored data. To illustrate the effects of drop-out on our proposed estimator from Theorem 2.10 (Chapter 2) which does not account for indivividuals lost to follow-up, Table 5 shows results from a simulation study of complete-case analysis. Under the same framework as in section 2.4 we generate 5000 Monte Carlo samples; n error-prone observations were drawn from the true population as in section 2.4 with θ and ϕ ranging from $(1.00, 1.00)$ to $(0.90, 0.90)$. We consider a sample for which 35% of the observations are lost to follow-up. Table 5 presents the results for $n = 250$. First consider the case of no misclassification (θ and ϕ equal to 1.0), even in the case of no misclassification error, the proposed estimator is biased to the true survival probabilities. In fact for all specified ranges of θ and ϕ the proposed estimator is biased. The results of Table 5 clearly show that new methods must be developed to handle individuals with incomplete follow-up.

The Kaplan-Meier (KM) product limit estimator Kaplan and Meier (1958) is a commonly used method for estimating the survival function in the presence of censoring. Greenwood's formula (Greenwood, 1926), is widely utilized to obtain an estimate of the variance of the KM estimator. However, when the outcome of interest is subject to misclassification, the true survival distribution is latent; thus, the KM estimator produces inaccurate conclusions. We approach the problem of estimating the true (latent) survival distribution by constructing a Kaplan-Meier-like estimator; this estimator will take into account the misclassification probabilities (NPV and PPV) and will also handle the issue of right censoring. We provide an expression for the variance of the proposed estimator and a formula for the variance estimate. Unlike Greenwood's variance formula, this variance estimate does not assume that the conditional probabilities of survival are asymptotically independent.

This article is organized as follows. We continue with the same notation, data, and assumptions as in Section 2.2. In Section 3.2 we propose an estimate of the true survival function and derive its formulation. We also derive the variance of our true survival rate estimator using methods due to Breslow and Crowley (1974). In addition, we show the consistency of our estimator using methods provided by Gill (1983). We evaluate the large-sample properties of the proposed methods through simulations in Section 3.3. In Section 3.4 we appply our proposed methods to analyze data from the VIRAHEP-C study. We conclude our analysis with a discussion in Section 3.5. A brief overview of future work is introduced in Section 4.

3.2 INFERENCE FROM INCOMPLETE DATA: PRESENCE OF DROP-OUT

To deal with censoring first note that no subjects have an event at the begining of the follow-up (t_0). Thus $P^*_{01} = P(T^* \leq t_0) = P(E^*_0 = 1) = 0$. From Equation (2.6) of chapter (2) the probability that an event does not occur by time t_1 is

$$\begin{aligned} P^*_{10} = P(E^*_1 = 0) &= \theta P_{10} + (1-\phi)P_{11} \\ &= (1-\phi) - (1-\theta-\phi)P_{10}. \end{aligned}$$

Setting j=2 in Theorem 2.10, Chapter 2, and taking the compliment, the unconditional probability that a true event does not occur by time t_2 is,

$$P^*_{20} = P^*_{20|10}P^*_{10}.$$

Now, using Lemma 2.3.1,

$$P^*_{20} = \{\theta\zeta^{(1)} + (1-\phi)(1-\zeta^{(1)})\}P^*_{10},$$

where $\zeta^{(1)} = \theta P_{20|10}(\frac{P_{10}}{P^*_{10}})$. Further simplification results in the expression

$$P^*_{20} = (1-\phi)^2 - (1-\phi)(1-\theta-\phi)P_{10} - \theta(1-\theta-\phi)P_{20}.$$

A similar derivation shows that the unconditional probability that a true event does not occur by time t_3 is,

$$P^*_{30} = (1-\phi)^3 - (1-\phi)^2(1-\theta-\phi)P_{10} + (1-\phi)\theta(\theta+\phi-1)P_{20} - \theta(1-\theta-\phi)P_{30}.$$

The following theorem provides a general expression for the probability of not having an event by a specific time:

Theorem 2. *Under assumptions (2.1)-(2.4), the probability of not having a true event by time t_k can be expressed as:*

$$\begin{aligned} P^*_{k0} &= \sum_{j=0}^{k} P_{j0}(1-\phi)^{k-j}\theta^{1(j>1)}(\theta+\phi-1)^{1(j>0)}, k = 1,2,\dots,K, \qquad (3.1) \\ &= \mathbf{a}^T_k(\theta,\phi)\mathbf{P}_0, \end{aligned}$$

where $\mathbf{P}_0 = (P_{00}, P_{10}, P_{20}, \ldots, P_{K0})^T$, $\mathbf{a}_k(\theta,\phi)^T = [a_{0k}(\theta,\phi), a_{1k}(\theta,\phi), \ldots, a_{Kk}(\theta,\phi)]$, and

$$a_{jk} = \begin{cases} (1-\phi)^{k-j}\theta^{1(j>1)}(\theta+\phi-1)^{1(j>0)}, j = 0, 1, \ldots, k \\ 0, j = k+1, \ldots, K. \end{cases} \tag{3.2}$$

Theorem 3.1 gives an expression for the true survival rates (P^*_{j0}) as a formulation of the error prone survival vector ($\mathbf{P}_0$). Thus, P^*_{j0} can be estimated through $\mathbf{P}_0$, elements of which can be estimated by the standard Kaplan-Meier approach namely,

$$\hat{P}_{j0} = \prod_{k \le j} \left(1 - \frac{\sum_{i=1}^{n} E_{ik}}{\sum_{i=1}^{n} \left(1 - E_{i(k-1)}\right)}\right), j = 1, 2, \ldots, K, \tag{3.3}$$

Thus an expression for the estimate of the true survival distribution is given as:

$$\begin{aligned} \hat{P}^*_{k0} &= \sum_{j=0}^{k} \hat{P}_{j0}(1-\phi)^{k-j}\theta^{1(j>1)}(\theta+\phi-1)^{1(j>0)}, k = 1, 2, \ldots, K, \\ &= \mathbf{a}_k^T(\theta,\phi)\hat{\mathbf{P}}_0, \end{aligned} \tag{3.4}$$

where $\hat{\mathbf{P}}_0 = (\hat{P}_{00}, \hat{P}_{10}, \hat{P}_{20}, \ldots, \hat{P}_{K0})^T$.

Theorem 3. *(Consistency) Under assumptions (2.1)-(2.4), the estimators defined in (3.4) are consistent.*

Proof. The result follows from the fact that KM estimator $\hat{\mathbf{P}}_0$ of $\mathbf{P}_0$ are consistent (Gill, 1983) and that the estimator $\hat{P}^*_{k0}$ is a linear combination of $\hat{\mathbf{P}}_0$. □

The next step is to derive a formula for the variance of our survival rate estimator. We adapt the techniques of Breslow and Crowley (1974) to estimate the variance of our survival rate estimator. They studied properties of the life table and product limit estimates under random censorship and proposed rigorous derivations of many of its formal large sample properties. Their study is particularly important to our investigation because the type of life table considered in their paper is the cohort table used for estimation of a survival distribution from right censored data. A useful result of their work is the derivation of the asymptotic covariance between K-M estimates at different time points. We use this result to derive the asymptotic variance of the KM estimates and thus obtain the asymptotic covariance matrix of our proposed estimator in the presence of right censoring and event misclassification.

Theorem 4. *(Asymptotic normality) Under assumptions (2.1)-(2.4), the estimators defined in (3.4) are asymptotically normal with mean P^*_{k0} and variance*

$$Var(\hat{P}^*_{k0}) = \mathbf{a}^T_k(\theta, \phi) Cov(\hat{\mathbf{P}}_0)\mathbf{a}_k(\theta, \phi),$$

where,

$$Cov(\hat{\mathbf{P}}_0) = \begin{pmatrix} Var(\hat{P}_{10}) & Cov(\hat{P}_{10}, \hat{P}_{20}) & \dots & Cov(\hat{P}_{10}, \hat{P}_{K0}) \\ Cov(\hat{P}_{20}, \hat{P}_{10}) & Var(\hat{P}_{20}) & \dots & Cov(\hat{P}_{20}, \hat{P}_{K0}) \\ & \vdots & \ddots & \vdots \\ Cov(\hat{P}_{K0}, \hat{P}_{10}) & \dots & \dots & Var(\hat{P}_{K0}) \end{pmatrix};$$

the elements of $Cov(\hat{\mathbf{P}}_0)$ are given by $\frac{Cov(Z_j, Z_k)}{n}$.

Proof. By Theorem 5 of Breslow and Crowley (1974), Let $t_K < \infty$ satisfy $P_{K1} < 1$. Then the random variable $\sqrt{n}(\hat{P}_{j0} - P_{j0})$, for $0 < j < k < K$, converges weakly to a mean zero normal random variable Z_j, moreover,

$$Cov(Z_j, Z_k) = (1 - P_{j1})(1 - P_{k1}) \sum_{t=0}^{j} (1 - P_{t1})^{-2}(1 - H)^{-1} P(E_t = 1), j \leq k$$

where H is right censoring distribution drawn independently of P_{j1}; Theorem 5 of Breslow and Crowley (1974) proved the Kaplan-Meier estimator to be asymptotically normal. Our estimator of the true survival distribution is a linear combination of the Kaplan-Meier estimator, therefore, it follows that $\hat{P}^*_{k0}$ is also asymptotically normal. □

The j^{th} diagonal element of $Cov(\hat{\mathbf{P}}_0)$ is estimated by Greenwood's formula for variance of the cumulative probability of survival given by,

$$\hat{Var}(\hat{P}_{j0}) = \hat{P}_{j0}^2 \prod_{k \le j} \left(\frac{\sum_{i=1}^n E_{ik}}{\sum_{i=1}^n (1 - E_{i(k-1)})(\sum_{i=1}^n (1 - E_{i(k-1)}) - \sum_{i=1}^n E_{ik})} \right), j = 1, 2, \ldots, K, \tag{3.5}$$

and the off-diagonal elements of $Cov(\hat{\mathbf{P}}_0)$ is estimated by the following,

$$\hat{Cov}(\hat{P}_{j0}, \hat{P}_{k0}) = \frac{(1 - \hat{P}_{k1})}{(1 - \hat{P}_{j1})} \hat{Var}(\hat{P}_{j0}), j < k; j = 1, \ldots, K. \tag{3.6}$$

3.3 SIMULATION STUDY

We evaluate the large sample properties of our proposed method in small to moderately large samples. As in the case with no lost-to-follow-up, we simulated data from a population with a design similar to the VIRAHEP-C study. Each individual was followed and evaluated a maximum of K = 8 times or until an event is observed, at which point that individual is no longer followed. The true survival distribution of T^*, the true time to first event in the absence of classification error is specified by the true survival probabilities of the event at the 8 evaluation times, namely, $\mathbf{P}_0^* = (1 - \mathbf{P}_1^*) = (0.550, 0.400, 0.350, 0.325, 0.300, 0.275, 0.250, 0.200)^T$. For given values of $\mathbf{P}_0^*$, θ and ϕ, one can obtain the observed conditional probabilities through the results given in Lemma 2.3.1. At the first evaluation time t_j, $P_{10} = \frac{(1-\phi) - P_{10}^*}{(1-\theta-\phi)}$. For time points $t_2, t_3, \ldots, t_K$

$$P_{j0|(j-1)0,\ldots,10} = \frac{\sum_{m=1}^{j-1} \prod_{l=0}^{m-1} P^*_{l0|(l-1)0,..,10}(1 - P^*_{m0|(m-1)0,..,10}) + \phi \prod_{m=1}^{j-1} P^*_{m0|(m-1)0,..,10} - P^*_{j1}}{\theta(\theta + \phi - 1) \prod_{m=1}^{j-1} P_{m0|(m-1)0,..,10}}.$$

Note that in this data generation process, the parameters (θ, ϕ and $\mathbf{P}_0^*$) need to be chosen carefully so that the probabilities lie between 0 and 1.

These conditional probabilities were used to generate 5000 Monte Carlo samples; n error-prone observations were drawn from the true population described above with θ and ϕ ranging from $(1.00, 1.00)$ to $(0.90, 0.80)$. Additionally, a number of participants were allowed to drop out for the purpose of illustration based on the following uniform distribution: $U \sim (a, b)$. Choices of parameters $(a, b) = (0, 15)$ gave us approximate drop-out rate of 35%. Table 6 presents the results for $n = 250$. First consider the case of no misclassification (θ and ϕ equal to 1.0). The proposed estimator is identical to the KM estimator, as expected. The proposed estimator of the true survival probabilities are unbiased. The standard errors of the estimator are close to the Monte-Carlo standard error, showing that the estimated variance is consistent. The coverage probabilities at all time points closely matched the nominal confidence of 95%.

We introduce misclassification errors, first consider changing ϕ keeping θ fixed. When θ=1 and $\phi = 0.9$, the KM estimator is biased; at timepoint t_1 it is 4.7% compared to that at timepoint t_8 (0.2%). The proposed estimator of the true survival probabilities are unbiased. The standard error of this estimator are close to the Monte-Carlo standard error, showing that the estimated variance of the proposed estimator is consistent. The coverage probabilities at all time points closely matched the nominal confidence of 95%. In the case where $\theta = 1$ and $\phi = 0.8$ the KM estimator shows even larger bias compared to the previous scenario. The bias at timepoint t_1 is 10.8% whereas at timepoint t_{10} it is 0.9%. On the other hand, even with such decrease in PPV the proposed estimator of the true survival probabilities and their standard error remain unbiased. The coverage probabilities of 95% confidence intervals range between 94.5% and 95.2%.

When θ was reduced to 0.95 with ϕ fixed at 1.0, the bias of the KM estimator is 3.1% at timepoint t_1, 3.8% at timepoint t_4 and at timepoint t_8 the bias is 2.4%. The proposed estimator of the true survival probabilities are unbiased, its standard error matched the Monte-Carlo standard error, and the coverage probabilities were between 94.1% and 94.9%. Table 6 shows that in the presence of error-prone events the estimates from the KM method are biased. This bias increases as the PPV and NPV of the diagnostic tool decreases. Table 7 presents the results for a sample size of 500. The overall conclusion remains the same as in Table 6.

In the presence of error-prone events, the KM method is biased in estimating the true survival rates. However, there are interesting results, namely the cases where θ is fixed at 1.0 and $\phi = 0.9$ or 0.8, in the later timepoints the KM method is just as good as our proposed estimator in estimating the true survival rates. This is because at later timepoints, most participants have truly had the event of interest, therefore, a missclassification rate of predicting the occurence of an event (1-PPV) of 10% or 20% has little effect on estimating the true survival probabilities by the KM method.

3.4 ANALYSIS OF THE VIRAHEP-C DATA

VIRAHEP-C was a multicenter, collaborative clinical trial, sponsored by NIDDK-NIH, that was designed to test the hypothesis that African Americans respond less well to antiviral therapy than Caucasian patients. A total of 401 chronically infected patients with Hepatitis C virus (HCV) of genotype 1 were enrolled in the VIRAHEP-C study (Conjeevaram et al, 2006). Viral levels were measured at Days 1, 2, 3, 7, and weeks 2, 3, 4, 8, 12, 24 and 48. One of the aims of the study was to investigate the time to viral negativity. True viral negativity is defined as HCV RNA in serum at or below limit of detection (50 IU/ml) by a qualitative assay. Therefore, in the notation of this paper, $E^* = \{\text{viral levels} \leq 50 \text{ IU/ml by qualitative assay}\}$. The potentially misclassified events are obtained from the quantitative assay, hence $E = \{\text{viral levels} \leq 600 \text{ IU/ml by quantitative assay}\}$. Viral level measurements at Day 3 and Week 3 were discontinued due to limited resources after the study recruited about one-third of the participants. We therefore excluded those time points from our analysis. In addition, we limited our analysis to the evaluation points up to 24 weeks. Thus, the final data consist of the following visits: Days 1, 2, 7, and weeks 2, 4, 8, 12 and 24.

The definition of the true and observed event above implies that if the quantitative assay detects a non-event ($E = 0$) then the qualitative assay will also be in agreement ($E^* = 0$). This is because if viral levels were greater than 600 IU/ml by the quantitative PCR, they will be detected by the qualitative PCR. Therefore, $\theta = P(E^* = 0 \mid E = 0) = 1$. However, if the quantitative PCR detect viral levels of less then 600 IU/ml, only a fraction of the results will be in agreement with the qualitative PCR. Therefore, $\phi = P(E^* = 1 \mid E = 1) < 1$. To calculate ϕ, we used the data from the timepoints at which both qualitative and quantitative assay were performed. Only 30% of the negative results from the quantitative assay (i.e. viral level $<$ 600 IU/ml) were also negative by the qualitative assay (i.e. viral level $<$ 50 IU/ml). Therefore ϕ was estimated as 0.3.

In our framework, once an individual is observed to have the event of interest, follow-up ends. If an evaluation was missing but was followed by an occurrence or nonoccurrence of an event, then the missing evaluation result was set to a nonoccurrence, this is because we are interested in the time to the first event. The total number of participants in our data analysis is 401 [196 African Americans (AA); 205 Caucasians (CA)]; 32% of participants were censored.

The results of the data analysis are shown in Table 8. If the less sensitive quantitative assay results are used and a KM estimator is used to estimate the survival distribution of time to viral negativity, then at Weeks 2 and 24, 89.2% and 27.5% of the patients, respectively, remain viral positive. These numbers underestimate the true survival function as seen by the corresponding estimates of 95.3% and 53.2% respectively using our proposed method; the KM estimator and the proposed estimator deviates substantially as time progresses and the event rates increase.

We take a further investigation into both methods by testing for differences between African Americans and Caucasians; tests of statistical significance were calculated by Wald's method. For example, the KM estimate of the probability of survival at Day 1 for AA and CA are 99.5% and 97.1% respectively whereas our proposed estimates of the probability of survival at the same time of interest are 99.8% for AA and 99.1% for CA (Table 8). Figure 3 shows a graphical representation of the proportion remaining viral positive at different time points by race. Although the difference between the estimated survival were relatively larger for the EDF estimates towards the end compared to the proposed estimates, the differences were not statistically significant by either method.

3.5 DISCUSSION

In the analysis of time-to-event data with a binary outcome, there are times when the outcome is misclassified. For instance, in clinical diagnosis, if the diagnostic tool used to measure the event of interest is not perfect, the result may not be indicative of the participant's true event status. However, we had to deal with more than just misclassified events; we had to deal with incomplete data as well. We have shown that when there is misclassification, the Kaplan-Meier estimator is biased in estimating the true survival rates. Thus, the problem of event misclassification led us to investigate and develop methods to address this issue. Retention rates of less than 100% are common occurrences in studies that accrue large number of participants with a lengthy follow up period, hence, methods to effectively analyze incomplete data are of utmost importance.

We studied the issue of incomplete data and misclassified outcomes. We proposed a new method to estimate the survival probabilities with no bias by integrating the NPV and PPV of the diagnostic tool into the Kaplan-Meier estimator. An estimator of the variance of the proposed estimator is also provided and shown to be efficient through simulation studies. If misclassification rates are high, i.e., NPV and PPV are low, the proposed method may not produce valid estimates of the true survival rates; poor diagnostic tools are rarely used in practice, thus, this may not be a problem. Our method provides clinical investigators with a tool to accurately estimate the survival probabilities in the presence of misclassified events and incomplete data.

3.6 TABLES AND FIGURES

Table 5: Simulation results for estimating $P(T^* > t_k, k = 1, 4, 8)$ based on 5000 Monte Carlo (MC) samples of size 250. EST is the MC mean of the proposed estimate assuming θ and ϕ known, SE is the MC mean of the estimated standard errors, MCSE is the standard error of MC estimates, CP is the empirical coverage probablity.

35% Censoring			$(\theta = 1.00)$				$(\theta = 0.95)$				$(\theta = 0.90)$			
	Parameter	Truth	EST	MCSE	SE	CP%	EST	MCSE	SE	CP%	EST	MCSE	SE	CP%
$(\phi = 1)$	$P(T^* > t_1)$	0.550	0.378	0.037	0.037	0.70	0.376	0.036	0.036	0.24	0.374	0.036	0.036	0.20
	$P(T^* > t_4)$	0.325	0.100	0.023	0.022	00.0	0.105	0.023	0.023	00.0	0.110	0.023	0.022	00.0
	$P(T^* > t_8)$	0.200	0.000	0.000	0.000	00.0	0.000	0.000	0.000	00.0	0.000	0.000	0.000	00.0
$(\phi = 0.9)$	$P(T^* > t_1)$	0.550	0.387	0.032	0.032	0.20	0.382	0.031	0.031	0.08	0.378	0.030	0.030	06.0
	$P(T^* > t_4)$	0.325	0.105	0.023	0.023	00.0	0.108	0.023	0.022	00.0	0.114	0.022	0.022	00.0
	$P(T^* > t_8)$	0.200	0.004	0.001	0.001	00.0	0.004	0.001	0.001	00.0	0.004	0.001	0.001	00.0

Table 6: Simulation results for estimating $P(T^* > t_k, k = 1, 4, 8)$ based on 5000 Monte Carlo (MC) samples of size 250. EST is the MC mean of the proposed estimate assuming θ and ϕ known, SE is the MC mean of the estimated standard errors, MCSE is the standard error of MC estimates, CP is the empirical coverage probablity, KM is the Kaplan-Meier estimator ignoring misclassification.

35% Censoring			$(\theta = 1.00)$					$(\theta = 0.95)$					$(\theta = 0.90)$				
	Parameter	Truth	KM	EST	MCSE	SE	CP%	KM	EST	MCSE	SE	CP%	KM	EST	MCSE	SE	CP%
$(\phi = 1)$	$P(T^* > t_1)$	0.550	0.552	0.552	0.032	0.033	94.9	0.581	0.552	0.030	0.030	94.7	0.614	0.553	0.029	0.029	94.4
	$P(T^* > t_4)$	0.325	0.328	0.328	0.032	0.032	95.1	0.363	0.327	0.030	0.029	94.9	0.404	0.327	0.027	0.027	94.6
	$P(T^* > t_8)$	0.200	0.203	0.203	0.031	0.031	94.7	0.224	0.202	0.030	0.029	94.1	0.250	0.203	0.028	0.028	94.6
$(\phi = 0.9)$	$P(T^* > t_1)$	0.550	0.503	0.553	0.030	0.029	94.1	0.531	0.552	0.028	0.028	94.6	0.565	0.552	0.026	0.025	94.8
	$P(T^* > t_4)$	0.325	0.326	0.328	0.032	0.031	94.5	0.362	0.327	0.029	0.028	94.5	0.406	0.327	0.026	0.026	94.7
	$P(T^* > t_8)$	0.200	0.198	0.203	0.010	0.031	94.7	0.220	0.203	0.028	0.029	94.9	0.247	0.203	0.027	0.026	94.5
$(\phi = 0.8)$	$P(T^* > t_1)$	0.550	0.442	0.553	0.026	0.026	94.9	0.471	0.553	0.024	0.024	95.2	0.503	0.552	0.023	0.022	94.3
	$P(T^* > t_4)$	0.325	0.322	0.328	0.030	0.031	95.2	0.361	0.328	0.028	0.028	94.9	0.408	0.327	0.026	0.026	94.7
	$P(T^* > t_8)$	0.200	0.191	0.203	0.030	0.030	94.5	0.214	0.203	0.028	0.028	94.8	0.242	0.203	0.025	0.026	94.9

Table 7: Simulation results for estimating $P(T^* > t_k, k = 1, 4, 8)$ based on 5000 Monte Carlo (MC) samples of size 500. EST is the MC mean of the proposed estimate assuming θ and ϕ known, SE is the MC mean of the estimated standard errors, MCSE is the standard error of MC estimates, CP is the empirical coverage probablity, KM is the Kaplan-Meier estimator ignoring misclassification.

35% Censoring			($\theta = 1.00$)					($\theta = 0.95$)					($\theta = 0.90$)				
	Parameter	Truth	KM	EST	MCSE	SE	CP%	KM	EST	MCSE	SE	CP%	KM	EST	MCSE	SE	CP%
($\phi = 1$)	$P(T^* > t_1)$	0.550	0.553	0.553	0.023	0.022	94.4	0.581	0.552	0.022	0.022	94.5	0.614	0.552	0.020	0.020	94.9
	$P(T^* > t_4)$	0.325	0.328	0.328	0.023	0.023	94.8	0.363	0.327	0.021	0.022	94.6	0.405	0.328	0.018	0.019	95.4
	$P(T^* > t_8)$	0.200	0.203	0.203	0.022	0.022	94.5	0.225	0.203	0.021	0.021	94.5	0.251	0.203	0.019	0.020	95.2
($\phi = 0.9$)	$P(T^* > t_1)$	0.550	0.503	0.553	0.021	0.021	94.8	0.532	0.552	0.019	0.019	94.6	0.565	0.552	0.018	0.018	94.5
	$P(T^* > t_4)$	0.325	0.326	0.328	0.022	0.022	95.0	0.362	0.328	0.021	0.020	94.6	0.406	0.327	0.019	0.019	94.5
	$P(T^* > t_8)$	0.200	0.198	0.203	0.022	0.022	94.6	0.220	0.203	0.021	0.020	94.3	0.247	0.203	0.019	0.019	94.9
($\phi = 0.8$)	$P(T^* > t_1)$	0.550	0.442	0.554	0.018	0.018	94.7	0.470	0.553	0.016	0.017	95.2	0.503	0.552	0.016	0.016	94.4
	$P(T^* > t_4)$	0.325	0.323	0.329	0.022	0.022	94.7	0.362	0.328	0.019	0.020	95.2	0.408	0.327	0.018	0.018	94.7
	$P(T^* > t_8)$	0.200	0.191	0.204	0.021	0.021	94.9	0.215	0.204	0.019	0.020	95.1	0.242	0.203	0.017	0.018	95.1

Table 8: Analysis results of the estimation of survival probabilities for time to viral negativity at selected time points. All is the overall estimated survival estimates by both methods. AA (n=196) and CA (n=205) stands for African Americans and Caucasians, respectively. KM is the Kaplan-Meier estimator ignoring misclassification. Proposed is our proposed estimate of the true survival. Right below the estimated survival probabilities are the 95% confidence intervals of the estimates; p-value compares survival rates between groups at the designated time point.

	KM				Proposed			
Parameter	All	AA	CA	p-value	All	AA	CA	p-value
$P(T^* > 1)$	0.983	0.995	0.971	0.23	0.995	0.998	0.991	0.23
	(0.969,0.995)	(0.985,1.00)	(0.948,0.994)		(0.991,0.999)	(0.995, 1.000)	(0.984, 0.998)	
$P(T^* > 14)$	0.892	0.922	0.862	0.24	0.953	0.967	0.939	0.24
	(0.861,0.922)	(0.885,0.960)	(0.815,0.910)		(0.938,0.967)	(0.950, 0.984)	(0.915, 0.962)	
$P(T^* > 168)$	0.275	0.367	0.185	0.10	0.532	0.610	0.458	0.07
	(0.228,0.321)	(0.296,0.439)	(0.127,0.242)		(0.499,0.564)	(0.563, 0.657)	(0.416, 0.499)	

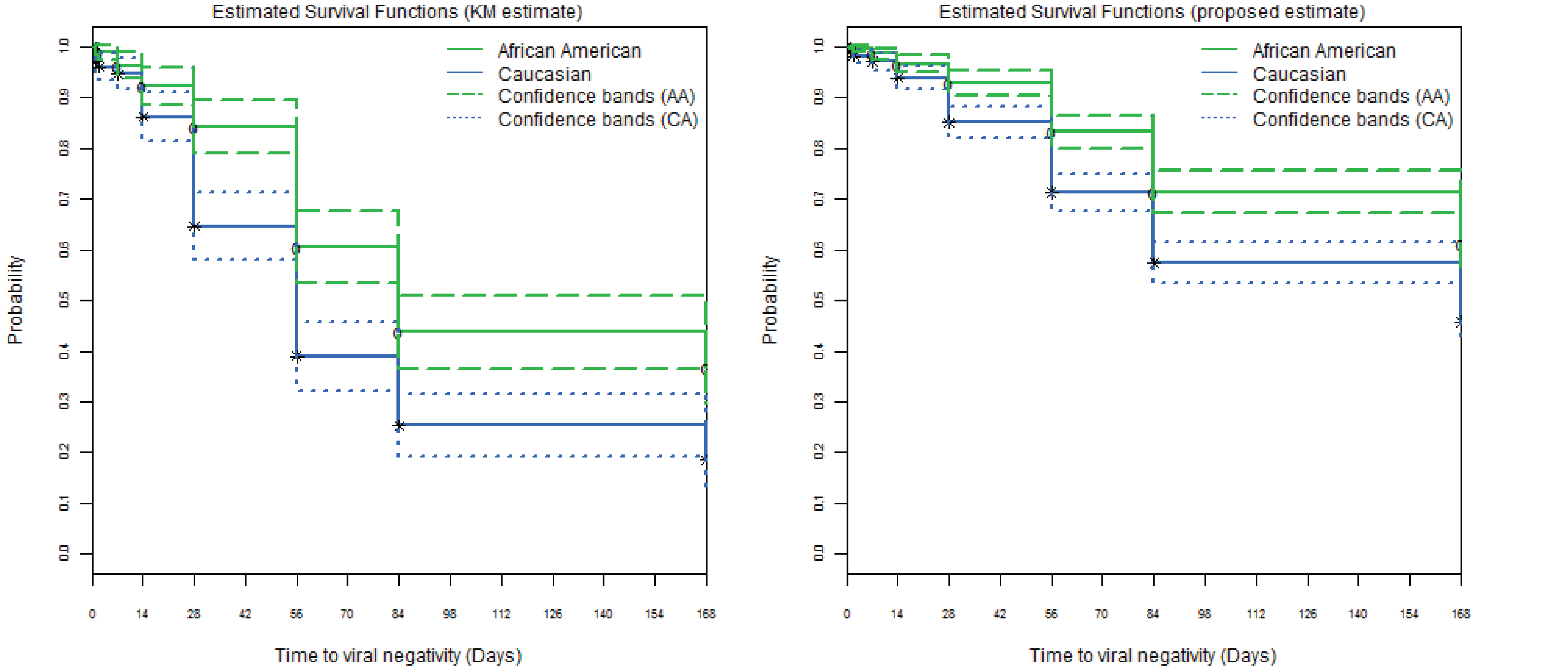

Figure 3: Estimated survival curves for time to viral negativity at selected time points for the VIRAHEP-C study. (Left panel: KM estimates; right panel: proposed method).

4.0 CONCLUSION

4.1 CONCLUSION AND DISCUSSION

The work presented in this paper is centered on correcting for error in clinical diagnosis. In the absence of a gold standard test, diagnostic results are prone to misclassification. We showed that the Kaplan-Meier estimator is biased in estimating the true survival distribution when events are prone to misclassification, as a result, we derived an unbiased and consistent estimator of the true survival distribution and we showed it to be asymptotically normal.

Future work on the issue of misclassified events in the analysis of time-to-event data can take several paths. An interesting course will be to assume no prior knowledge of the classification probabilities, the NPV and PPV could be directly estimated from the observed data. Another interesting problem would be to construct a discrete survival model, hence, estimate hazard ratios. Furthermore, an interesting problem is the development of a logrank test to compare survival distributions.

Estimating the true distribution of time to an event such as time to symptom resolution among subgroups of population with certain characteristics is important in improving public health. When the event is measured with error, the actual distribution cannot be estimated without bias, providing an inaccurate picture of the population. The new methods provide investigators with a tool to accurately estimate the survival probabilities in the presence of misclassified events. Our method offers the possibility of obtaining accurate measures of survival despite the use of a less expensive diagnostic test; hence, a cheaper study could be conducted.

4.2 PUBLIC HEALTH SIGNIFICANCE

Estimating the true distribution of time to an event such as time to symptom resolution among subgroups of population with certain characteristics is important in public health. When the event is measured with error, the actual distribution cannot be estimated without bias, providing an inaccurate picture of the population. The new methods provide clinical investigators with a tool to accurately estimate the survival probabilities in the presence of misclassified events.

BIBLIOGRAPHY

Raji Balasubramanian and Stephen W. Lagakos. Estimation of the timing of perinatal transmission of hiv. *Biometrics*, 57:1048–1058, 2001.

Raji Balasubramanian and Stephen W. Lagakos. Estimation of a failure time distribution based on imperfect diagnostic tests. *Biometrika*, 90:171–182, 2003.

M. Banerjee and J.A. Wellner. Confidence intervals for current status data. *Scandinavian Journal of Statistics*, 32:405–424, 2005.

CDC. Hepatitis c information for health professionals. "http://www.cdc.gov/hepatitis/HCV/HCVfaq.htm", March 2012.

Hari S. Conjeevaram, Michael W. Fried, Lennox J. Jeffers, Norah A. Terrault, Thelma E. WileyLucas, Nezam Afdhal, Robert S. Brown, Steven H. Belle, Jay H. Hoofnagle, David E. Kleiner, and Charles D. Howell. Peginterferon and ribavirin treatment in african american and caucasian american patients with hepatitis c genotype 1. *Gastroenterology*, 131:470–477, 2006.

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

Richard Gill. Large sample behavior of the product-limit estimator on the whole line. *The Annals of Statistics*, 11(1):49–58, 1983.

G.Jongbloed. The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics*, 7:310–321, 1998.

E.P. Simard G.M. McQuillan W.L. Kuhnert M.J. Alter G.L. Armstrong, A. Wasley. The prevalence of hepatitis c virus infection in the united states, 1999 through 2002. *Annals of Internal Medicine*, 144:705–714, 2006.

Major Greenwood. The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26, 1926.

Stephen P. Jenkins. Survival analysis. *Unpublished Manuscript, Institute for Social and Economic Research*, University of Essex, Colchester, 2008.

Nicholas P. Jewell and Mark J. van der Laan. Current status data: Review, recent developments and open problems. *U.C. Berkely Division of Biostatistics Working Paper Series*, 113, 2002.

Stephen Pianko Stuart C. Gordon Andrea E. Reid Jules Dienstag Timothy Morgan Ruji Yao Janice Albrecht John G. Mchutchison, Thierry Poynard. The impact of interferon plus ribavirin on response to therapy in black patients with chronic hepatitis c. *Gastroenterology*, 119(5):1317–1323, 2000.

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

John P. Klein and Melvin L. Moeschberger. *Survival Analysis*. Springer, New York, 2003.

T. Wong M. Wilkinson L. Y. Lee, C.Y. W. Tong. New therapies for chronic hepatitis c infection: a systematic review of evidence from clinical trials. *International Journal of Clinical Practice*, 66:342–355, 2012.

Kung-Yee Liang and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Laurence S. Magder and James P. Hughes. Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146:195–203, 1997.

P. Mccullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, New York, 1989.

Karen McKeown and Nicholas P. Jewell. Misclassification of current status data. *Lifetime Data Analysis*, 16:215–230, 2010.

Guy McKhann. Clinical diagnosis of alzheimer's disease: report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease. *Neurology*, 34:939–944, 1984.

Amalia S. Meier, Barbara A. Richardson, and James P. Hughes. Discrete proportional hazards models for mismeasured outcomes. *Biometrics*, 59(4):947–954, 2003.

John M. Neuhaus. Bias and efficiency loss due to misclassified response in binary regression. *Biometrika*, 86:843–855, 1999.

John M. Neuhaus. Analysis of clustered and longitudinal binary data subject to response misclassification. *Biometrics*, 58:675–683, 2002.

Amy Racine-Poon and David G. Hoel. Nonparametric estimation of the survival function when cause of death is uncertain. *Biometrics*, 40:1151–1158, 1984.

Barbara A. Richardson and James P. Hughes. Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics*, 1:341–354, 2000.

Victor G. Rosas and James P. Hughes. Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. *UW Biostatistics Working Paper Series*, 364, 2010.

Steven M. Snapinn. Survival analysis with uncertain endpoints. *Biometrics*, 54:209–218, 1998.

Leonard A. Stefanski and Dennis D. Boos. The calculus of m-estimation. *The American Statistician*, 56(1):29–38, 2002.

Nancy A. Obuchowski Xiao-Hua Zhou, Donna K. McClish. *Statistical Methods in Diagnostic Medicine*. Wiley, New York, 2002.

Scott L. Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130, 1986.