STATISTICAL METHODS FOR DEALING WITH OUTCOME MISCLASSIFICATION IN STUDIES WITH COMPETING RISKS SURVIVAL OUTCOMES

Philani Brian Mpofu

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Biostatistics,
Indiana University
February 2020

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Constantin Yiannoutsos, Ph.D., Co-Chair

Doctoral Committee

Giorgos Bakoyannis, Ph.D., Co-Chair

Wanzhu Tu, Ph.D.

September 19, 2019

Yiqing Song, Ph.D.

DEDICATION

To my parents, Joseph and Ruth Mpofu. And to my great teacher, Professor Ming-Wen An.

ACKNOWLEDGMENTS

I am eternally grateful to all the individuals, institutions and entities that made this work possible. Specifically, I would like to thank the National Institutes of Health and the East African consortium of the International Epidemiologic Databases to Evaluate HIV/AIDS (IeDEA) for funding my doctoral research under the supervision of Professor Constantin Yiannoutsos. I am especially grateful to Professor Constantin because he took a chance on me, introduced me to epidemiologic and statistical research in HIV/AIDS, and supported me, without wavering, during my doctoral studies. Under his tutelage I gained not only a passion for public health, but also a strong sense of purpose in my pursuits. To him I say, "ndinotenda, ngiyabonga, thank you".

My research was also made possible by Dr. Giorgos Bakoyannis, the co-chair of my research committee. In guiding me through the research process, Dr. Bakoyannis maintained a saintly-level of kindness and patience, and for that, I am grateful. As a a result of his generosity, my research skills in statistical methodology grew immensely. I am also grateful for the collaboration with and mentorship from Professor Wanzhu Tu and Professor Yiqing Song, who also served as advisors on my doctoral research committee.

Aside from developing statistical theory, my research involved analyzing data collected from people living with HIV/AIDS (PLWH) who are part of IeDEA East Africa. This aspect of my research was made possible by Beverly Musick, who prepared the data for analysis and provided many helpful insights on how to analyze the data.

I would also to like to extend my gratitude to Indiana University and Vassar College for funding my education, and providing me with the optimal conditions for learning and growing. I am grateful to the faculty at the Department of Biostatistics at Indiana University Fairbanks School of Public Health for the fantastic training in the theory and practice of Biostatistics: I am confident that the skills that I acquired will carry me far in my medical statistics career. To Vassar College, I shall forever be in debt for the invaluable undergraduate education.

Last but not least, I would like to thank my family for the unconditional love and support. Thank you Mum, Saphi ("Dad"), Babomncane Andries, Mamoncane Sitshengisiwe, Sekuru Victor, Gogo Nomsa, and the rest of my family, who for the lack of space I shall not individually name. It really took a village to get me here.

Philani Brian Mpofu

STATISTICAL METHODS FOR DEALING WITH OUTCOME MISCLASSIFICATION IN STUDIES WITH COMPETING RISKS SURVIVAL OUTCOMES

In studies with competing risks outcomes, misidentifying the event-type responsible for the observed failure is, by definition, an act of misclassification. Several authors have established that such misclassification can bias competing risks statistical analyses, and have proposed statistical remedies to aid correct modeling. Generally, these rely on adjusting the estimation process using information about outcome misclassification, but invariably assume that outcome misclassification is non-differential among study subjects regardless of their individual characteristics. In addition, current methods tend to adjust for the misclassification within a semi-parametric framework of modeling competing risks data. Building on the existing literature, in this dissertation, we explore the parametric modeling of competing risks data in the presence of outcome misclassification, be it differential or non-differential. Specifically, we develop parametric pseudo-likelihood-based approaches for modeling cause-specific hazards while adjusting for misclassification information that is obtained either through data internal or external to the current study (respectively, internal or external-validation sampling). Data from either type of validation sampling are used to model predictive values or misclassification probabilities, which, in turn, are used to adjust the cause-specific hazard models. We show that the resulting pseudo-likelihood estimates are consistent and asymptotically normal, and verify these theoretical properties using simulation studies. Lastly, we illustrate the proposed methods using data from a study involving people living with HIV/AIDS (PLWH)in the East-African consortium of the

International Epidemiologic Databases for the Evaluation of HIV/AIDS (IeDEA EA). In this example, death is frequently misclassified as disengagement from care as many deaths go unreported to health facilities caring for these patients. In this application, we model the cause-specific hazards of death and disengagement from care among PLWH after they initiate anti-retroviral treatment, while adjusting for death misclassification.

Constantin Yiannoutsos, Ph.D., Co-Chair

Giorgos Bakoyannis, Ph.D., Co-Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Translating the Research Aim to a Statistical Aim

The East-African International Epidemiologic Databases to Evaluate HIV/AIDS (IeDEA EA) is data consortium is funded by the US National Institutes of Health for the purpose of “collecting, merging, harmonizing, sharing and analyzing” data from people living with HIV/AIDS (PLWH) in East Africa (Web Page 2019). The countries that are part of this data consortium include Kenya, Uganda and Tanzania.

One goal for treatment programs that contribute to IeDEA EA is to retain PLWH in HIV care so as to reduce HIV-related mortality. Part of this mission entails understanding the risk factors of death and disengagement from care among patients who receive HIV care. Clarity on the risk factors can help modify the treatment programs in ways that reduce attrition from care and death among HIV patients (Schafer and Graham 2002; Bakoyannis and Yiannoutsos 2015).

The study of the risks factors of death and disengagement from care requires statistical methods within competing-risks survival analysis. Namely, the modeling of cause-specific hazards. To see why this is case, I will provide the reader with a gentle and general introduction to competing risks survival analysis. The narrative will also describe some of the contextual challenges for HIV treatment programs that render standard competing risks methodology insufficient for addressing the modeling problems at hand.

1.1 An Introduction to Competing-risks Survival Analysis

Competing risks survival analysis is a branch of time-to-event analysis wherein an individual (study unit) can fail from any one event that is within a set of mutually-exclusive competing events (Kalbfleisch and Prentice 2011). Of interest is the time to the first occurring event-type. For example, in a mortality study where interest lies in the time to death due either cancer or heart disease, the causes of death are considered to be competing risks. Observing death due to cancer precludes us from observing death due to heart disease, and vice versa.

In the language of stochastic processes, a competing-risks system can be viewed as the simplest form of a multi-state process wherein the initial state of the process is the only transient state, and all the other states, corresponding to competing events, are absorbing states(Aalen, Borgan, and Gjessing 2008). Such a process can be visualized as shown in Figure 1.1.

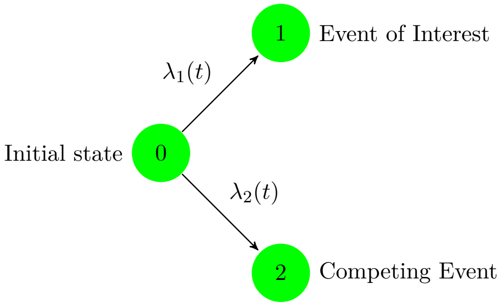

Figure 1.1: Multi-state process process representing competing risks

It is also worth noting that non-mutually exclusive events can also be considered competing risks if the time to event is computed based whichever event comes first (Austin, Lee, and Fine 2016). Death and disengagement from care in our motivating study, for example, are not mutually exclusive: Death can still be observed after a patient disengages from care (the opposite is, however, not true).

1.1.1 Definitions of Important quantities

The presence of competing risks requires the definition of quantities/measures beyond those encountered in standard survival analysis. I will introduce the reader to some of these quantities with strong emphasis placed on those commonly used in biomedical studies. In this introduction and subsquent chapters of this dissertation, I shall use the terms event-type, cause, and cause of failure interchangeably.

Assuming we have a m-cause competing-risks system wherein a subject can fail from any one of m causes, let the true cause of failure be represented by $C \in \{1, 2, ..., m\}$. Let U be the failure time; V be the censoring time, and T be the right-censored failure time, where $T = \min(U, V)$. Assume that U and V are independent, and that censoring distribution is independent of the cause of failure. Lastly, let $\boldsymbol{Z}$ be the subject characteristics. For each of the n subjects, $i = 1, 2..., n$, we observe independent 3-tuples of the form $(T_i, C_i, \boldsymbol{Z}_i)$.

In standard survival analysis, the distribution of failure times is usually specified using either the hazard function $\lambda(t; \boldsymbol{Z})$, the survival function $S(t|\boldsymbol{Z})$, or the probability density function $f(t|Z)$ (Kalbfleisch and Prentice 2011). These functions are still relevant in competing risks survival analysis, provided the competing events are treated as a composite outcome, with time to event defined as the time to any of the competing events. For such a composite outcome of competing events, the technical definitions of the aforementioned (survival analysis) functions stay the same, for example, the hazard function is defined as shown by Equation 1.1:

$$\lambda(t; \boldsymbol{Z}) = \lim_{h \to 0} \frac{P[t \leq T < t + h | T \geq t,\, \boldsymbol{Z}]}{h} \tag{1.1}$$

and the survival function is defined as shown in Equation 1.2:

$$S(t|\boldsymbol{Z}) = P[T > t|\boldsymbol{Z}] = \exp\left(-\int_0^t \lambda(u|\boldsymbol{Z})du\right) \tag{1.2}$$

The study the time to event for each of the mutually-exclusive events that comprise a competing-risks process requires functions including: the cause-specific hazard function, the (sub) density function, and the cumulative incidence function.

The *cause-specific hazard* of cause-j at time t is given by Equation 4.2.

$$\lambda_j(t; \boldsymbol{Z}) = \lim_{h \to 0} \frac{P[t \leq T \leq t + h,\, J = j | T \geq t,\, \boldsymbol{Z}]}{h} \tag{1.3}$$

for $j = 1, 2..., m$.

Colloquially, $\lambda_j(t; \boldsymbol{Z})$ is defined as the instantaneous rate for failure due to cause j at time t given the covariate pattern $\boldsymbol{Z}$, *in the presence of other causes of failure* (Kalbfleisch and Prentice 2011). Notice that this definition recognizes that study units live in a world where they can fail from other causes besides cause j.

If only one cause of failure can occur, the overall hazard as defined by Equation 1.1 is equal to the sum of the cause-specific hazards associated with the mutually-exclusive events that comprise the competing-risks system. That is,

$$\lambda(t; \boldsymbol{Z}) = \sum_{j=1}^{m} \lambda_j(t; \boldsymbol{Z})$$

Proof:

$$\begin{aligned}
\lambda(t; \boldsymbol{Z}) &= \lim_{h \to 0} \frac{P[t \leq T < t + h | T \geq t,\, \boldsymbol{Z}]}{h} \\
&= \lim_{h \to 0} \frac{P[\cup_{j=1}^{m} (t \leq T < t + h, J = j) | T \geq t,\, \boldsymbol{Z}]}{h}, \text{ since only one cause of failure can occur} \\
&= \lim_{h \to 0} \frac{\sum_{j=1}^{m} P[t \leq T < t + h, J = j | T \geq t,\, \boldsymbol{Z}]}{h} \\
&= \sum_{j=1}^{m} \lim_{h \to 0} \frac{P[t \leq T < t + h, J = j | T \geq t,\, \boldsymbol{Z}]}{h} \\
&= \sum_{j=1}^{m} \lambda_j(t; \boldsymbol{Z})
\end{aligned}$$

Recalling that, $S(t|\boldsymbol{Z}) = \exp(-\int_0^t \lambda(u|\boldsymbol{Z}) du)$, it would follow that under competing risks, the survival function can also be defined as follows:

$$S(t|\boldsymbol{Z}) = P[T > t|\boldsymbol{Z}] = \exp\left[-\int_0^t \sum_{j=1}^{m} \lambda_j(u|\boldsymbol{Z}) du\right]$$

According to Kalbfleisch and Prentice, the *sub-density function* for the time to event-type j, given $\boldsymbol{Z}$, is defined as shown by Equation 1.4 (Kalbfleisch and Prentice 2011).

$$f_j(t; \boldsymbol{Z}) = \lim_{h \to 0} \frac{P[t \leq T \leq t + h, J = j|\boldsymbol{Z}]}{h} = \lambda_j(t; \boldsymbol{Z}) S(t; \boldsymbol{Z}) \tag{1.4}$$

Proof:

$$\begin{aligned}
f_j(t;\boldsymbol{Z}) &= \lim_{h\to 0}\frac{P[t\le T<t+h, J=j|\boldsymbol{Z}]}{h}\\
&= \lim_{h\to 0}\left[\frac{P[(t\le T<t+h, J=j)\cap(T\ge t)|\boldsymbol{Z}]}{h}\right]\\
&+ \lim_{h\to 0}\left[\frac{P[(t\le T<t+h, J=j)\cap(T<t)|\boldsymbol{Z}]}{h}\right]\\
&= \lim_{h\to 0}\frac{P[(t\le T<t+h, J=j)\cap(T\ge t)|\boldsymbol{Z}]}{h}\\
&= \lim_{h\to 0}\frac{P[t\le T<t+h, J=j|T\ge t,\boldsymbol{Z}]P[T\ge t|\boldsymbol{Z}]}{h}\\
&= \lim_{h\to 0}\frac{P[t\le T<t+h, J=j|T\ge t,\boldsymbol{Z}]S(t;\boldsymbol{Z})}{h}\\
&= S(t;\boldsymbol{Z})\lim_{h\to 0}\frac{P[t\le T<t+h, J=j|T\ge t,\boldsymbol{Z}]}{h}\\
&= \lambda_j(t;\boldsymbol{Z})S(t;\boldsymbol{Z})
\end{aligned}$$

Of note is that the sum of the *sub-density functions* is equal to the *overall density* for the time to any event (a composite of the competing events). That is, $f(t|\boldsymbol{Z})=\sum_{j=1}^{m} f_j(t|\boldsymbol{Z})$. The proof of this relationship is provided below:

$$\begin{aligned}
\sum_{j=1}^{m} f_j(t|\boldsymbol{Z}) &= \sum_{j=1}^{m}\lambda_j(t|\boldsymbol{Z})S(t|\boldsymbol{Z})\\
&= S(t|\boldsymbol{Z})\sum_{j=1}^{m}\lambda_j(t|\boldsymbol{Z})\\
&= S(t|\boldsymbol{Z})\lambda(t|\boldsymbol{Z})\\
&= f(t|\boldsymbol{Z})
\end{aligned}$$

Lastly, the *cumulative incidence function* for cause-j time t is defined as follows:

$$
\begin{aligned}
F_j(t;\boldsymbol{Z}) &= P[T \leq t, J = j; \boldsymbol{Z}] \\
&= \int_0^t f_j(u;\boldsymbol{Z})du \\
&= \int_0^t \lambda_j(u;\boldsymbol{Z})S(u;\boldsymbol{Z})du \\
&= \int_0^t \lambda_j(u;\boldsymbol{Z}) \exp\left[-\int_0^t \sum_{j=1}^{m} \lambda_j(u;\boldsymbol{Z})du\right] du
\end{aligned}
\tag{1.5}
$$

For all $j \in \{1, 2, ..., m\}$, an important constraint is that $\sum_{j=1}^{m} F_j(t;\boldsymbol{Z}) \leq 1$.

Competing risks functions as presented by Equations 1.3, 1.4 and 1.5 have interpretations that preserve the fact the event of interest, j, exists in a world where other event-types are competing with event-type j. Moreover, when defining the competing risks functions, one need not make assumptions about the inter-relationships among event-types that comprise the competing risks process.

1.1.2 Competing Risks in Biomedical Studies

In biomedical studies, interest usually lies in modeling cause-specific hazards and cumulative incidence functions. These quantities nicely align with questions that are encountered in biomedical studies involving competing events. The cumulative incidence function, for example, provides an absolute measure of the accumulated risk of failure due a particular cause, which in turn can be used for making clinical predictions and decisions (Hinchliffe, Abrams, and Lambert 2013). Cause-specific hazards, on the other hand, are used to identify the factors that influence the rate of occurrence of a particular event, in the presence of other competing events (Hinchliffe, Abrams, and Lambert 2013; Austin, Lee, and Fine 2016).

As already noted, in the motivating study involving PLWH in East Africa, one research aim is to identify the factors that are associated with the incidence of death and disengagement from care among patients who initiate anti-retroviral therapy (ART). From a public health standpoint, it is in the best interests of treatment programs to ensure low mortality and low attrition among patients who initiate care. In other words, treatment programs would like for patients to go for as long as possible without disengaging from care or dying: Death and disengagement from care are undesirable outcomes. In order to study the time to observing an undesirable outcome, death and disengagement from care are treated as competing for the status of *first undesirable event*. From this vantage point, the research aim can be addressed by modeling the cause-specific hazards of the respective causes of failure. Under an assumed statistical model, usually this would entail using maximum likelihood estimation to estimate parameters associated with the cause-specific hazards.

1.1.3 Likelihood

For each of the n subjects, $i = 1, 2..., n$, we observe $(T_i, C_i, \boldsymbol{Z}_i)$, where the observations are assumed to be independent and identically distributed. The observed cause-of-failure, C_i, can also be expressed as vector of dummy (binary) random variables by defining the following:

1. $\delta_{ij} = \mathrm{I}[C_i = j]$ to be the indicator that subject i to fail due to cause j
2. $\delta_i = \sum_{j=1}^{m} \delta_{ij}$ to be the any-cause failure indicator for subject i,

so that for each subject i we observe, $(T_i, \delta_i, \boldsymbol{Z}_i)$.

Given the observed data, under right-censoring, the likelihood is defined as follows:

$$
\begin{aligned}
L &\propto \prod_{i=1}^{n} f(t_i; \boldsymbol{Z}_i)^{\delta_i} [S(t_i; \boldsymbol{Z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} f(t_i, c_i = j; \boldsymbol{Z}_i) \right)^{\delta_i} [S(t_i; \boldsymbol{Z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f_j(t_i; \boldsymbol{Z}_i)^{\delta_{ij}} \right)^{\delta_i} [S(t_i; \boldsymbol{Z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} [\lambda_j(t_i; \boldsymbol{Z}_i) S(t_i; \boldsymbol{Z}_i)]^{\delta_{ij}} \right)^{\delta_i} [S(t_i; \boldsymbol{Z}_i)]^{1-\delta_i} \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} \lambda_j(t_i; \boldsymbol{Z}_i)^{\delta_{ij}} \right)^{\delta_i} S(t_i; \boldsymbol{Z}_i) \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} \lambda_j(t_i; \boldsymbol{Z}_i)^{\delta_{ij}} \right) \exp\left[-\sum_{j=1}^{m} \int_0^{t_i} \lambda_j(u; \boldsymbol{Z}_i) du \right] \\
&= \prod_{i=1}^{n} \left(\prod_{j=1}^{m} \lambda_j(t_i; \boldsymbol{Z}_i)^{\delta_{ij}} \exp\left[- \int_0^{t_i} \lambda_j(u; \boldsymbol{Z}_i) du \right] \right) \\
&= \prod_{j=1}^{m} \left(\prod_{i=1}^{n} \lambda_j(t_i; \boldsymbol{Z}_i)^{\delta_{ij}} \exp\left[- \int_0^{t_i} \lambda_j(u; \boldsymbol{Z}_i) du \right] \right) \\
&= \prod_{j=1}^{m} L_j
\end{aligned}
\tag{1.6}
$$

From likelihood 1.6, it can be deduced that:

1. The likelihood is a function of the cause-specific hazards, λ_j.
2. The likelihood factors into separate components for each failure type.
3. L_j is likelihood component corresponding to λ_j, assuming all events besides cause-j are treated as censored events.
4. If there are no common parameters among the different causes of failure, then maximum likelihood estimation can be applied to each of the components of the likelihood separately.

For more on the derivation of likelihood 1.6, and the deductions (1-4) above, reader can refer to (Kalbfleisch and Prentice 2011).

1.1.4 Model Specification

The cause-specific hazards that define likelihood 1.6 can be specified parametrically, semi-parametrically or non-parametrically. That said, semi-parametric modeling usually prevails due to the popularity of the Cox proportional hazards model. The Cox model has wide appeal because it relies on fewer assumptions than parametric models (Nardi and Schemper 2003).

Under the proportional hazards assumption, the cause-specific hazard function for cause-j is defined as follows:

$$\lambda_j(t; \boldsymbol{\theta}_j, \boldsymbol{Z}) = \lambda_{0j}(t) \exp\left(\boldsymbol{Z}^T \boldsymbol{\theta}_j\right)$$

for $j = 1, 2, ..., m$, where:

a. $\lambda_{0j}(t)$ is an arbitrary function that defines the baseline cause-specific hazard for cause-j at time t.

b. $\boldsymbol{Z}$ is a matrix of covariates that are assumed to explain the cause-specific hazard for cause-j.

c. $\boldsymbol{\theta}_j$ is a vector of coefficients reflecting the changes in the log hazards for marginal changes in the covariates.

1.2 Death Under-reporting in HIV/AIDS Treatment Programs

Thus far, I have noted that IeDEA East Africa is interested in identifying the risk factors of death and disengagement from care. I also noted these risk factors may be identified by modeling cause-specific hazards, assuming death and disengagement from care are competing risks. After this, I introduced the reader to the basics of modeling cause-specific hazards. I will now present why ordinary statistical methods for modeling cause-specific hazards may not suffice in IeDEA East Africa.

Treatment programs that contribute data to IeDEA East Africa often face a challenge of death under-reporting or under-ascertainment (Elvin H Geng et al. 2011; Yiannoutsos et al. 2008). This death under-reporting is actually a form of outcome misclassification as some patients who end up being classified as disengaged from care may actually be dead (Bakoyannis and Yiannoutsos 2015). Such outcome misclassification is problematic as it may lead to estimation bias and loss of power when performing competing analyses (Van Rompaye, Jaffar, and Goetghebeur 2012; Hinchliffe, Abrams, and Lambert 2013; Bakoyannis and Yiannoutsos 2015). We, therefore, need to be wary of death misclassification when modeling the cause-specific hazards of death and disengagement from care among PLWH enrolled in IeDEA.

To establish some clarity about the gravity of the problem of outcome misclassification when modeling cause-specific hazards, let's take a short detour and explore the problem via simulations.

1.2.1 Studying the Effects of Misclassification on Modeling Cause-specific Hazards

Take for example a censoring-free, two-cause competing risks system where subjects are followed until they fail from either cause-1 or cause-2. The true cause of failure is represented by $C \in \{1, 2\}$. Given the possibility of outcome misclassification, we observe $C^* \in \{1, 2\}$, where C^* is not necessarily consistent with the true outcome, C. Outcome misclassification for this two-cause system is illustrated in Figure 1.2.

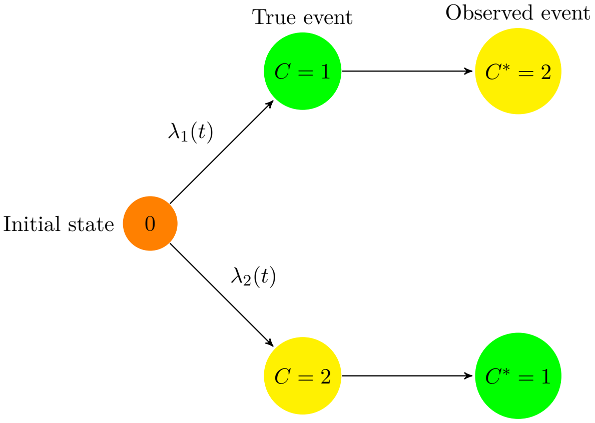

Figure 1.2: Illustration of bi-directional outcome misclassification

Assume that the respective cause-specific hazards for cause-1 $(C = 1)$ and cause-2 $(C = 2)$ are defined as follows:

a. $\lambda_1(t; \theta_1, Z) = \exp(Z\theta_1)$

b. $\lambda_2(t; \theta_2, Z) = \exp(Z\theta_2)$

where $Z \sim N(0, 1)$, $\theta_1 = 1$ and $\theta_2 = -0.5$.

In addition, let's assume that there is uni-directional misclassification, such that $P[C^* = 1|C = 2] = \tau$ and $P[C^* = 2|C = 1] = 0$. Colloquially, this means that some subjects are observed as having failed from cause-1 when, in fact, they failed from cause-2.

I varied τ from 0% to 20% in steps of 2%. At each simulation setting (i.e at each τ), I generated 1000 datasets of sample size 1000. Using these datasets, I modeled the cause-specific hazard for cause-1, and noted the average estimate: $\bar{\theta}_1 = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\theta}_{1i}$; the bias percent of the average estimate: $100 \times \frac{\bar{\theta}_1 - 1}{1}\%$; the asymptotic standard error of the average estimate: $\text{SE}\left(\bar{\theta}_1\right) = \frac{1}{1000}\sum_{i=1}^{1000} SE\left(\hat{\theta}_{1i}\right)$; the Monte-Carlo standard deviation: $\sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}\left(\hat{\theta}_{1i} - \bar{\theta}_1\right)^2}$; and the 95% coverage probability. The results of the simulation study were as presented in Table 1.1.

	mis_rate	truth	estimate	bias_perc	ase	mcsd	cp
1	0.000	1.000	1.000	0.040	0.056	0.054	95.300
2	2.000	1.000	0.970	2.970	0.055	0.057	89.900
3	4.000	1.000	0.942	5.850	0.055	0.056	79.400
4	6.000	1.000	0.915	8.480	0.054	0.057	62.500
5	8.000	1.000	0.888	11.180	0.054	0.054	45.400
6	10.000	1.000	0.860	13.990	0.053	0.054	23.900
7	12.000	1.000	0.839	16.080	0.052	0.052	13.900
8	14.000	1.000	0.813	18.700	0.052	0.054	5.600
9	16.000	1.000	0.790	20.980	0.051	0.055	3.100
10	18.000	1.000	0.767	23.330	0.051	0.053	1.400
11	20.000	1.000	0.745	25.520	0.050	0.052	0.200

Table 1.1: Results of simulation study examining the the effect of outcome misclassification on estimation

The simulation illustrates that as the extent of misclassification, among those who truly failed from cause-2, increases from 0% to 20%, the bias of the point-estimate increases, and the coverage probability decreases (thereby indicating an increase in the Type-I error away from the desired 5% level.). The simulation results were graphed as shown in Figure 1.3.

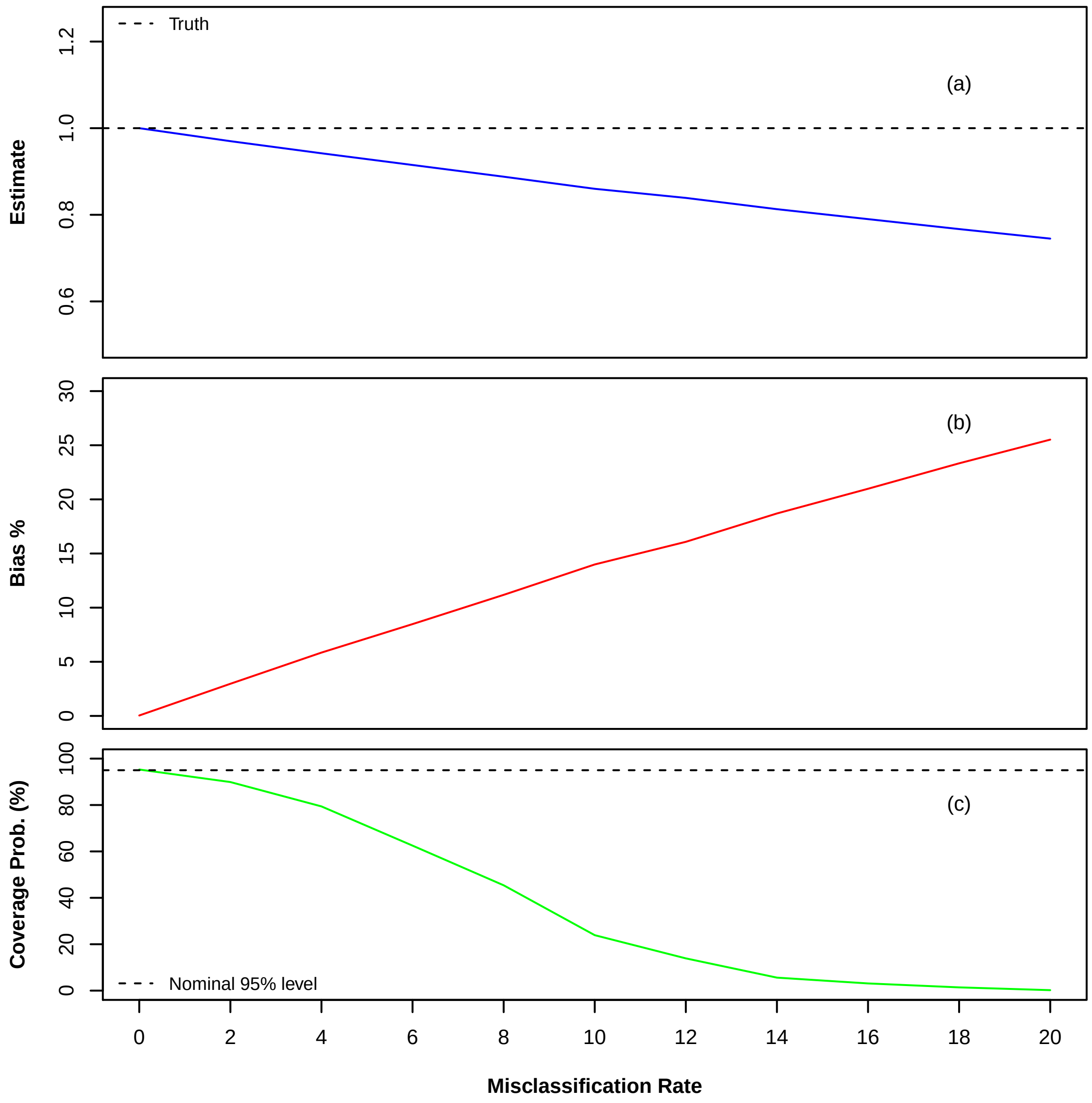

Figure 1.3: Plots of the simulation results. Panel (a) depicts how the estimates change as the extent of misclassification increases. Panel (b) shows the change in the bias percent as misclassification increases. Panel (c) shows the change in coverage probability as misclassification increases.

1.3 Cause-specific Hazards in the Presence of Misclassification

Since outcome misclassification may lead to incorrect modeling of cause-specific hazards as shown in Section 1.2.1, the statistical aim has to be modified to align with the contextual challenges. Although the research aim remains that of identifying the risk factors of death and disengagement from care, the statistical aim becomes that of correctly modeling cause-specific hazards in the presence misclassification among the competing events. This statistical aim is crux of the dissertation, and entails two sequential aspects: First, the quantification of outcome misclassification, and second, the adjustment of the estimating procedure for outcome misclassification.

In the fulfilling of the statistical aim, I shall place emphasis on developing statistically-principled methods that not only align well with the research aim, but are also easy to understand, implement, and share with research partners such as the UNAIDS. Research partners typically use mathematical modeling, therefore, I will only explore parametric modeling solutions. One such solution has been proposed Gravel et. al (2018) for the purpose of modeling cause-specific hazards in the presence of misclassification. The solution by Gravel et al. requires a full-likelihood specification, and depends on the availability of an internal-validation sample. An internal validation sample serves as source of misclassification information, and is created by re-ascertaining outcomes on a subset of the main-study sample using a gold-standard approach (R. J. Carroll et al. 2006). Other authors have proposed methods in the non-parametric realm. For example, Van Rompaye et al. (2012) proposed adjusting Cox models of cause-specific hazards using misclassification probabilities. The approach by Van Rompaye et al. (2012) assumes that misclassification probabilities are known and non-differential. Bakoyannis et al. (2019) proposed specifying the misclassification problem as a missing cause of failure problem if an internal validation sample is available. Ha and Tsodikov (2012), on the other hand, adopted a fully non-parametric for modeling cause-specific hazards while adjusting for misclassification (Ha and Tsodikov 2012). The shortcomings in the existing literature will be extensively presented in Chapter 3 of this dissertation. For now, I will only present basic structures of how I intend to solve the problem of modeling cause-specific hazards in the presence of misclassification. The structures presented rely on validation sampling, internal or external, for information on misclassification.

1.3.1 Internal-validation Based Solution

The outcome-misclassification problem that has been described for IeDEA, ideally, can be remedied by re-ascertaining the outcomes for everyone deemed disengaged from care. Such a remedy would result in error-free data, that can then be used to model the cause-specific hazards of death and disengagement. Although ideal, such as solution is infeasible as it is expensive to employ, especially for resource-poor treatment programs that are part of IeDEA East Africa.

For IeDEA, the next best solution is internal validation or double-sampling (Tenenbein 1970; Greenland 1988 Rosner, Spiegelman, and Willett (1990); Spiegelman 2010). This involves re-ascertaining, through outreach, the vital-status data on a sub-sample of those initially deemed to be disengaged from care(E. H. Geng et al. 2008; Yiannoutsos et al. 2008; An et al. 2009). Patient outreach is considered to be a gold-standard outcome ascertainment approach as it results in more accurate outcome data than the initial outcome-ascertainment procedure.

To understand the utility of internal validation, such as the one described for IeDEA

EA, let's step back and examine the information that is generated through internal validation or double-sampling. When double-sampling is performed, the resulting validation sample contains useful information about the concordance or lack thereof among the *true* and *observed/misclassified* outcomes. *True* outcomes are those ascertained through a gold-approach, and *observed/misclassified* outcomes are those ascertained using an error-prone approach. Using data from the validated sub-sample of the main study sample, at minimum one can model:

1. *Predictive values*: $p_{jk}(\boldsymbol{\eta}_k) = P[C = j|C^* = k, \boldsymbol{\eta}_k]$
2. *Misclassification probabilities*: $\pi_{jk}(\boldsymbol{\beta}_k) = P[C^* = j|C = k, \boldsymbol{\beta}_k]$

where $j, k \in \{1, 2\}$.

Predictive values and misclassification probabilities are measures of concordance or lack thereof between the true and the observed outcomes. The difference between these quantities is the conditioning or what is assumed to be known. When modeling predictive values, in this case, the observed outcomes are assumed to known. And when modeling misclassification probabilities, the true outcomes are assumed to be known. Predictive values and misclassification probabilities carry the information required to adjust to the likelihood in a manner that enables correct estimation. Whether to use predictive values or misclassification probabilities as an adjustment depends on whether one wants to frame the problem as either a missing-data problem or a misclassification problem.

1.3.1.1 Framing Problem as a Missing-data Problem

For a two-cause example where the true outcome $C \in \{1, 2\}$ and the observed outcome $C^* \in \{1, 2\}$, when double-sampling is performed, the resulting data takes on a structure depicted by Table 1.2.

	Outcome	
Double-Sampled	Observed	True
Yes	1	1
		2
No	🟦	Missing
Yes	2	2
		1
No	🟥	Missing

Table 1.2: Data structure after double sampling

Table 1.2 also illustrates that double-sampling results in a dataset where some of the outcome data are missing-by-design. For example, if the 50% double-sampling is performed, then 50% of the true outcome data will be missing-by-design.

Let $R_i = 1$ indicate that the true outcome for subject i was observed(in other words, the subject's outcome was successfully validated). And also recall that, when modeling cause-specific hazards, the goal is to maximize a log-likelihood of the form presented in Equation 1.7.

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{m} \sum_{i=1}^{n} \left\{ \delta_{ij} \log \lambda_j(t_i; \boldsymbol{\theta}_j \boldsymbol{Z}_i) - \int_0^{t_i} \lambda_j(u; \boldsymbol{\theta}_j \boldsymbol{Z}_i) du \right\} \tag{1.7}$$

In Equation 1.7, δ_{ij} is observable only when $R = 1$. As a result, the log-likelihood can be re-written as shown by Equation 1.8.

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{m}\sum_{i=1}^{n}\left\{(R_i\delta_{ij} + (1 - R_i) \times \delta_{ij})\log\lambda_j(t_i;\boldsymbol{\theta}_j\boldsymbol{Z}_i) - \int_0^{t_i}\lambda_j(u;\boldsymbol{\theta}_j\boldsymbol{Z}_i)du\right\} \tag{1.8}$$

Equation 1.8 illustrates that the events from subjects where $R_i = 0$ are not considered when modeling cause-specific hazards. That is, subjects who are missing *true* outcome values are excluded. Therefore, there is a missing-data problem.

The missing-data problem that has been identified can be solved by recognizing that in log-likelihood 1.8, δ_{ij} is linear in the log-likelihood. Such linearity ensures that consistent estimation can also be performed by replacing missing values of δ_{ij} by their conditional expectations. In this case, replacement with $E[\delta_j|\delta^*, \boldsymbol{Z}] = P[C = j|C^*, \boldsymbol{Z}]$–the predictive values that can be modeled using data from internal-validation sampling as noted in Section 1.3.1. Before making the aforementioned replacement for missing data, one needs to deliberate about *mechanisms of missingness.*

1.3.1.2 Missing data mechanisms

Prior to performing estimation in the presence of missing data, one needs ask why the data are missing, and whether the missing data have implications to estimation and inference (J. Carpenter and Kenward 2012). After making sense of why data are missing, one needs to make assumptions about the random process that results in the missing data (Rubin 1976). These random processes are also known as *missingness mechanisms.* Rubin proposed that missingness mechanisms be classified into 3 groups, namely:

1. Missing completely at random (MCAR)
2. Missing at random (MAR)

3. Missing not at random (MNAR)

To define the missingness mechanisms in the manner set forth by Rubin, let $Y_{comp} = (Y_{obs}, Y_{miss})$, where Y_{comp} represents complete outcome data, Y_{obs} represents observed outcome data, and Y_{miss} represents missing outcome data. In addition, let $M = 1$ indicate that data are missing, and X represent other observed data.

Data are said to be missing completely at random (MCAR) if:

$$P[M = 1|Y_{obs}, Y_{miss}, X] = P[M = 1]$$

That is, missingness is independent of all observed data. In such a situation, one can ignore missing data, and proceed with estimation (Rubin 1976; Aalen, Borgan, and Gjessing 2008; R. J. Little and Rubin 2014).

Data are missing at random (MAR) if missingness is dependent on the observed data and independent of the missing data, that is:

$$P[M = 1|Y_{obs}, Y_{miss}, X] = P[M = 1|Y_{obs}, X]$$

When one has MAR data, one can still use likelihood-based methods for perform parameter estimation, provided the parameter that defines the missing data is distinct from the parameter of interest (Rubin 1976; Schafer and Graham 2002; R. J. Little and Rubin 2014).

Lastly, data are said to be missing not at random (MNAR) if missingness is dependent on the missing data. That is:

$$P[M = 1|Y_{obs}, Y_{miss}, X] \neq P[M = 1|Y_{obs}, X]$$

In the motivating study involving PLWH in East Africa, it is untenable to assume that data are MCAR because not everybody chosen for internal validation (outreach) is available to provide a response. As a result, the missingness assumption is relaxed MAR: That is, we allow missingness to be explained by observed patient characteristics.

1.3.1.3 Implications MAR in IeDEA context

Recalling that $R = 1$ indicates that the true outcome has been observed, it follows that $R = 0$ indicates that the true outcome is missing. Given this notation, under a two-cause competing risks process, the MAR assumption is written as follows,

$$P[R = 0|C^* \neq 0, C = j, \boldsymbol{Z}] = P[R = 0|C^* \neq 0, \boldsymbol{Z}]$$

for $j \in \{1, 2\}$. According to (Bakoyannis, Siannis, and Touloumi 2010), such a MAR assumption implies that:

$$P[C = j|C^* > 0, R = 0, \boldsymbol{Z}] = P[C = j|C^* > 0, R = 1, \boldsymbol{Z}] = P[C = j|C^* > 0, \boldsymbol{Z}] \qquad (1.9)$$

In words, Equation 1.9 means that the predictive value model from those whose outcomes were validated is the same as the one for those whose outcomes were not validated. Therefore, under the MAR assumption, the missing outcome data, in log-likelihood 1.8, can be replaced by predictive values computed using a predictive-value model derived for those whose outcome data were validated.

1.3.1.4 Framing the Problem as a Misclassification Problem

The problem of modeling cause-specific hazards in the presence of misclassification can also be solved by adjusting estimation using misclassification probabilities. The utility of outcome misclassification probabilities becomes clear when we consider the cause-specific hazards of the *observed/misclassified* outcome.

By definition, the cause-specific hazard for observed cause-j is,

$$\lambda_j^*(t; \boldsymbol{Z}) = \lim_{h \to 0} \frac{P\left(t \leq T < t+h, C^* = j | T \geq t, \boldsymbol{Z}\right)}{h} \tag{1.10}$$

The cause-specific hazard as shown by Equation 1.10 can be re-written as follows:

$$\begin{aligned}
\lambda_j^*(t; \boldsymbol{Z}) &= \lim_{h \to 0} \frac{P\left(t \leq T < t+h, C^* = j | T \geq t, \boldsymbol{Z}\right)}{h} \\
&= \lim_{h \to 0} \frac{\sum_{i=1}^{k} P\left(t \leq T < t+h, C^* = j, C = i | T \geq t, \boldsymbol{Z}\right)}{h}, \text{ by law of total probability,} \\
&= \lim_{h \to 0} \frac{\sum_{i=1}^{k} P\left(t \leq T < t+h, C = i | T \geq t, \boldsymbol{Z}\right) P\left(C^* = j | t \leq T < t+h, C = i, \boldsymbol{Z}\right)}{h} \\
&= \sum_{i=1}^{k} \lambda_i(t; \boldsymbol{Z}) P\left(C^* = j | T = t, C = i, \boldsymbol{Z}\right)
\end{aligned}$$

For example, in a two-cause setting, where C, $C^* \in \{1, 2\}$, the cause-specific hazard for observed cause-1 given $\boldsymbol{Z}$ is:

$$\lambda_1^*(t; \boldsymbol{Z}) = \lambda_1(t; \boldsymbol{Z}) P\left(C^* = 1 | T = t, C = 1, \boldsymbol{Z}\right) + \lambda_2(t) P\left(C^* = 1 | T = t, C = 2, \boldsymbol{Z}\right).$$

Colloquially, this means that the observed cause-specific hazard under misclassification is a linear combination of true cause-specific hazards weighted by the misclassfication probabilities. These misclassification probabilities can be estimated using the internal-validation sample as defined in Section 1.3.1.

Subject to outcome misclassification, in the two-cause system, the likelihood is:

$$
\begin{aligned}
L &= \prod_{i=1}^{n} f(t_i, c_i^*; \boldsymbol{Z}_i) \\
&= \prod_{i=1}^{n} [f_j^*(t_i; \boldsymbol{Z}_i)]^{\delta_i^*} [S(t_i; \boldsymbol{Z}_i)]^{1-\delta_i^*} \\
&= \prod_{i=1}^{n} [\lambda_j^*(t_i; \boldsymbol{Z}_i) S^*(t_i; \boldsymbol{Z}_i)]^{\delta_i^*} [S^*(t_i; \boldsymbol{Z}_i)]^{1-\delta_i^*} \\
&= \prod_{i=1}^{n} [\lambda_j^*(t_i; \boldsymbol{Z}_i)]^{\delta_i^*} S^*(t_i; \boldsymbol{Z}_i) \\
&= \prod_{i=1}^{n} [\lambda_j^*(t_i; \boldsymbol{Z}_i)]^{\delta_i^*} \exp\left[-\sum_{j=1}^{2} \int_0^{t_i} \lambda_j^*(u; \boldsymbol{Z}_i) du \right] \\
&= \prod_{i=1}^{n} \left[\sum_{k=1}^{2} \lambda_k(t_i; \boldsymbol{Z}_i) P\left(C_i^* = j | T_i = t_i, C_i = k, \boldsymbol{Z}_i\right) \right]^{\delta_i^*} \times \exp\left[-\sum_{j=1}^{2} \int_0^{t_i} \lambda_j(u; \boldsymbol{Z}_i) du \right] \\
&= \prod_{i=1}^{n} \left[\sum_{k=1}^{2} \lambda_k(t_i; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) P\left(C_i^* = j | T_i = t_i, C_i = k, \boldsymbol{Z}_i\right) \right]^{\delta_{ij}^*} \\
&\times \prod_{j=1}^{2} \exp\left[-\int_0^{t_i} \lambda_j(u; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) du \right] \\
&= \prod_{j=1}^{2} \prod_{i=1}^{n} \left[\sum_{k=1}^{2} \lambda_k(t_i; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) P\left(C_i^* = j | T_i = t_i, C_i = k, \boldsymbol{Z}_i\right) \right]^{\delta_{ij}^*} \times \exp\left[-\int_0^{t_i} \lambda_j(u; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) du \right] \\
&= \prod_{j=1}^{2} \prod_{i=1}^{n} \left[\sum_{k=1}^{2} \lambda_k(t_i; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) \pi_{jk}^*(\mathbf{Z}_i, \boldsymbol{\beta}_k) \right]^{\delta_{ij}^*} \times \exp\left[-\int_0^{t_i} \lambda_j(u; \boldsymbol{\theta}_k, \boldsymbol{Z}_i) du \right]
\end{aligned}
\tag{1.11}
$$

For a fixed $\boldsymbol{\beta}$ we can compute estimates of $\boldsymbol{\theta}$ through maximum likelihood estimation. In a real data analysis, $\boldsymbol{\beta}$ can be estimated using an internal-validation sample such as one described in Section 1.3.1.

At this point, the question that reader should be asking is that of how we should be using the internal-validation sample to compute the misclassification probabilities. In particular, should one ignore the non-validated portion of the sample when estimating misclassification probabilities? I will not answer this questions here, rather I will devote Chapter 2 of this dissertation to exploring the question of estimating outcome misclassification probabilities when one has a validation sample. I devote a whole chapter because the problem at hand is not only a statistical problem, but also a problem of resource constraints. Performing internal-validation/double-sample can be expensive endeavor, as such, data generated through validation sampling should be used as efficiently as possible.

1.3.2 External-validation Based Solution

In last section, I presented internal-validation/double-sampling as a possible solution of dealing with the detrimental effects of outcome misclassification in competing risks analyses. However, I did not touch upon the constraints that may rule out the use internal-validation or double-sampling. The use of internal-validation/double-sampling can be hindered by financial constraints. This is especially true for the resource-poor treatment programs that contribute data to IeDEA East Africa. Validating patient outcomes is expensive as it may require treatment programs to hire, train and pay additional community-health workers, and buy or hire transportation to enable health workers to travel to the homes of patients. Consequently, not all treatment programs can validate patients' vital-status data.

To deal with the outcome misclassification challenge, treatment programs that do not have validation sampling may have to rely on misclassification information from treatment programs with outcome validation. This reliance on misclassification information from external settings is a form of *external validation*. Misclassification information is borrowed from external settings assuming the *transportability* of misclassification models across different settings (Justice, Covinsky, and Berlin 1999; Lyles et al. 2011; R. J. Carroll et al. 2006; Wu et al. 2019). Under transportability, we assume that the misclassification-model coefficients estimated in an external study, with validation sampling, are the same as the coefficients in the current study (that has no validation sampling) (Lyles et al. 2011). For treatment programs within IeDEA East Africa, the scheme for borrowing misclassification probabilities across treatment programs is presented in Figure 1.4.

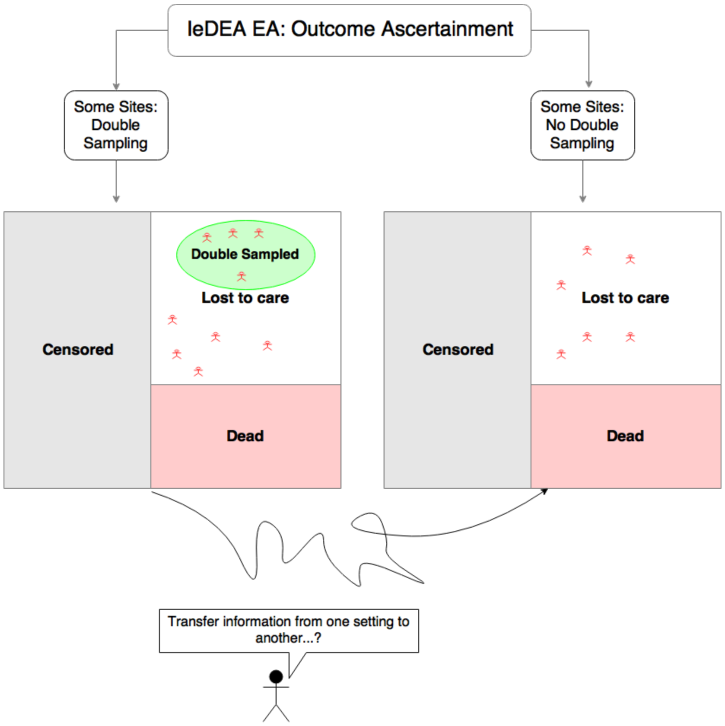
IeDEA EA: Outcome Ascertainment
Some Sites: Double Sampling
Some Sites: No Double Sampling
Censored
Double Sampled
Lost to care
Dead
Censored
Lost to care
Dead
Transfer information from one setting to another...?

Figure 1.4: External validation scheme for treatment programs within IeDEA East Africa consortium.

Notice that, although, we can use the validated sample from the external study to model both predictive values and misclassification probabilities, we only transport misclassification probabilities to the current study. We do not transport predictive value models to external settings/studies as this requires us to assume that event prevalence is the same across different settings: Such an assumption is stringent; thus, unlikely to be satisfied in reality. On the other hand, the transfer of misclassification probabilities from one setting to another does not require identical prevalences across the different settings.

The last thing that the reader should note is that when one relies on external studies/settings for outcome misclassification information, the problem of modeling cause-specific hazards can only be framed as a misclassification problem. In contrast, when the information on outcome misclassification is generated internal to the study, via say double sampling, the problem of modeling cause-specific hazards can be formulated as either a missing-data problem or a misclassification problem.

1.4 Concluding Remarks

Thus far, I have introduced the reader to competing risks survival analysis, and detriments of misclassifying competing events to correct estimation and inference. I also presented how internal-validation sampling can be used as a remedy when faced with challenge of misclassification in the competing events. Internal-validation sampling is a viable solution as it is a cost-effective way to generate quantities that are required to adjust competing risks analyses for misclassification. These quantities include: predictive values and misclassification probabilities. I showed that, when modeling cause-specific hazards in the presence of outcome misclassification, augmenting predictive values or misclassification probabilities into likelihood resulted in likelihood formulations that contained cause-specific hazard parameters of interest. By deriving these formulations, at minimum, I highlighted likelihood-based structures for solving the problem of modeling cause-specific hazards in the presence of outcome misclassification. With IeDEA-EA HIV treatment programs in mind, I will devote this dissertation to exploring solutions to misclassification based on the theoretical-building blocks that were highlighted in the introduction.

Moreover, in the introduction, I noted part of the mandate for IeDEA East Africa is to analyze and share findings from data collected from HIV programs serving East African countries. Such analysis and data sharing assists in evidence-based planning and implementation of treatment programs, thereby encouraging the efficient use of scarce resources. I will, therefore, attempt to frame the statistical solutions in a manner that supports IeDEA EA's public health mandate. The solutions will be presented sequentially in this dissertation's chapters. In Chapter 2, I will present a method for estimating misclassification probabilities in the presence of an internal-validation sample. In Chapter 3, I will develop a parametric method for modeling cause-specific hazards while adjusting for misclassification probabilities that will be estimated outside the study-sample of interest. Finally, in Chapter 4, I will perform a comprehensive data analysis to examine the application of statistical methods developed in Chapter 2 and Chapter 3. Data for the analysis will come from, and will be used with permission from IeDEA East Africa. I am hopeful that the knowledge generated in this work will support the goal of understanding the factors that influence death and disengagement-from-care in IeDEA East Africa, which in turn may inform actionable changes to treatment/care programs.

CHAPTER 2

A Pseudo-likelihood Method for Estimating Misclassification Probabilities When Outcome Data Are Partially Observed

Outcome misclassification occurs frequently in binary-outcome clinical studies and can result in biased estimation of quantities such as the incidence and prevalence. A number of remedies have been proposed to address the potential misclassification of the outcomes in such data. The majority of these remedies lies in the estimation of misclassification probabilities, which are in turn used to adjust analyses for outcome misclassification. A number of authors advocate using a gold-standard procedure on a sample internal to the study to learn about the extent of the misclassification. With this type of internal validation, the problem of quantifying the misclassification also becomes a missing data problem as, by design, the true outcomes are only ascertained on a subset of the entire study sample. Although, the process of estimating misclassification probabilities appears simple conceptually, the estimation methods proposed so far have several methodological and practical shortcomings. Most methods rely on missing outcome data to be missing completely at random (MCAR), a rather stringent assumption which is unlikely to hold in practice. Some of the existing methods also tend to be computationally-intensive. To address these issues, in this chapter, I propose a computationally-efficient, easy-to-implement, pseudo-likelihood estimator of the misclassification probabilities under a missing at random (MAR) assumption, in studies with an available internal validation sample. The corresponding estimates can be directly utilized by methods for misclassification adjustment. I describe the consistency and asymptotic distributional properties of the resulting estimate, and derive a closed-form estimator of its variance. The estimator is also extended to settings with clustered data. The finite-sample performance of this estimator is evaluated via simulations. Using real-world data, I illustrate how the proposed method can be used to estimate misclassification probabilities. I also show how the estimated misclassification probabilities can be used in an external study to adjust for possible misclassification bias in the framework of competing risks.

2.1 Introduction

Outcome misclassification in binary data leads to bias, and thereby poses a significant threat to the validity of epidemiological and clinical studies (Bross 1954; Barron 1977; Magder and Hughes 1997; Neuhaus 1999; Lyles et al. 2011; Edwards et al. 2013). The effect of this bias can be ameliorated by adjusting estimators for possible misclassification(Lyles et al. 2011; Tang et al. 2015; Lyles and Lin 2010). One way to make this adjustment, is to have *a priori* knowledge about the misclassification probabilities. However, the extent of misclassification is rarely known beforehand so it must be estimated.

A frequently used approach to obtain information about the extent of misclassification is *internal validation or double-sampling* (Greenland 1988). In this approach, the true outcomes for a small subset of study participants are ascertained using a gold-standard outcome-ascertainment procedure (Tenenbein 1970). Based on this internal validated sample, misclassification probabilities can be estimated by comparing the observed (and potentially misclassified) outcomes with the outcomes obtained through the gold-standard procedure. Then the resulting misclassification probabilities can be used to adjust estimators in the current study or in other studies where, for some reason, internal validation sampling is not possible. This latter use of the misclassification probabilities is known as external validation, because the validation sample is obtained outside the study of interest (Spiegelman, Carroll, and Kipnis 2001).

The motivation for the exploration into the estimation of misclassification probabilities is a large study in sub-Saharan Africa consisting of people living with HIV/AIDS (PLWH) that receive care at various health facilities participating in the East-African International Epidemiology Databases for the Evaluation of AIDS (IeDEA) consortium. One challenge that arises in the monitoring and evaluation of care-program effectiveness is the underreporting of death (Egger et al. 2011; Brinkhof et al. 2010; Bakoyannis and Yiannoutsos 2015). Unreported deaths are typically classified as disengagements from care by the program staff. The underreporting of death leads to an underestimation of mortality and an overestimation of rates of disengagement from care. This problem can be remedied by internal validation, wherein a more accurate (but also more expensive) method is used to ascertain the true patient outcomes in a portion of subjects who have been initially declared as having disengaged from care because they failed to attend to their clinic visits (internal validation sample) (Tenenbein 1970). The reason for selecting a subset of the study population for exhaustive outcome validation is that internal validation, although desirable, cannot be performed on a large scale because of resource and other feasibility constraints. As a result, the outcomes in the remaining patients are missing by design (Wacholder 1996; Zhao, Lawless, and McLeish 2009). In spite of this inherent missingness, internal validation is an efficient way to identify misclassification probabilities, which in turn can be used to adjust statistical estimators that target parameters such as prevalence, cause-specific hazards, cumulative incidence and so on. We can also employ external validation, where information on the misclassification (of death as disengagement from care) can be obtained from a validation sample outside the main study (Spiegelman, Carroll, and Kipnis 2001).

In the East-Africa IeDEA cohort, the internal validation scheme involves intensive tracing in the community of a subset of patients considered disengaged from care, and active ascertainment of their vital status (E. H. Geng et al. 2008; An et al. 2009; Yiannoutsos et al. 2008). This patient tracing procedure is the gold-standard outcome ascertainment approach alluded to earlier. It results in much more accurate vital-status data than initially captured by routine review of patients' medical records. However, as described earlier, the data resulting from tracing are missing by design for patients where no tracing was performed. In addition, these data are also affected by non-response as some patients cannot be successfully traced.

When an internal validation sample is available, most authors use only the internal validation sample to estimate the extent of outcome misclassification. By so doing, they implicitly assume that outcome data on the non-validation sample are missing completely at random (MCAR) (Magder and Hughes 1997; Chen 2000; Pepe 1992). That is, the probability of missingness is independent of both the observed characteristics of the patients and the unobserved outcomes (Rubin 1976). In reality, MCAR is rarely justifiable. Other authors attempt to resolve this problem by augmenting the validated and the non-validated samples allowing for the use of the entire study sample in the estimation procedure. However, such data augmentation methods like the expectation-maximization (EM) algorithm, and multiple imputation can be difficult to use. For example, in order to use the EM algorithm, one needs to correctly set up the expectation and maximization steps and correctly derive the variance estimator. On the other hand, multiple imputation can be complicated if the imputation and the analysis models are not congenial (Meng 1994), that is if the imputation model does not contain all the variables in the analysis model including the response variable of interest, and auxiliary covariates that may be related to the variables being imputed. The need for compatibility between the analysis and imputation models is a common pitfall when it comes to using multiple imputation (Tilling et al. 2016). The consequence of this is that the Rubin's variance estimator is biased (Robins and Wang 2000), and this ultimately leads to invalid inference.

To address many of the methodological and practical shortcomings of existing methods, I propose a pseudo-likelihood approach for estimating outcome misclassification (unidirectional or bi-directional) probabilities when some of the binary outcome data are missing both by design and by non-response. This method relaxes the MCAR assumption, which is untenable in the study context because not everyone who is sampled for internal validation is available to provide data. Instead, data are assumed to be missing at random (MAR), allowing missingness to be related to observed data and observed subject characteristics (Rubin 1976). Furthermore, unlike Rubin's multiple imputation, I allow for auxiliary covariates that may be related to the probability of missingness and can make the MAR assumption more plausible in practice (Lu and Tsiatis 2001). The proposed method is easy to implement and relies on existing software. Moreover, the method is computationally efficient and can thus be used with the large data sets frequently encountered in large epidemiological studies. An added benefit of the proposed method is that it can be can be easily extended to clustered data settings, such as multi-site treatment programs.

This chpater proceeds as follows: In Section 2.2, I present some of the data assumptions and notation. In Section 2.3, I present the likelihood given the data, describe the pseudo-likelihood function and derive the large-sample properties of the resulting pseudo-likelihood estimator. I also present an extension to a clustered data setting in Section 2.3. In Section 2.4, I evaluate the finite-sample properties of estimator using a simulation study. In Section 2.5, I present a data application to illustrate the estimation of misclassification probabilities. In Section 2.6, I illustrate the use of misclassification probabilities estimated in Section 2.5

to make adjustments for potential misclassification in external studies with no outcome validation. I conclude with a brief discussion of the findings in Section 2.7.

2.2 Notation and Assumptions

This paper focuses on binary data that occur within a competing risks setting, as a result the *event types* shall be referred to as *causes of failure.* Assume that individuals succumb to two competing causes of failure (events) , say, cause 1 and cause 2. Also assume that the method of ascertaining the cause of failure is subject to error, so that the observed and true causes of failure are not always the same. We can think of the observed causes as "surrogates" for the true causes of failure. Let $C^* \in \{1, 2\}$ represent the observed, and potentially misclassified cause of failure, and $C \in \{1, 2\}$ represent the true cause of failure. Henceforth, "*observed causes of failure*" are those ascertained through a standard method that is subject to error, and "*true causes of failure*" are those that ascertained using a gold-standard method that is more accurate than the standard method.

In general, the conditional misclassification probabilities are represented as follows:

1. $\mathrm{P}\left[C^* = 1 | C = 2, \boldsymbol{X}, \boldsymbol{\beta}_2\right] = \pi^*_{12}(\boldsymbol{\beta}_2; \boldsymbol{X})$
2. $\mathrm{P}\left[C^* = 2 | C = 1, \boldsymbol{X}, \boldsymbol{\beta}_1\right] = \pi^*_{21}(\boldsymbol{\beta}_1; \boldsymbol{X})$,

where, $\boldsymbol{X}$ represents a matrix of subject characteristics, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ represents the association between misclassification probabilities and subject characteristics. In order to simplify our exposition, assume that both misclassification probabilities depend on the same set of covariates $\boldsymbol{X}$. It is also worth noting that the misclassification probabilities defined above can be seen as the complements of subject-level sensitivities of a diagnostic/classification method.

With real-world applications in mind, I will model misclassification probabilities using parametric logistic regression. In epidemiology, logistic regression is popular because the resulting relationship between the log-odds and covariates has an intuitive interpretation. The logit models for the true misclassification probabilities are defined below:

$$\log\left[\frac{\pi^*_{12}}{1-\pi^*_{12}}\right] = \boldsymbol{X}^T\boldsymbol{\beta}_2 \tag{2.1}$$

$$\log\left[\frac{\pi^*_{21}}{1-\pi^*_{21}}\right] = \boldsymbol{X}^T\boldsymbol{\beta}_1 \tag{2.2}$$

where $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^q$, and $\boldsymbol{X}_{n\times q}$.

It is worth reiterating that the binary-outcome misclassification problem of interest occurs within competing risks, as such, most of the notation and data set-up will mimic that of competing risks literature. Lets consider a study where each subject is followed until he/she fails from either cause 1 or cause 2, or is censored. Observing cause 1 precludes us from observing cause 2, and vice-versa. Outcomes are re-ascertained on a sub-sample from those observed to fail from either cause 1 or cause 2. We refer to this outcome re-ascertainment as internal validation or double-sampling. There is no need for outcome validation among those censored, as censoring is assumed to be correctly ascertained. Let R_i be the indicator that the true outcome is known, with $R_i = 1$ indicating that the subject i was successfully double-sampled or censored. The true outcome, C_i, is only observed if subject i is successfully double sampled, or is censored $(C_i = 0)$. For each subject, $i = 1, 2, ..., n$, we observe $\{C^*_i, X_i = (T_i, X^*_i), R_i, (C_i \,\text{if}\, R_i = 1)\}$, where,

1. R_i: is the indicator function that the outcome for subject i is known;

2. $C_i \in \{0, 1, 2\}$: is the true cause of failure, observed only if $R_i = 1$ or if subject i is censored;
3. $C_i^* \in \{0, 1, 2\}$: is the observed cause of failure, $C_i^* = 0$ if subject i is censored;
4. $\boldsymbol{X}_i^*$: are observed covariates for subject i, excluding time contribution to study;
5. V_i: is the censoring time;
6. U_i: is the time to cause 1 or cause 2.
7. $T_i = \min(U_i, V_i)$: is the time contributed to study by subject i;
8. $\boldsymbol{X}_i$: are the observed covariates for subject i, including time contribution to study;

We assume that the censoring time is independent of failure time and the cause of failure, that is, $(T, C) \perp V$. We also assume that subject characteristics $\boldsymbol{X}$ are measured without error.

Ideally, double-sampling should be able to validate the outcomes on the entire subsample selected for validation. In reality, this is unlikely to be true because not everyone who is double-sampled is available, so the true outcome cannot be ascertained. This lack of response may not be completely at random. I relax this assumption and assume that missing data due to non-response are missing at random (MAR) so that, among non-censored subjects, the probability that the true outcome is missing (unknown) may depend on observed subject characteristics measured prior to censoring and not on the unobserved true outcome. That is, I assume that $\mathrm{P}[R_i = 0 | C_i, C_i^* > 0, \boldsymbol{X}_i] = \mathrm{P}[R_i = 0 | C_i^* > 0, \boldsymbol{X}_i]$.

2.3 Likelihood

Under the assumptions presented above, the log-likelihood of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)^T \in \mathbb{R}^{2q}$ based on the observed data is

$$
\begin{aligned}
l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_{1i} \left\{ (1 - \delta_{1i}^{*})\, \boldsymbol{X}_i^T \boldsymbol{\beta}_1 - \log\left(1 + \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_1\right)\right) \right\} \\
+ \sum_{i=1}^{n} \delta_{2i} \left\{ \delta_{1i}^{*} \boldsymbol{X}_i^T \boldsymbol{\beta}_2 - \log\left(1 + \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_2\right)\right) \right\} \qquad (2.3)
\end{aligned}
$$

where $\delta_{1i} = I[C_i = 1]$ and $\delta_{2i} = I[C_i = 2]$, $\delta_{1i}^{*} = I[C_i^{*} = 1]$ and $\delta_{2i}^{*} = I[C_i^{*} = 2]$ are the true and observed event indicators. The derivation of the full-likelihood from which log-likelihood (2.3) is obtained can be found in Section 2.8.1.

In the above full log-likelihood (2.3), notice that δ_{1i} and δ_{2i} are only observable among those who were successfully double-sampled or censored, that is, some subset of $\{i = 1, 2, ..., n\}$. As a result, maximum likelihood estimation is not straightforward. We can proceed with maximum likelihood estimation by setting up an EM algorithm (Magder and Hughes 1997; Dempster, Laird, and Rubin 1977). This can be challenging for even for people with formal statistical training as it requires customized programming. In addition, the EM algorithm is computationally-expensive, particularly with the large databases involved in the motivating HIV study. To overcome these shortcomings, I proceed by first formulating the objective function as a pseudo/estimated likelihood.

2.3.1 Setting up the pseudo-likelihood

I begin by recognizing that the binary-outcome indicators, δ_{1i} and δ_{2i}, are linear in the log-likelihood as shown in Equation 2.3. As a result of this linearity I can still perform consistent estimation by replacing the missing true values using their conditional expectations given the observed data. That is, among those missing true outcome values, δ_{ji} is replaced by $\mathrm{E}\left[\delta_{ji} | \delta_{ki}^{*}, \boldsymbol{Z}_i\right] = p_{jk}(\boldsymbol{\gamma}_k; \boldsymbol{Z}_i)$ for $j, k \in \{1, 2\}$. In the context of a real data-analysis, estimation proceeds by replacing δ_{1i} and δ_{2i}, in the full log-likelihood 2.3 by $\tilde{\delta}_{1i}$ and $\tilde{\delta}_{2i}$ respectively, where:

$$\tilde{\delta}_{1i} = R_i \times \delta_{1i} + (1 - R_i) \times \left[p_{11}(\hat{\boldsymbol{\gamma}}_1; \boldsymbol{Z}_i)^{\delta^*_{1i}} p_{12}(\hat{\boldsymbol{\gamma}}_2; \boldsymbol{Z}_i)^{1-\delta^*_{1i}}\right] \tag{2.4}$$

and

$$\tilde{\delta}_{2i} = R_i \times \delta_{2i} + (1 - R_i) \times \left[p_{21}(\hat{\boldsymbol{\gamma}}_1; \boldsymbol{Z}_i)^{\delta^*_{1i}} p_{22}(\hat{\boldsymbol{\gamma}}_2; \boldsymbol{Z}_i)^{1-\delta^*_{1i}}\right] \tag{2.5}$$

Here $\hat{\boldsymbol{\gamma}}_1, \hat{\boldsymbol{\gamma}}_2 \in \mathbb{R}^d$; $\boldsymbol{Z}_i$ is a $1 \times d$ matrix containing the characteristics for subject i; and $p_{jk}(\hat{\boldsymbol{\gamma}}_k; \boldsymbol{Z}_i) = P(C_i = j | C^*_i = k, \boldsymbol{Z}_i, \hat{\boldsymbol{\gamma}}_k)$ for $j, k \in \{1, 2\}$ is the estimated conditional probability of the true cause $C = j$ given the observed cause $C^* = k$, for $k \in \{1, 2\}$, $\sum_{j=1}^{2} p_{jk}(\hat{\boldsymbol{\gamma}}_k; \boldsymbol{Z}_i) = 1$. Henceforth, for all j and k, I shall refer to $p_{jk}(\boldsymbol{\gamma}_k; \boldsymbol{Z}_i)$ as the predictive values of the standard diagnostic/classification procedure. The phrase "predictive value" is used in a similar manner as in traditional diagnostic testing literature, where, for example P[Diseased|Positive test result] is called the positive predictive value of a diagnostic test. If subject i is not censored, and δ_{1i} and δ_{2i} are not observed, the "true" cause indicators in the likelihood are replaced by the estimated predictive values. Missing outcome data are assumed to be missing at random(MAR). That is, among the non-censored, the probability that the true cause is missing conditional on the observed cause is independent of the true cause of failure. That is,

$$\mathrm{P}[R_i = 0 | C_i, C^*_i > 0, \boldsymbol{Z}_i] = \mathrm{P}[R_i = 0 | C^*_i > 0, \boldsymbol{Z}_i]$$

It is also worth noting that based on notation defined in Section 2.2, $I[R = 0]$, is also a missing value indicator. The covariate matrix $\boldsymbol{Z}$ may also include auxiliary covariates that make the MAR assumption plausible. From the MAR assumption defined above, without losing generality, it immediately follows that:

$$\begin{aligned}\mathrm{P}[C_i = 1 | R_i = 0, C_i^* > 0, \boldsymbol{Z}_i] &= \mathrm{P}[C_i = 1 | R_i = 1, C_i^* > 0, \boldsymbol{Z}_i] \\ &= \mathrm{P}[C_i = 1 | C_i^* > 0, \boldsymbol{Z}_i] \qquad (2.6)\end{aligned}$$

In other words, under the MAR assumption, among the non-censored, the predictive value model is the same among those who were double-sampled and those who were not double-sampled (Bakoyannis, Siannis, and Touloumi 2010). From a data analysis perspective, this means predictive values, estimated using data from those whose outcomes were validated, can be used to inform the predictive-value estimates among those whose outcomes were not validated (that is, provided the validated and unvalidated subjects are drawn from the same population).

I also estimate predictive values, $p_{jk}(\boldsymbol{\gamma}; \boldsymbol{Z})$ for $j, k \in \{1, 2\}$, parametrically using logistic regression as follows:

$$p_{12}[\boldsymbol{\gamma}_2; \boldsymbol{Z}] = \frac{\exp(\boldsymbol{Z}^T \boldsymbol{\gamma}_2)}{1 + \exp(\boldsymbol{Z}^T \boldsymbol{\gamma}_2)} \qquad (2.7)$$

$$p_{21}[\boldsymbol{\gamma}_1; \boldsymbol{Z}] = \frac{\exp(\boldsymbol{Z}^T \boldsymbol{\gamma}_1)}{1 + \exp(\boldsymbol{Z}^T \boldsymbol{\gamma}_1)} \qquad (2.8)$$

where $\boldsymbol{Z}$ is a matrix of subject characteristics. Note that, the subject characteristics in $\boldsymbol{Z}$ need not be the same as those in $\boldsymbol{X}$, the set covariates used to build the misclassification models 2.1 and 2.2.

When δ_{1i} and δ_{2i} are replaced with $\tilde{\delta}_{1i}$ and $\tilde{\delta}_{2i}$ respectively, the resulting pseudo-log-likelihood(estimated log-likelihood) is:

$$l(\boldsymbol{\beta};\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \tilde{\delta}_{1i} \left\{ \left(1-\delta_{1i}^{*}\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_1 - \log\left(1+\exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_1\right)\right) \right\} \\ + \sum_{i=1}^{n} \tilde{\delta}_{2i} \left\{ \delta_{1i}^{*} \boldsymbol{X}_i^T \boldsymbol{\beta}_2 - \log\left(1+\exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_2\right)\right) \right\} \quad (2.9)$$

where the overall parameter is $(\boldsymbol{\beta},\boldsymbol{\gamma}) \in \mathbb{R}^{2(q+d)}$, with $\boldsymbol{\beta}$ representing the parameter of interest, and $\boldsymbol{\gamma}$ the nuisance parameter. The parameter $\boldsymbol{\gamma}$ is estimated by fitting logistic regression models using the internal validation data as stated above. Assuming the logistic regression models are correctly specified, $\hat{\boldsymbol{\gamma}}$ will converge in probability to $\boldsymbol{\gamma}$. When $\hat{\boldsymbol{\gamma}}$ is plugged into the log-likelihood, the problem reduces to that of optimizing $l(\boldsymbol{\beta},\hat{\boldsymbol{\gamma}})$, a pseudo-log-likelihood, and the resulting estimates are called pseudo-likelihood estimates.

The maximum pseudo-likelihood estimate (MPLE), is such that, the average score function is equal to zero, that is,

$$\boldsymbol{\Psi}_n^{(1)}(\boldsymbol{\beta}_1,\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \tilde{\delta}_{1i}(\hat{\boldsymbol{\gamma}}) \left[(1-\delta_{1i}^{*}) - \frac{\exp\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)}{1+\exp\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)} \right] = 0 \quad (2.10)$$

$$\boldsymbol{\Psi}_n^{(2)}(\boldsymbol{\beta}_2,\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \tilde{\delta}_{2i}(\hat{\boldsymbol{\gamma}}) \left[\delta_{1i}^{*} - \frac{\exp\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_2\right)}{1+\exp\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_2\right)} \right] = 0 \quad (2.11)$$

Generally, the average score function is of the form $\boldsymbol{\Psi}_n(\boldsymbol{\beta},\hat{\boldsymbol{\gamma}})$. Henceforth, I shall focus on $\hat{\boldsymbol{\beta}}_n$, the general estimator of $\boldsymbol{\beta}$.

2.3.2 Asymptotics

The asymptotic properties of the proposed pseudo-likelihood estimator were established under the same regularity conditions as those presented by Gong and Samaniego (1981) (Gong and Samaniego 1981), Parke (1986) (Parke 1986), and Bakoyannis et al. (2018) (Bakoyannis, Zhang, and Yiannoutsos 2018). Particularly, the regularity conditions are the same as those in standard maximum likelihood theory, with the exception being the following two conditions:

1. $\hat{\boldsymbol{\gamma}}_n \xrightarrow{p} \boldsymbol{\gamma}$ as $n \to \infty$;
2. The ratio of the size (n) of the main sample to the size (n_v) of the validation sample is fixed. That is, $\lim_{n\to\infty} \frac{n}{n_v} = s$.

It was shown that $\hat{\boldsymbol{\beta}}_n$ is consistent estimator of $\boldsymbol{\beta}$. The detailed proof for consistency can be found in the Section 2.8.2. Additionally, it can be shown that:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$$

where $\boldsymbol{\Omega} = \boldsymbol{I}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + s.\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\mathbf{W}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$ with

$$\begin{aligned}\mathbf{W}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{X}, \boldsymbol{Z}) = {} & \mathrm{E}\left[\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{Z})\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|X)^T\right] \\ & + \mathrm{E}\left[\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|\boldsymbol{X})\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{Z})^T\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^T\right] \\ & + \mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^T\end{aligned}$$

where $\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \left[\frac{d}{d\gamma}\Psi_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\right]$, $s = \lim_{n\to\infty} \frac{n}{n_v}$, with n_v being the size of the validation sample. $\boldsymbol{\Omega}$ can be estimated by replacing the parameter $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ with their consistent estimators so that

$$\hat{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}(\boldsymbol{X}_i|\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_{n_v})\tilde{\psi}(\boldsymbol{X}_i|\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_{n_v})^T$$

where $\tilde{\psi}(\boldsymbol{X}_i|\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = -\boldsymbol{I}^{-1}\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\left[\dot{l}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0|\boldsymbol{X}_i) + \sqrt{s}.\mathbf{R}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{X}_i)\right]$. A detailed proof for asymptotic normality can be found in Section 2.8.3.

2.3.3 Clustered Data

The estimation method that has been described thus far, only applies to cross-sectional data where individuals are considered to be independent. In this section, I show that the proposed method can be extended easily to clustered data settings, where subjects nested within a cluster are considered to be correlated. The observed data will remain as described in Section 2.2, with the only difference being that subjects will now have a cluster identifier. Assuming that there are m clusters of varying sample sizes, let i be the cluster index, such that $i = 1, 2, ..., m$, and let j be the subject index within cluster, such that $j = 1, 2, ..., n_i$. Also assume finite sample size within clusters, that is, $n_i < \infty$, and that the clusters are independent. Overall, there are $n = \sum_{i=1}^{m} n_i$ subjects considered. Under a working independence assumption among clusters, asymptotic arguments will be same as described in Section 2.3, with clusters now being the primary sample units. Generally, at the maximum pseudo-likelihood estimate, $\sum_{i=1}^{m}\sum_{j=1}^{n_i} U_{ij}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m) = \sum_{i=1}^{m} U_{i.}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m) = 0$. By the central limit theorem,

$$\sqrt{m}\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0\right) \rightarrow N(\mathbf{0}, \boldsymbol{\Omega})$$

where, $\boldsymbol{\Omega} = \mathrm{E}[\tilde{\psi}_{i.}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\tilde{\psi}_{i.}^{T}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)]$, with

$$\tilde{\psi}_{i.}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = -\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)U_{i.}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) - \sqrt{s}\mathbf{R}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)I^{-1}(\boldsymbol{\gamma}_0)U_{i.}(\boldsymbol{\gamma}_0)$$

Also, $\mathbf{R}(\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \frac{1}{m}\sum_{i=1}^{m}\frac{d}{d\gamma}U_{i.}(\boldsymbol{\beta}_0,\boldsymbol{\gamma})|_{\gamma=\gamma_0}$ and $s = \frac{n_i}{n_{i(v)}}$ as cluster size, n_i,increases to ∞ for all $i = 1, 2, ..., m$. $n_{i(v)}$ is number of the double-sampled individuals within cluster i. Empirically, $\boldsymbol{\Omega}$ is estimated by replacing parameters with their consistent estimates, that is,

$$\hat{\boldsymbol{\Omega}} = \mathrm{E}[\tilde{\psi}_{i.}(\hat{\boldsymbol{\beta}}_m,\hat{\boldsymbol{\gamma}}_m)\tilde{\psi}_{i.}^{T}(\hat{\boldsymbol{\beta}}_m,\hat{\boldsymbol{\gamma}}_m)] = \frac{1}{m}\sum_{i=1}^{m}\tilde{\psi}_{i.}(\hat{\boldsymbol{\beta}}_m,\hat{\boldsymbol{\gamma}}_m)\tilde{\psi}_{i.}^{T}(\hat{\boldsymbol{\beta}}_m,\hat{\boldsymbol{\gamma}}_m)$$

2.3.4 Implementing the pseudo-likelihood estimation in `R`

It is fairly simple to set up the estimating equation represented by Equation 2.10 (or 2.11) in `R`. This entails fitting a logistic regression model using the `glm` function, where the binary-outcome is δ^*_{1i}, and the `weights` option of the `glm` function is set to $\tilde{\delta}_{1i} = R_i\delta_{1i} + (1 - R_i)[1 - p_{21}(\hat{\gamma}_1; \boldsymbol{Z}_i)]^{\delta^*_{1i}} p_{12}(\hat{\gamma}_2; \boldsymbol{Z}_i)^{1-\delta^*_{1i}}$, for subject indices $i = 1, ..n$. A subject whose outcome was validated is weighted based on his/her validated outcome(0 versus 1), otherwise they will be weighted based on an estimated predictive value between zero and one. The `glm` function in `R`, however, does not return correct standard error estimates: When it computes standard errors, it ignores the additional variability due the estimation of predictive values, $p_{jk}(\boldsymbol{\gamma}; \boldsymbol{Z})$. One needs to manually code the closed-form variance estimator in `R`. Details about the composition of the closed-form estimator are presented in sub-section 2.3.2. Alternatively, one could appeal to parametric bootstrapping in order to propagate, into the estimation, the variability due the estimation of predictive values.

2.4 Simulation Study

The finite sample properties of the pseudo-likelihood estimator were explored using a simulation study. The details of the simulation study are described below.

2.4.1 Simulating the *true cause of failure*

We simulate the competing risks data using the method developed by Beyersmann et al. (2009) (Beyersmann et al. 2009). Assume the failure time T is distributed according to the Weibull distribution with parameters $\alpha > 0$ and $\lambda > 0$, that is, $T \sim \mathrm{W}(\alpha, \lambda)$. Under a competing risks scenario with two causes of failure, $C \in \{1, 2\}$, the cause-specific hazards are: $h_i(t) = \alpha_i \lambda_i t^{\alpha_i - 1}$, for $i = 1, 2$. The overall survival function is

$$S(t) = \exp\left(-\sum_{i=1}^{2} \int_0^t \alpha_i \lambda_i u^{\alpha_i - 1} du\right) = \exp\left(-\sum_{i=1}^{2} \lambda_i t^{\alpha_i}\right)$$

Given that an event occurs at time $T = t$, the probability that the cause of failure is cause $i \in \{1, 2\}$ is

$$\mathrm{P}[C = i | T = t] = \frac{\alpha_i \lambda_i t^{\alpha_i - 1}}{\sum_{i=1}^{2} \alpha_i \lambda_i t^{\alpha_i - 1}}$$

Assuming proportional hazards, the survival distribution is

$$S_T(t|Z) = \exp\left[-\sum_{i=1}^{2} \lambda_i t^{\alpha_i} \exp(Z^T \kappa_i)\right]$$

and

$$\mathrm{P}[C = i | T = t, \boldsymbol{Z}, \boldsymbol{\kappa}] = \frac{h_i(t) \exp(\boldsymbol{Z}^T \boldsymbol{\kappa}_i)}{\sum_{i=1}^{2} h_i(t) \exp(\boldsymbol{Z}^T \boldsymbol{\kappa}_i)}$$

where $\boldsymbol{Z} = (z_1, z_2)$ is the matrix of covariates.

If a subject does not experience either cause 1 or cause 2, the subject will be right censored. Assume that censoring time, $V \sim \mathrm{Exp}(\eta)$. For subject j, the survival data are $\{\min(T_j, V_j), C_j = i\}$, for $j = 1, 2, ..., n$, $i = 0, 1, 2$; $C_j = 0$ if subject j is censored. Setting $\alpha_1 = \alpha_2 = \alpha$, the formula for failure time is derived by inversion to be $t = \left[\frac{-\log(1-U)}{\sum_{i=1}^{2} \lambda_i \exp(\boldsymbol{Z}_{\boldsymbol{i}}^T \boldsymbol{\kappa}_i)}\right]^{\frac{1}{\alpha}}$, where $U \sim \mathrm{U}(0, 1)$. If the $\min(T_j = t, V_j = v) = v$, then subject j is considered to have been censored, that is $C_j = 0$, otherwise subject experiences either cause 1 or 2 at time t. Given that subject j experienced failure at time t, the probability that he/she failed due to cause 1 is:

$$\mathrm{P}[C_j = 1 | T_j = t, \boldsymbol{Z}_j, \boldsymbol{\kappa}] = \frac{\alpha \lambda_1 t^{\alpha - 1} \exp(\boldsymbol{Z}_j^T \boldsymbol{\kappa}_1)}{\sum_{i=1}^{2} \alpha \lambda_i t^{\alpha - 1} \exp(\boldsymbol{Z}_j^T \boldsymbol{\kappa}_i)} \tag{2.12}$$

Define D_j as the indicator function that subject j fails from cause 1, otherwise fails from cause 2. Using the probability in (2.12), generate true cause of failure 1 from $D_j \sim \mathit{Bernoulli}\,(\mathrm{P}[C_j = 1 | T_j = t, Z_j])$. For a non-censored subject j,

$$C_j = 1 \times I(D_j = 1) + 2 \times I(D_j = 0)$$

2.4.2 Simulating the *observed cause of failure*

Instead of $C \in \{1, 2\}$, we observe $C^* \in \{1, 2\}$, where C^* and C are the observed and true causes of failure respectively and are not necessarily the same due to misclassification. Assume that those who are censored are never misclassified, that is, $C = 0$ if and only if $C^* = 0$. Additionally, let $\mathrm{P}[C^* = 2|C = 1] = \pi^*_{21}$, and $\mathrm{P}[C^* = 1|C = 2] = \pi^*_{12}$ be the misclassification probabilities as defined in Section 2.2, with true models of log-odds of misclassification defined as

$$\log\left(\frac{\pi^*_{21}}{1-\pi^*_{21}}\right) = \boldsymbol{X\beta}_1 = \frac{\exp(\beta_{01} + \beta_{21}t + \beta_{21}z_1 + \beta_{31}z_2)}{1 + \exp(\beta_{01} + \beta_{21}t + \beta_{21}z_1 + \beta_{31}z_2)}$$

and

$$\log\left(\frac{\pi^*_{12}}{1-\pi^*_{12}}\right) = \boldsymbol{X\beta}_2 = \frac{\exp(\beta_{02} + \beta_{12}t + \beta_{22}z_1 + \beta_{32}z_2)}{1 + \exp(\beta_{02} + \beta_{12}t + \beta_{22}z_1 + \beta_{32}z_2)}$$

where $\boldsymbol{X}_{n\times 4} = [\mathbf{1}, \boldsymbol{Z}_1, \boldsymbol{Z}_2, \mathbf{t}]$. We generate

$$M_j \sim \mathit{Bernoulli}\left(I(C_j = 1) \times \pi^*_{21j} + I(C_j = 2) \times \pi^*_{12j}\right)$$

, the misclassification indicator for subject j where $M_i = 1$ indicates that the outcome is misclassified, that is, the observed outcome is not the same as the true outcome. The observed cause of failure for subject j, C^*_j, is then defined as follows:

$$C^*_j = \begin{cases} C_j & \text{if } M_j = 0 \\ 1 \times I(C_j = 2) + 2 \times I(C_j = 1) & \text{if } M_j = 1 \end{cases}$$

2.4.3 True outcomes missing at random (MAR)

In addition to exploring a situation where data are missing completely at random (MCAR), the simulation study also explores a situation where data are missing at random (MAR). In this case, MAR will arise from the non-response among some of the double-sampled subjects. In particular, investigation was done for a situation where the probability of being successfully double-sampled is about 80%, and deviations from that probability are explained by an auxiliary variable A. Although they may not be of interest in the study, auxiliary covariates make the MAR assumption plausible (Hardt, Herke, and Leonhart 2012). Here the auxiliary variable, A is associated with both outcome misclassification and the missingness in the true cause of failure. A is defined as follows:

$$A = I[C = C^*] \times Ber(0.3) + I[C \neq C^*] \times Ber(0.45)$$

Among the double-sampled, the probability that the double-sampling is successful (true outcome is not missing) is given by

$$\mathrm{P}[R = 1|A = a] = \frac{\exp\left(\log(4) - a\right)}{1 + \exp\left(\log(4) - a\right)}$$

2.4.4 Conducting the simulation study

Simulation parameters were set as follows:

a) *Misclassification parameters*: $\boldsymbol{\beta}_1 = (-0.4, -0.4, 0.5, -0.5)$, $\boldsymbol{\beta}_2 = (-0.4, -0.4, 0.5, -0.5)$;

b) *Weibull proportional hazards parameters*: $\kappa_1 = (0.5, 1)$, $\kappa_2 = (-0.5, 0.5)$;

c) *Weibull shape parameter*: $\alpha = 2$;

d) *Weibull scale parameters*: $\lambda_1 = 0.75$, $\lambda_2 = 1$;

e) *Exponential censoring parameter* $\eta = 0.6$;

f) *Subject characteristics*: $Z_1 \sim U(0, 1)$, $Z_2 \sim N(0, 1)$.

Simulations were performed using datasets of sample size 5000, and varied the following conditions:

i) *double-sampling proportion* (20% versus 50%), with those who are double-sampled only drawn from the non-censored portion of the sample;

ii) *missing outcome imputation* (no imputation versus imputation). Not imputing is tantamount to performing a complete case analysis wherein only those who are successfully double-sampled are considered. And, imputing entails using our proposed pseudo-likelihood method of estimation.

iii) *missingness mechanism* (MCAR versus MAR). MCAR data occur when the double-sampling among the non-censored is 100% successful (missingness of true outcomes is completely by design). On the other hand, MAR data are simulated as described in subsection 2.4.3.

iv) *predictive value model specification* (correct versus incorrect). When $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, for example $\boldsymbol{\beta}_2 = (-0.3, 0.2, 0.5, 0.5)$, it is no longer correct to use logistic regression to model the predictive values. That is, when $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$, the proposed logistic models 2.7 and 2.8 are no longer suitable because the linearity assumption between the logit function and the covariates is violated. Using logistic regression to model the predictive values will therefore be a form of model misspecification. The proof of this assertion is provided in Appendix subsection 2.8.5. For the MCAR, the covariates that were entered into the predictive value models were the same as those for the proposed misclassification model. For the MAR case, the predictive value model also included the auxiliary covariate, A, in addition to the covariates entered into the proposed misclassification model. An additional thing to note is that the correct predictive-value model specification coincides with a case where the misclassification models for cause 1 and cause 2 are the same. On the other hand, incorrect model specification coincides with a case where the misclassification models for cause 1 and cause 2 are different.

For each of the 16 simulation conditions, 1000 replications were performed and the following quantities were compute: average estimate, $\hat{\beta}_{average} = \frac{1}{1000}\sum_{l=1}^{1000}\hat{\beta}_l$; absolute percent bias of average estimate, $100 \times |\frac{\hat{\beta}_{average}-\beta}{\beta}|$; Monte-Carlo standard deviation (MCSD) ,$\sqrt{\frac{1}{1000-1}\sum_{l=1}^{1000}\left(\hat{\beta}_l - \beta\right)^2}$; asymptotic standard error (ASE); 95% coverage probability(CP); the relative efficiency (RE) of the complete-case estimator versus the pseudo-likelihood estimator. Estimation was repeated using the EM algorithm, and the computational efficiency of the EM algorithm was compared to that of the proposed pseudo-likelihood method. Lastly, the simulation study was repeated under a clustered data setting. Data for clustered setting were simulated using the marginal-model approach within the `SimCorMultRes` `R` package (Touloumis 2016). Cluster size was set at 50, and the intra-cluster correlation was set at 0.5. Under a correct model specification, as described in condition (iv), simulations were performed for 100, 200, 400 and 800 clusters.

2.4.5 Simulation Results

Results under MCAR

The datasets used in simulations with 20% double-sampling under MCAR are summarized in Figure 2.1. In the 1000 simulation datasets, on average, 37.23% of those who truly failed from cause 1 were observed as failing from cause 2; and, 40.59% of those who truly failed from cause 2 were observed as failing from cause 1. When missing data were MCAR, at 20% double sampling, the complete-case and pseudo-likelihood estimators in general showed good finite-sample performance as the estimates had small bias and attained coverage close to the nominal level. The asymptotic standard errors (ASE) were also close to the Monte Carlo standard deviations thereby increasing confidence in the closed-form variance estimator. These observations held true both under correct and incorrect specifications of the predictive value models. That being said, the pseudo-likelihood estimator was between 55.6% and 90.3% more efficient than the complete-case estimator when the misclassification models for both causes of failure were the same. When the misclassification models for cause 1 and cause 2 were different, the pseudo-likelihood estimator was between 58.1% and 96.8% more efficient than the complete-case estimator.

At 50% double-sampling, both the complete-case and pseudo-likelihood estimators gained efficiency compared to those derived at at 20% double-sampling. This gain in efficiency also came with an attenuation of the relative efficiency gains between the pseudo-likelihood and complete-case estimators. The results of simulations at 20% and 50% double-sampling are presented in Table 2.1.

Results under MAR

When missing data were MAR, the actual level of double-sampling fell short of the planned double-sampling, as the simulation allowed for some non-response (e.g., patient who were double sampled but were not successfully traced in our motivating example). For example, when 20% double sampling was planned, about 13.6% of the non-censored observations were successfully double-sampled. Under MAR, the pseudo-likelihood estimator continued to show the same good finite sample properties as those seen in MCAR. That is, the pseudo-likelihood estimates had small bias, the standard error estimates where close to the Monte-Carlo standard deviations and the estimates attained coverage close to the nominal 95% level. On the other hand, under the auxiliary-variable dependent MAR setting, the complete-case estimator showed more bias than the pseudo-likelihood estimator. The results of simulations performed under MAR are presented in Table 2.2.

Computational efficiency

The comparison of results from the EM and pseudo-likelihood methods is presented in Table 2.3. Compared to the maximum likelihood estimator generated by the EM algorithm, the pseudo-likelihood estimator was generally less efficient with a relative efficiency deficit between 10% and 20%. On the other hand, the estimates derived from the EM algorithm had smaller variability than those from the proposed pseudo-likelihood estimation method.

That being said, the proposed pseudo-likelihood estimation method was computationally faster than the EM algorithm. To compare computational efficiency, I ran a series of experiments where the samples size was increased from 5000 to 10000, 20000, 50000 and 100000 while holding the double-sampling proportion among the non-censored observations at 20% and compared the time it took for the EM and pseudo-likelihood-based methods to converge. The pseudo-likelihood approach was performed in `R` software using the `glm` function with the appropriate weighting specified. The EM algorithm, on the other hand, was programmed into `R` by the study authors. The starting values for the EM algorithm were simulated from a $\mathit{Uniform}(0,1)$ distribution. All the experiments were performed in `R version 3.4.1` on a computer with the following technical specifications: {64 bit, Intel(R) Core(TM) i5-3470 CPU @ 3.2GHz, 8GB Ram}. At all the experimental conditions, the pseudo-likelihood-based approach was found to converge significantly faster than the EM algorithm. The results of comparing the computational speeds of the EM algorithm and the pseudo-likelihood approach are presented in Figure 2.2. At the different sample sizes, and under the computational restrictions of the computer used, the pseudo-likelihood approach was found to converge, on average, 93.6 times faster than the EM algorithm.

Clustered data

The simulation results under clustered data settings are presented in Table 2.4. Similar to the cross-sectional data setting, the pseudo-likelihood method resulted in estimates with small bias both under correct and incorrect predictive-value model specifications. As expected, ignoring the clustering aspect of the data had an impact on variance estimation and coverage: Ignoring the clustering led to underestimation of the variance and under-coverage. When the clustering structure was recognized, the proposed variance estimator was able to correctly estimate the variance of the pseudo-likelihood estimator as shown by the small discrepancies between the asymptotic standard errors and the Monte-Carlo standard deviations (this observation was more apparent with increasing number of clusters).

2.5 Application 1: Estimating misclassification probabilities

2.5.1 Notation

In this application, C^* is defined as the observed cause of failure, and C as the true cause of failure. The observed cause C^* is ascertained by an error-prone approach that results in the under-reporting of death. C, on the other hand, is correctly ascertained. Formally,

$$C^* = \begin{cases} 0 & \text{if censored} \\ 1 & \text{if death is observed} \\ 2 & \text{if disengagement from care is observed} \end{cases}$$

and

$$C = \begin{cases} 0 & \text{if censored} \\ 1 & \text{if true status is death} \\ 2 & \text{if true status is disengagement from care} \end{cases}$$

2.5.2 Goal

The goal of this statistical analysis is to model the probability of classifying subjects as disengaged from care when they are in fact dead, conditional on a set of covariates, that is, $\mathrm{P}[C^* = 2|C = 1;\ \mathit{covariates}]$.

2.5.3 Data

I consider a study consisting of cohorts of PLWH that contribute data to the International Epidemiology Databases for the Evaluation of HIV/AIDS (IeDEA) in East Africa. In this study, patients are followed prospectively from antiretroviral therapy (ART) initiation until death, disengagement from care, or censoring. A patient is considered disengaged from care, if he/she has no recorded visit in the period spanning his/her last visit and two months after the next scheduled visit. There is possible misclassification in this study as some subjects are classified as disengaged from care when they are, in fact, deceased. The outcome of some of the patients who are observed as disengaged from care (i.e., those with $C^* = 2$) is validated by tracing them in the community (double-sampling). Through validation, the true outcome C is observed for these patients, thereby providing information on outcome misclassification. In this analysis, only uni-directional outcome misclassification is considered (i.e., an observed death cannot be a misclassified disengagement). It is worth restating that the proposed method can also work for bi-directional misclassification.

Our analysis of outcome misclassification consisted of 31,179 participants enrolled at the care facilities of AMPATH (Academic Model Providing Access to Healthcare) who had been observed as either dead or disengaged from care (i.e., non-censored). Of these, 28,460(91%) were observed as disengaged from care by the healthcare workers. Outcome validation was performed on 4238(14.9%) of those observed as disengaged from care: Among these cases, 1143(27%) were found to be actually deceased. After outcome validation, the death count increased from 2719 to 3862, meaning that 29.6%(1143/3862) of deaths had initially been misclassified as disengagements from care. The characteristics of patients involved in the misclassification model are summarized in Table 2.5.

2.5.4 Methods

The misclassification probabilities were modeled using the pseudo-likelihood method presented in this paper. First, the predictive value of death $\mathrm{P}[C = 1|C^* = 2;\ covariates]$ were modelled using the 4238 subjects who were observed as disengaged from care and whose outcomes were validated through double-sampling. It was not necessary to model the predictive value for disengagement because observed deaths were always correctly ascertained, so that $\mathrm{P}[C = 2|C^* = 1;\ covariates] = 0$.

The covariates considered included gender(male versus female), age at ART initiation, CD4 count at ART initiation and time contributed to the study (in months). The functional forms of the covariates and overall goodness-of-fit were verified using the Supremum goodness-of-fit test (D. Y. Lin, Wei, and Ying 2002). There was evidence that the proposed predictive value model fit the data well (goodness-of-fit test p-value=0.169).

Using the same set of covariates considered in the predictive value model, a model for the misclassification probabilities, $\mathrm{P}[C^* = 2|C = 1,\ covariates]$ was fit. Model goodness-of-fit was assessed using the Supremum goodness-of-fit test at the 0.05 alpha level.

2.5.5 Results

The misclassification models resulting from performing a complete-case analysis and a pseudo-likelihood-based analysis are presented in Table 2.6. There was evidence that the proposed model was a good fit to the data (goodness-of-fit test p-value=0.641). The complete-case analysis consisted of $3,862(12\% \, \mathrm{of} \, 31,179)$ subjects with verified deaths. In the pseudo-likelihood estimation, $3,862$ subjects with verified deaths were each assigned weight $= 1$, whereas the remaining $24,222 (78\% \text{ of } 31,179)$ subjects, without verified outcomes, were weighted based on modeled predictive values ($0 < \text{weight} < 1$).

At the 0.05 alpha level, the complete-case model suggested a significant association between death misclassification and square-root of CD4 count at ART initiation, age at ART initiation, and time spent in the study. The pseudo-likelihood model suggested significant associations between death misclassification and gender, the square root of CD4 count at ART initiation and time spent in the study. In this case, the association between death misclassification and time was found to be time-dependent, therefore the time spent in the study was entered into model in a piece-wise linear form. Before month 3, there was a positive association between death misclassification and study time, and this positive association began to attenuate beyond month 3. By month 12, the association between death misclassification and study time had become negative. Beyond month 12, the log odds of death misclassification were found to decline by 0.01 units for each additional month of follow-up, holding constant all the other factors. It is also worth noting that, as expected, the estimates from the pseudo-likelihood method had smaller standard errors than those from the complete-case analysis.

2.6 Application 2: Adjusting for misclassification probabilities from an external study

This section illustrates how the misclassification probabilities estimated from an external study can be used in a situation where no outcome validation has been performed. Misclassification probabilities derived from a treatment program with an available internal-validation sample are used inform the possible misclassification in a treatment program that does not have outcome validation. This borrowed information is then used to adjust the observed estimator of the cumulative incidence of death in the program without a validation sample. In the motivating study, the AMPATH program traced its patients in the community, but the FACES (Family AIDS Care & Education Services) program did not. The differential death misclassification in AMPATH was modeled as shown in Section 2.5. Similar modeling could not be performed in the FACES cohort because of the lack of validation data. Under the transportability assumption, we assumed the death misclassification model for FACES was the same as that in AMPATH. The resulting misclassification probabilities were then used to adjust the observed cumulative incidence of death at FACES for possible death misclassification.

The external validation analysis was performed using data from 3886 patients enrolled in FACES. Of these 73 (1.88%) were observed as deceased, 1541(39.66%) were observed as disengaged from care, and 2272 (58.47%) were censored. None of the observed disengagements were validated in the FACES cohort. Using the misclassification probabilities from the pseudo-likelihood method as shown in Table 2.6, the cumulative incidence of death in the FACES cohort was adjusted for possible death misclassification. The results of the adjustment are shown in Figure 2.3. The technical details for adjusting the cumulative incidence function for misclassification can be found in (Bakoyannis and Yiannoutsos 2015). In the FACES cohort, the na"{ı}ve cumulative incidence function estimate of mortality at 12 months after ART initiation was about 1.9%, whereas the misclassification-adjusted cumulative incidence function estimate of mortality at 12 months was about 6.4%. That is, the misclassification-adjusted mortality was about 3.37 times the unadjusted mortality within the first year of follow-up.

2.7 Discussion

In this chapter, I present a pseudo-likelihood method of estimating binary misclassification probabilities in the presence of an internal validation sample. I note that internal validation allows for the identification of the extent to which a diagnostic procedure/classifier fails to correctly classify the outcomes. Internal validation of outcomes tends to be very expensive; it is, therefore, only performed on a subset of the main study sample. Moreover, not every study unit that is earmarked for validation is available to provide an outcome. Consequently, when using data with internal validation, researchers invariably contend with both missing-by-design and non-response analytic challenges.

With these considerations in mind, I formulated the problem of estimating misclassification probabilities for binary outcome data in the presence of internal validation as a missing data problem. Under the missing at random (MAR) assumption, I proposed a method that relies on imputing the missing binary outcomes among the non-validated observations using predictive values estimated from observations with outcome validation. This imputation changes the likelihood into a pseudo-likelihood, and the estimation of the parameters of interest involves the maximization of the corresponding pseudo-log-likelihood (estimated log-likelihood). The resulting maximum pseudo-likelihood estimates were found to have good large and finite-sample properties both in cross-sectional and clustered data settings wherein clustered units are correlated. The resulting estimates had small bias and their variance resulted in correct coverage probabilities. The closed-form variance estimator developed in this paper, accounts for variability due to the data generating process, estimation of predictive values that were imputed and estimation of misclassification probabilities. Our simulations also showed that the pseudo-likelihood estimates were substantially more efficient than the complete-case estimates. This gain in efficiency is due to the fact that the pseudo-likelihood method allows for the use of the entire study sample during estimation of misclassification probabilities. The observed gain in the efficiency of estimates is not a trivial matter, especially if one considers the costs associated with collecting and validating the data. By running a complete-case analysis, one only uses the validated data, which are only a fraction of the full study sample resulting in significant loss of statistical efficiency. We also saw that bias can become a problem for complete-case analysis when the missingness was explained by auxiliary covariates. Under similar circumstances, the pseudo-likelihood estimator had small bias because it depended on predictive values that adjusted for auxiliary covariates. In using our proposed pseudo-likelihood estimator, one can possibly make gains in both estimation and precision.

That being said, the proposed pseudo-likelihood approach is not a "panacea" or the only solution. One could either use the EM algorithm or multiple imputation to address the missing data problems addressed in this paper. Multiple imputation can be directly implemented in many statistical software without much programming from the analyst. The main challenge when using multiple imputation is that one has to contend with the congeniality issue (Meng 1994). That is, one has to ensure compatibility between the imputation and the analysis models (Tilling et al. 2016). The lack of congeniality can lead to biased variance estimation when using multiple imputation (Robins and Wang, 2000). One need not contend with the somewhat "esoteric" concept of congeniality when using the pseudo-likelihood approach. In a comparison of the EM algorithm to the pseudo-likelihood approach, simulations showed that the EM algorithm results in maximum likelihood estimates which are more efficient than the maximum pseudo-likelihood likelihood estimates from our proposed method. That said, the EM algorithm is much more difficult to implement compared to our method which can be implemented with off-the-shelf software. The EM algorithm is also more computationally intensive. In a series of simulation experiments at increasing sample sizes, the pseudo-likelihood method was found to be, on average, 93.6 times faster than the EM algorithm. For studies that involve large datasets, and in simulation analyses that require many replications, it may be worthwhile to use the proposed pseudo-likelihood estimation in order to speed up computation, notwithstanding the gains in statistical efficiency afforded by the EM algorithm, especially given the ease of implementation via existing statistical software. The pseudo-likelihood described in this article can be easily implemented using the `glm` function in the `R` software. The variance estimator of the pseudo-likelihood estimator is also relatively easy to define in the `R` software.

One may take issue with our use of parametric estimation, since misspecification of the conditional mean model can lead to inconsistent estimates. Our decision to present a parametric method was driven largely by pragmatic considerations. In practice, logistic regression is widely used to model binary outcome data, and is accessible to practitioners with different levels of statistical training. It may be worthwhile to consider flexible penalized parametric models such as those discussed by Zhang and Little (2009) (G. Zhang and Little 2009) to build the predictive model used in imputing the values for the non-validated observations. In addition, when fitting the predictive value models, practitioners need to be wary of auxiliary covariates that make the MAR assumption plausible: Omitting important auxiliary covariates when building the predictive value models can bias the pseudo-likelihood estimation.

I hope the reader is convinced that the process of estimating of misclassification probabilities is one that should be undertaken carefully. In the presence of validation sampling, many practitioners only use the validated sample to learn about the extent of misclassification. In this chapter, I have shown that the discarding of the unvalidated observations not only may lead to loss in efficiency but in some instances may lead to biased estimation of the targeted misclassification probabilities. These findings suggest that, at minimum, practitioner needs to be more deliberative when estimating misclassification probabilities. The reason for the added caution/deliberation is that misclassification probabilities play an important role in adjusting statistical estimators of interest for misclassification bias. In our motivating example consisting of cohorts of patients from IeDEA East Africa, one important goal is that of correctly modeling quantities such as the cause-specific hazards and the cumulative incidence functions. This goal is, however, complicated by death under-reporting, as some patients are considered disengaged from care when they are, in fact, deceased. Using the collected data as-is may lead to the underestimation of the cumulative incidence of death, which in turn can have important implications on aspects of treatment-program such as funding, implementation, and so on. In order to reduce the extent of death-underreporting, IeDEA East Africa has made a large investment in validating the outcomes of some patients considered disengaged from care by tracing them in their communities. This validation yields information that can be used adjust na"{ı}ve estimates of the cumulative incidence of death. In our application consisting of patients from AMPATH, the presence of validation sample allowed us to estimate differential death-misclassification probabilities as efficiently as possible. The same estimation, however, could not be done in FACES cohort because FACES did not perform outcome validation. We, therefore, had to rely on misclassification information from AMPATH to make misclassification adjustments on the cumulative incidence of death at FACES, assuming transportability of misclassification. After adjustment, the 12-month mortality at FACES was estimated to be about 6.4%–a value that was least 3-fold higher than the naive 12-month cumulative incidence of about 1.9%. This change, in our opinion, delineates the importance of statistically principled ways of estimating misclassification probabilities.

2.8 Appendix

2.8.1 Likelihood

Based on the data assumptions presented in Section 2.2, the full-likelihood of the observed data is derived as follows:

$$
\begin{aligned}
L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) &= \prod_{i=1}^{n} \mathrm{P}[T_i, C_i^*, X_i^*] \\
&\propto \prod_{i=1}^{n} \mathrm{P}[T_i, C_i^* | X_i^*] \\
&= \prod_{i=1}^{n} \mathrm{P}[T_i, C_i^*, C_i = 1 | X_i^*]^{I[C_i=1]} \mathrm{P}[T_i, C_i^*, C_i = 2 | X_i^*]^{I[C_i=2]} \\
&= \prod_{i=1}^{n} \left(\mathrm{P}[T_i | X_i^*] \mathrm{P}[C_i^*, C_i = 1 | X_i, T_i]\right)^{I[C_i=1]} \left(\mathrm{P}[T_i | X_i^*] \mathrm{P}[C_i^*, C = 2 | X_i^*, T_i]\right)^{I[C_i=2]} \\
&\propto \prod_{i=1}^{n} \left(\mathrm{P}[C_i^*, C_i = 1 | X_i^*, T_i]\right)^{I[C_i=1]} \left(\mathrm{P}[C_i^*, C_i = 2 | X_i^*, T_i]\right)^{I[C_i=2]} \\
&= \prod_{i=1}^{n} \left(\mathrm{P}[C_i = 1 | X_i^*, T_i] \mathrm{P}[C_i^* | X_i^*, T_i, C_i = 1]\right)^{I[C_i=1]} \left(\mathrm{P}[C_i = 2 | X_i^*, T_i] \mathrm{P}[C_i^* | X_i^*, T_i, C_i = 2]\right)^{I[C_i=2]} \\
&\propto \prod_{i=1}^{n} \left(\mathrm{P}[C_i^* | X_i^*, T_i, C_i = 1]\right)^{I[C_i=1]} \left(\mathrm{P}[C_i^* | X_i^*, T_i, C_i = 2]\right)^{I[C_i=2]} \\
&= \prod_{i=1}^{n} \Big(\mathrm{P}[C_i^* = 1 | X_i^*, T_i, C_i = 1]^{I[C_i^*=1]} \mathrm{P}[C_i^* = 2 | X_i^*, T_i, C_i = 1]^{I[C_i^*=2]}\Big)^{I[C_i=1]} \\
&\times \prod_{i=1}^{n} \Big(\mathrm{P}[C_i^* = 1 | X_i^*, T_i, C_i = 2]^{I[C_i^*=1]} \mathrm{P}[C_i^* = 2 | X_i^*, T_i, C_i = 2]^{I[C_i^*=2]}\Big)^{I[C_i=2]} \\
&= \prod_{i=1}^{n} \Big\{(1 - \pi_{21}^*(\boldsymbol{\beta}_1; \boldsymbol{X}_i))^{I[C_i^*=1]} \, \pi_{21}^*(\boldsymbol{\beta}_1; \boldsymbol{X}_i)^{I[C_i^*=2]}\Big\}^{I[C_i=1]} \\
&\times \prod_{i=1}^{n} \Big\{\pi_{12}^*(\boldsymbol{\beta}_2; \boldsymbol{X}_i)^{I[C_i^*=1]} \, (1 - \pi_{12}^*(\boldsymbol{\beta}_2; \boldsymbol{X}_i))^{I[C_i^*=2]}\Big\}^{I[C_i=2]}
\end{aligned}
$$

2.8.2 Proving consistency

Our goal is show that $||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| \xrightarrow{p} 0$. Without loss of generality, let's begin by showing that the parameter of interest, $\boldsymbol{\beta}_1$ in the score equation 2.10 is identifiable–that is, it exists and is unique. Existence is shown by proving that, $\boldsymbol{\Psi}(\boldsymbol{\beta}_1, \boldsymbol{\gamma}) = 0$ if the mean model is correctly specified, and uniqueness by showing that $\frac{d}{d\boldsymbol{\beta}}\boldsymbol{\Psi}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1}$ is negative-definite.

Under the true model P_θ, the expected score contribution for i-th study unit is given by

$$\begin{aligned}
\boldsymbol{\Psi}(\boldsymbol{\beta}_1, \boldsymbol{\gamma}) &= P\psi_i(\boldsymbol{\beta}_1, \boldsymbol{\gamma}) \\
&= \mathrm{E}\left\{ \boldsymbol{X}_i^T \left\{ R_i\delta_{1i} + (1-R_i)[1 - p_{21}(\boldsymbol{\gamma}_1; \boldsymbol{Z}_i)]^{\delta_{1i}^*} p_{12}(\boldsymbol{\gamma}_2; \boldsymbol{Z}_i)^{1-\delta_{1i}^*} \right\} \left[(1-\delta_{1i}^*) - \frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} \right] \right\} \\
&= \mathrm{E}\left\{ \boldsymbol{X}_i^T g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) \left[(1-\delta_{1i}^*) - \frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} \right] \right\}, \text{ where,} \\
& g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) = \left\{ R_i\delta_{1i} + (1-R_i)[1 - p_{21}(\boldsymbol{\gamma}_1;\, \boldsymbol{Z}_i)]^{\delta_{1i}^*} p_{12}(\boldsymbol{\gamma}_2;\, \boldsymbol{Z}_i)^{1-\delta_{1i}^*} \right\}; \\
& \text{and}\, \Delta_{1i} = (\delta_{1i}, \delta_{1i}^*);\ \boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2), \\
&= \mathrm{E}_W\left\{ E\left[\boldsymbol{X}_i^T g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) \left((1-\delta_{1i}^*) - \frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} \right) \middle| \mathbf{W}_i = (\boldsymbol{X}_i \cup \boldsymbol{Z}_i) \right] \right\} \\
&= \mathrm{E}_W\left\{ \boldsymbol{X}_i^T E_{\Delta_1}\left\{ g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) \left[E\left(1 - \delta_{1i}^* \middle| \boldsymbol{X}_i, \delta_{1i} \right) - \frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} \right] \right\} \right\}, \\
& \text{if model is correctly specified,} \\
&= \mathrm{E}_W\left\{ \boldsymbol{X}_i^T E_{\Delta_1}\left\{ g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) \left[\frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} - \frac{\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}}{1+\exp{(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}} \right] \right\} \right\}, \\
&= 0.
\end{aligned}$$

In addition, the second derivative of the score function is clearly negative definite as shown below:

$$\frac{d}{d\boldsymbol{\beta}}\boldsymbol{\Psi}_n(\boldsymbol{\beta}, \boldsymbol{\gamma})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_1} = -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T g(\boldsymbol{\gamma}; R_i, \Delta_{1i}, \boldsymbol{Z}_i) \frac{\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)}{\left(1+\exp(\boldsymbol{X}_i^T\boldsymbol{\beta}_1)\right)^2}\boldsymbol{X}_i$$

Since existence and uniqueness conditions hold, conclude that $\boldsymbol{\beta}_1$ is identifiable.

Without loss of generality, $\boldsymbol{\beta}_2$ is also identifiable. We prove consistency by showing that the class of functions, $\{\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) : \boldsymbol{\beta} \in \mathcal{B}\}$, indexed by $\boldsymbol{\beta} \in \mathcal{B}$ is P Glivenko-Cantelli. With that proof, it would follow that $||\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0|| \xrightarrow{p} 0$. This is because, $\sup_{\boldsymbol{\beta}\in\mathcal{B}} ||\Psi_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n) - \Psi(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)|| \xrightarrow{p} 0$, since, by countable sub-additivity of norms,

$$\begin{aligned}
\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n) - \Psi(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| &= \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n) - \Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) + \Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \Psi(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| \\
&\leq \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n) - \Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| \\
&+ \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \Psi(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n) - \Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| + o_p(1), \text{ by the law of large numbers} \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|\Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) + (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0).\dot{\Psi}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \Psi_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| + o_p(\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|),
\end{aligned}$$

Then by Taylor series expansion, the above is

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathcal{B}} \|(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)\left[\dot{\Psi}_n(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) - \dot{\Psi}(\boldsymbol{\beta}, \boldsymbol{\gamma}_0) + \dot{\Psi}(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\right]\| + o_p(\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|) \\
&= \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|(\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0).o_p(1) + (\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0)\dot{\Psi}(\boldsymbol{\beta}, \boldsymbol{\gamma}_0)\| + o_p(\|\widehat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\|) \\
&= o_p(1)o_p(1) + O_p(1)o_p(1) + o_p(O_p(n^{-1/2})) \\
&= o_p(1)
\end{aligned}$$

2.8.3 Proving asymptotic Normality

At the maximum pseudo-likelihood estimate,

$$\begin{aligned}
0 &= \Psi_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_{n_v}) \\
&= \Psi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\gamma}_0) + \left[\Psi_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_{n_v}) - \Psi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\gamma}_0)\right] \qquad (2.13)
\end{aligned}$$

where n is the size of the main sample, n_v is the size of internal-validation sample and,

$\Psi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\gamma}_0)$ is the average of the estimated score function when $\boldsymbol{\gamma}_0$ is known. Through a number of Taylor Series expansions and algebraic steps on 2.13, it can be deduced that:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{I}^{-1}\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\left[\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|\boldsymbol{X}_i) + \sqrt{s}.\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{X}_i)\right] + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(\boldsymbol{X}_i|\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + o_p(1), \qquad (2.14)\end{aligned}$$

where $\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \left[\frac{d}{d\boldsymbol{\gamma}}\Psi_n(\boldsymbol{\beta}_0, \boldsymbol{\gamma})|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\right]$ is a $q \times d$ matrix; $s = \frac{n}{n_v}$ as $n \to \infty$ and

$$\tilde{\psi}(\boldsymbol{X}_i|\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \boldsymbol{I}^{-1}\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\left[\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|\boldsymbol{X}_i) + \sqrt{s}.\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{X}_i)\right]$$

By the central limit theorem:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(\boldsymbol{X}_i|\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + o_p(1) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega})$$

where

$$\boldsymbol{\Omega} \quad = \boldsymbol{I}^{-1}(\beta_0, \gamma_0) + s.\boldsymbol{I}^{-1}(\beta_0, \gamma_0)\mathbf{W}(\beta_0, \gamma_0, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{I}^{-1}(\beta_0, \gamma_0)$$

and,

$$\begin{aligned}\mathbf{W}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, \boldsymbol{X}, \boldsymbol{Z}) = {} & \mathrm{E}\left[\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{Z})\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|\boldsymbol{X})^T\right] \\ & + \mathrm{E}\left[\dot{l}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0|\boldsymbol{X})\dot{l}(\boldsymbol{\gamma}_0|\boldsymbol{Z})^T\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^T\right] \\ & + \mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\gamma}_0)\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)^T\end{aligned}$$

$\boldsymbol{\Omega}$ can be estimated by replacing the parameter $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ with their consistent estimators so that:

$$\hat{\boldsymbol{\Omega}}_n = \frac{1}{n}\sum_{i=1}^{n}\tilde{\psi}(\boldsymbol{X}_i|\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)\tilde{\psi}(\boldsymbol{X}_i|\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n)^T.$$

2.8.4 Clustered data

In the clustered-data setting, I will prove the asymptotic normality of the pseudo-likelihood estimator, and also derive the formula for the variance estimator. Assume that in each cluster $i \in \{1, 2, ..., m\}$ with n_i units, the number of doubled-sampled units is $n_{i(v)}$. Also assume that $\frac{n_i}{n_{i(v)}} = s$ as cluster size, n_i, increases to ∞ for all $i = 1, 2, ..., m$. Here, think of s as the inverse proportion of the cluster that is double sampled. At the optimal point, $\frac{1}{m}\sum_{i=1}^{m} U_{i.}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m) = 0$. Begin by recognizing that:

$$\begin{aligned}\frac{1}{m}\sum_{i=1}^{m} U_{i.}(\widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{\gamma}}_m) &= \frac{1}{m}\sum_{i=1}^{m}\left[U_{i.}(\widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{\gamma}}_m) + U_{i.}(\widehat{\boldsymbol{\beta}}_m, \boldsymbol{\gamma}_0) - U_{i.}(\widehat{\boldsymbol{\beta}}_m, \boldsymbol{\gamma}_0)\right] \\ &= \frac{1}{m}\sum_{i=1}^{m} U_{i.}(\widehat{\boldsymbol{\beta}}_m, \boldsymbol{\gamma}_0) + \frac{1}{m}\sum_{i=1}^{m}\left[U_{i.}(\widehat{\boldsymbol{\beta}}_m, \widehat{\boldsymbol{\gamma}}_m) - U_{i.}(\widehat{\boldsymbol{\beta}}_m, \boldsymbol{\gamma}_0)\right].\end{aligned}$$

After some Taylor series expansions and algebraic steps, it can be deduced that the influence function is given by:

$$\sqrt{m}\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{\psi}_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + o_p(1)$$

where $\tilde{\psi}_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = -\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)U_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) - \sqrt{s}.\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)I^{-1}(\boldsymbol{\gamma}_0)U_{i.}(\boldsymbol{\gamma}_0)$, with $\mathbf{R}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \frac{1}{m}\sum_{i=1}^{m}\frac{d}{d\gamma}U_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma})|_{\gamma=\gamma_0}$. It would then follow by the central limit theorem that

$$\sqrt{m}\left(\hat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0\right) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\tilde{\psi}_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) + o_p(1) \rightarrow N(\mathbf{0}, \boldsymbol{\Omega})$$

where, $\boldsymbol{\Omega} = \mathrm{E}[\tilde{\psi}_{i.}(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\tilde{\psi}_{i.}^T(\boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)]$. Empirically, $\boldsymbol{\Omega}$ is estimated by replacing parameters with their consistent estimates, that is,

$$\hat{\boldsymbol{\Omega}}_m = \mathrm{E}[\tilde{\psi}_{i.}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m)\tilde{\psi}_{i.}^T(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m)] = \frac{1}{m}\sum_{i=1}^{m}\tilde{\psi}_{i.}(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m)\tilde{\psi}_{i.}^T(\hat{\boldsymbol{\beta}}_m, \hat{\boldsymbol{\gamma}}_m)$$

2.8.5 Model misspecification

In the proposed pseudo-likelihood estimation, when true outcome is missing, it is replaced by its predictive value as defined above in sub-section 2.3.1. Also as stated in Section 2.3, I have chose to estimate the predictive values parametrically, despite the risk of biased estimation if the conditional mean model is misspecified. I conceded that, it would be up to the practitioner to perform appropriate goodness of fit tests, and, if need be, implement necessary model remedies. Now, I will formally explore the impact of misspecifying the predictive value model on parameter estimation (assuming there is no omission important variables).

Recall that the predictive value for true cause 1 given cause 2 is observed is defined as follows

$$\mathrm{P}[C=1|T=t,C^*=2,\boldsymbol{Z}]=\frac{\mathrm{P}[C^*=2|C=1,T=t,\boldsymbol{Z}]\mathrm{P}[C=1,T=t|\boldsymbol{Z}]}{\mathrm{P}[C^*=2|C=1,T=t,\boldsymbol{Z}]\mathrm{P}[C=1,T=t|\boldsymbol{Z}]+\mathrm{P}[C^*=2|C=2,T=t,\boldsymbol{Z}]\mathrm{P}[C=2,T=t|\boldsymbol{Z}]}$$

This predictive value equation can also be expressed as:

$$\mathrm{P}[C=1|T=t,C^*=2,\boldsymbol{Z}]=\frac{\frac{\lambda_1(t;\boldsymbol{Z})}{\lambda_1(t;\boldsymbol{Z})+\lambda_2(t;\boldsymbol{Z})}\pi^*_{21}(t;\boldsymbol{Z})}{\frac{\lambda_1(t;\boldsymbol{Z})}{\lambda_1(t;\boldsymbol{Z})+\lambda_2(t;\boldsymbol{Z})}\pi^*_{21}(t;\boldsymbol{Z})+\frac{\lambda_2(t;\boldsymbol{Z})}{\lambda_1(t;\boldsymbol{Z})+\lambda_2(t;\boldsymbol{Z})}\pi^*_{22}(t;\boldsymbol{Z})}$$

Recognizing that $\frac{\mathrm{P}[C=1|T=t,C^*=2,\boldsymbol{Z}]}{\mathrm{P}[C=2|T=t,C^*=2,\boldsymbol{Z}]}=\frac{\lambda_1(t;\boldsymbol{Z})\pi^*_{21}(t;\boldsymbol{Z})}{\lambda_2(t;\boldsymbol{Z})\pi^*_{22}(t;\boldsymbol{Z})}$, taking the log, and performing some algebraic operations, under the Weibull settings in Section 2.4.1, one can derive that

$$\log\left[\frac{\mathrm{P}[C=1|T=t,C^*=2,\boldsymbol{Z}]}{\mathrm{P}[C=2|T=t,C^*=2,\boldsymbol{Z}]}\right]=\log\left(\frac{\lambda_1}{\lambda_2}\right)+[\kappa_1^T-\kappa_2^T]\boldsymbol{Z}+\boldsymbol{Z}\boldsymbol{\beta}_1+\log\left[\frac{1+\exp(\boldsymbol{Z}\boldsymbol{\beta}_2)}{1+\exp(\boldsymbol{Z}\boldsymbol{\beta}_1)}\right]. \tag{2.15}$$

Based on Equation 2.15 above, the linear relationship between log odds and covariates is preserved only when $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$. That is, logistic regression is correct model for predictive values if $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$. Fitting a logistic model when $\boldsymbol{\beta}_1\neq\boldsymbol{\beta}_2$ is a form of model misspecification. I explored the impact of misspecification in the simulations by setting: $\boldsymbol{\beta}_2=(-0.3,0.2,0.5,0.5)$.

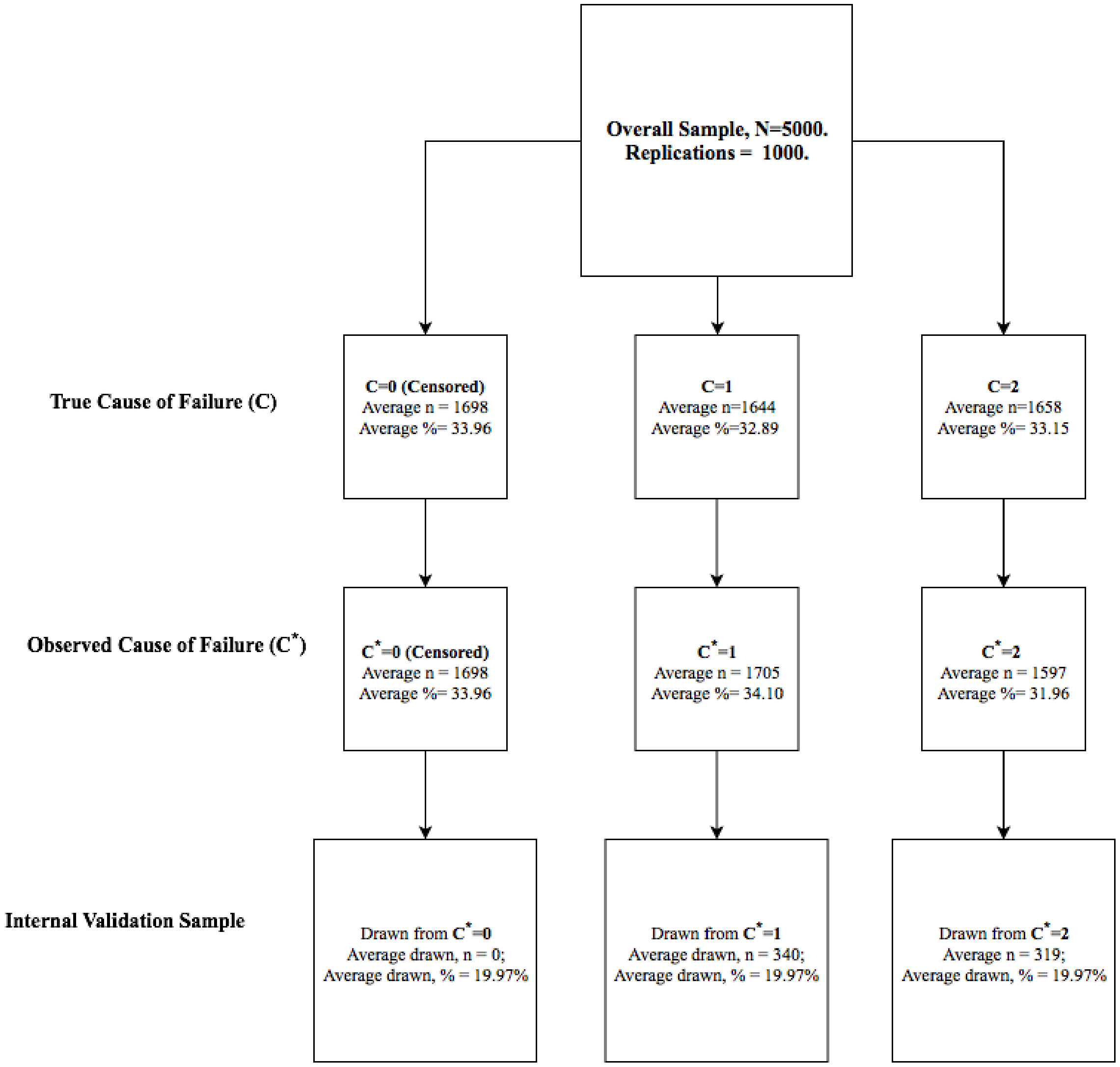

Figure 2.1: Summary of simulation samples used when double sampling was set at 20%.

				Complete Case Estimator					**Pseudo-likelihood Estimator**					
Model, (ds %)	Cause	Parameter	**Truth**	Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP	RE
Correct (20%)	1	β_{01}(Intercept)	**-0.4**	-0.421	5.25	0.353	0.343	0.944	-0.419	4.75	0.281	0.275	0.941	1.556
		β_{11} (t)	**-0.4**	-0.399	0.25	0.359	0.348	0.945	-0.398	0.50	0.284	0.275	0.931	1.601
		β_{21} (z_1)	**0.5**	0.525	5.00	0.414	0.416	0.952	0.520	4.00	0.325	0.324	0.941	1.649
		β_{31} (z_2)	**-0.5**	-0.499	0.20	0.144	0.146	0.945	-0.498	0.40	0.114	0.114	0.946	1.640
	2	β_{02}(Intercept)	**-0.4**	-0.408	2.00	0.298	0.292	0.950	-0.409	2.25	0.215	0.214	0.948	1.862
		β_{12} (t)	**-0.4**	-0.410	2.50	0.306	0.298	0.943	-0.409	2.25	0.222	0.216	0.949	1.903
		β_{22} (z_1)	**0.5**	0.517	3.40	0.416	0.406	0.942	0.518	3.60	0.314	0.309	0.948	1.726
		β_{32} (z_2)	**-0.5**	-0.506	1.20	0.145	0.142	0.945	-0.503	0.60	0.109	0.107	0.943	1.761
Incorrect (20%)	1	β_{01}(Intercept)	**-0.4**	-0.397	0.75	0.347	0.342	0.933	-0.399	0.25	0.281	0.272	0.940	1.581
		β_{11} (t)	**-0.4**	-0.418	4.50	0.358	0.349	0.942	-0.397	0.75	0.276	0.272	0.942	1.646
		β_{21} (z_1)	**0.5**	0.505	1.00	0.421	0.416	0.943	0.484	3.20	0.316	0.317	0.949	1.722
		β_{31} (z_2)	**-0.5**	-0.510	2.00	0.151	0.147	0.943	-0.497	0.60	0.111	0.109	0.937	1.819
	2	β_{02}(Intercept)	**-0.3**	-0.315	5.00	0.298	0.290	0.936	-0.302	0.67	0.213	0.207	0.937	1.963
		β_{12} (t)	**0.2**	0.215	7.50	0.296	0.296	0.947	0.206	3.00	0.213	0.211	0.939	1.968
		β_{22} (z_1)	**0.5**	0.514	2.80	0.415	0.401	0.944	0.496	0.80	0.301	0.298	0.947	1.811
		β_{32} (z_2)	**0.5**	0.517	3.40	0.139	0.141	0.955	0.503	0.60	0.101	0.102	0.948	1.911
Correct (50%)	1	β_{01}(Intercept)	**-0.4**	-0.394	1.50	0.214	0.214	0.952	-0.397	0.75	0.187	0.189	0.954	1.282
		β_{11} (t)	**-0.4**	-0.405	1.25	0.221	0.217	0.952	-0.400	0.00	0.190	0.190	0.956	1.304
		β_{21} (z_1)	**0.5**	0.491	1.80	0.262	0.260	0.951	0.492	1.60	0.231	0.226	0.947	1.324
		β_{31} (z_2)	**-0.5**	-0.501	0.20	0.089	0.092	0.950	-0.499	0.20	0.079	0.079	0.950	1.356
	2	β_{02}(Intercept)	**-0.4**	-0.406	1.50	0.181	0.183	0.952	-0.406	1.50	0.155	0.154	0.943	1.412
		β_{12} (t)	**-0.4**	-0.398	0.50	0.186	0.186	0.946	-0.399	0.25	0.157	0.156	0.956	1.422
		β_{22} (z_1)	**0.5**	0.506	1.20	0.254	0.255	0.950	0.506	1.20	0.218	0.218	0.943	1.368
		β_{32} (z_2)	**-0.5**	-0.499	0.20	0.089	0.089	0.949	-0.500	0.00	0.076	0.076	0.946	1.371
Incorrect (50%)	1	β_{01}(Intercept)	**-0.4**	-0.395	1.25	0.211	0.214	0.949	-0.394	1.50	0.183	0.188	0.964	1.296
		β_{11} (t)	**-0.4**	-0.401	0.25	0.219	0.216	0.943	-0.398	0.50	0.188	0.188	0.952	1.320
		β_{21} (z_1)	**0.5**	0.492	1.60	0.255	0.261	0.954	0.486	2.80	0.222	0.224	0.955	1.358
		β_{31} (z_2)	**-0.5**	-0.502	0.40	0.091	0.092	0.955	-0.498	0.40	0.077	0.078	0.953	1.391
	2	β_{02}(Intercept)	**-0.3**	-0.298	0.67	0.186	0.181	0.944	-0.295	1.67	0.157	0.151	0.932	1.437
		β_{12} (t)	**0.2**	0.202	1.00	0.191	0.184	0.943	0.205	2.50	0.161	0.152	0.928	1.465
		β_{22} (z_1)	**0.5**	0.493	1.40	0.254	0.251	0.947	0.483	3.40	0.218	0.213	0.943	1.389
		β_{32} (z_2)	**0.5**	0.505	1.00	0.090	0.088	0.931	0.502	0.40	0.077	0.074	0.934	1.414

Table 2.1: Comparison of finite-sample properties of complete-case estimator and the pseudo-likelihood estimator when data are missing completely at random (MCAR). Simulations were performed at 20% and 50% double-sampling and under correct and incorrect model specification. (ds%) represents the double-sampling percent among the non-censored observations.

				Complete Case Estimator					**Pseudo-likelihood Estimator**					
Model*,Planned DS%(Actual DS%)**	**Cause**	**Parameter**	**True Value**	Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP	RE
Correct, 20%(13.6%)	1	β_{01}(Intercept)	**-0.40**	-0.492	23.00	0.421	0.424	0.944	-0.420	5.00	0.336	0.331	0.951	1.570
		β_{11} (t)	**-0.40**	-0.404	1.00	0.436	0.432	0.947	-0.399	0.25	0.338	0.328	0.939	1.664
		β_{21} (z_1)	**0.50**	0.526	5.20	0.516	0.514	0.947	0.528	5.60	0.394	0.386	0.951	1.715
		β_{31} (z_2)	**-0.50**	-0.514	2.80	0.176	0.182	0.953	-0.509	1.80	0.133	0.136	0.960	1.751
	2	β_{02}(Intercept)	**-0.40**	-0.489	22.25	0.375	0.360	0.932	-0.415	3.75	0.264	0.251	0.929	2.018
		β_{12} (t)	**-0.40**	-0.405	1.25	0.398	0.368	0.934	-0.407	1.75	0.265	0.253	0.941	2.256
		β_{22} (z_1)	**0.50**	0.517	3.40	0.497	0.500	0.943	0.521	4.20	0.363	0.365	0.945	1.875
		β_{32} (z_2)	**-0.50**	-0.518	3.60	0.185	0.176	0.938	-0.510	2.00	0.132	0.128	0.942	1.964
Correct, 50%(34%)	1	β_{01}(Intercept)	**-0.4**	-0.459	14.75	0.261	0.263	0.943	-0.393	1.75	0.221	0.220	0.953	1.395
		β_{11} (t)	**-0.4**	-0.415	3.75	0.259	0.266	0.958	-0.408	2.00	0.218	0.219	0.951	1.412
		β_{21} (z_1)	**0.5**	0.490	2.00	0.322	0.320	0.947	0.495	1.00	0.263	0.261	0.947	1.499
		β_{31} (z_2)	**-0.5**	-0.512	2.40	0.114	0.112	0.943	-0.510	2.00	0.091	0.092	0.954	1.569
	2	β_{02}(Intercept)	**-0.4**	-0.460	15.00	0.225	0.224	0.940	-0.382	4.50	0.174	0.174	0.946	1.672
		β_{12} (t)	**-0.4**	-0.410	2.50	0.236	0.228	0.944	-0.411	2.75	0.178	0.176	0.946	1.758
		β_{22} (z_1)	**0.5**	0.487	2.60	0.311	0.312	0.952	0.478	4.40	0.256	0.250	0.939	1.476
		β_{32} (z_2)	**-0.5**	-0.512	2.40	0.113	0.109	0.938	-0.509	1.80	0.090	0.087	0.943	1.576
Incorrect, 20%(13.6%)	1	β_{01}(Intercept)	**-0.4**	-0.483	20.75	0.432	0.424	0.931	-0.403	0.75	0.326	0.325	0.950	1.756
		β_{11} (t)	**-0.4**	-0.407	1.75	0.421	0.431	0.950	-0.402	0.50	0.317	0.320	0.945	1.764
		β_{21} (z_1)	**0.5**	0.509	1.80	0.533	0.514	0.943	0.497	0.60	0.386	0.375	0.946	1.907
		β_{31} (z_2)	**-0.5**	-0.516	3.20	0.174	0.181	0.951	-0.506	1.20	0.124	0.128	0.954	1.969
	2	β_{02}(Intercept)	**-0.3**	-0.366	22.00	0.364	0.356	0.939	-0.291	3.00	0.247	0.242	0.946	2.172
		β_{12} (t)	**0.2**	0.191	4.50	0.362	0.364	0.955	0.196	2.00	0.245	0.246	0.946	2.183
		β_{22} (z_1)	**0.5**	0.504	0.80	0.480	0.490	0.955	0.485	3.00	0.340	0.351	0.965	1.993
		β_{32} (z_2)	**0.5**	0.506	1.20	0.168	0.172	0.949	0.499	0.20	0.114	0.119	0.959	2.172
Incorrect, 50%(34%)	1	β_{01}(Intercept)	**-0.4**	-0.475	18.75	0.266	0.262	0.941	-0.402	0.50	0.220	0.218	0.945	1.462
		β_{11} (t)	**-0.4**	-0.397	0.75	0.261	0.265	0.941	-0.396	1.00	0.210	0.216	0.951	1.545
		β_{21} (z_1)	**0.5**	0.501	0.20	0.324	0.319	0.949	0.495	1.00	0.267	0.256	0.927	1.473
		β_{31} (z_2)	**-0.5**	-0.508	1.60	0.111	0.112	0.951	-0.503	0.60	0.089	0.088	0.941	1.555
	2	β_{02}(Intercept)	**-0.3**	-0.372	24.00	0.227	0.222	0.937	-0.295	1.67	0.173	0.170	0.937	1.722
		β_{12} (t)	**0.2**	0.199	0.50	0.225	0.226	0.953	0.205	2.50	0.172	0.172	0.953	1.711
		β_{22} (z_1)	**0.5**	0.488	2.40	0.308	0.307	0.953	0.477	4.60	0.235	0.242	0.964	1.718
		β_{32} (z_2)	**0.5**	0.501	0.20	0.109	0.108	0.953	0.498	0.40	0.085	0.083	0.950	1.644

Table 2.2: Comparison of finite-sample properties of complete-case estimator and the pseudo-likelihood estimator when data are missing at random (MAR). Simulations were performed under correct and incorrect predictive-value model specifications(*). In each study, double-sampling (ds) was performed on either 20% or 50% of the non-censored, however due to subject non-response the actual double-sampling was smaller than the planned double-sampling (**). These simulations explore a situation where the actual double-sampling is about 80% of the planned double-sampling among the non-censored.

			Estimation Method						
			Expectation Maximization (EM)			**Pseudo-likelihood**			
Cause	Parameter	**True Value**	Estimate	% Bias	MCSD	Estimate	% Bias	MCSD	RE
1	β_{01} (Intercept)	**-0.4**	-0.415	3.75	0.266	-0.412	3.00	0.280	0.903
	β_{11} (t)	**-0.4**	-0.392	2.00	0.266	-0.405	1.25	0.266	1.000
	β_{21} (z_1)	**0.5**	0.512	2.40	0.302	0.517	3.40	0.337	0.803
	β_{31} (z_2)	**-0.5**	-0.501	0.20	0.105	-0.507	1.40	0.118	0.792
2	β_{02} (Intercept)	**-0.4**	-0.405	1.25	0.202	-0.399	0.25	0.215	0.883
	β_{12} (t)	**-0.4**	-0.405	1.25	0.209	-0.407	1.75	0.215	0.945
	β_{22} (z_1)	**0.5**	0.508	1.60	0.288	0.503	0.60	0.314	0.841
	β_{32} (z_2)	**-0.5**	-0.503	0.60	0.098	-0.505	1.00	0.111	0.779

Table 2.3: Simulation Results: Comparison of finite sample properties of maximum likelihood estimates from EM to pseudo-likelihood estimates. Sample size=5000; Double sampling percent is 20%.

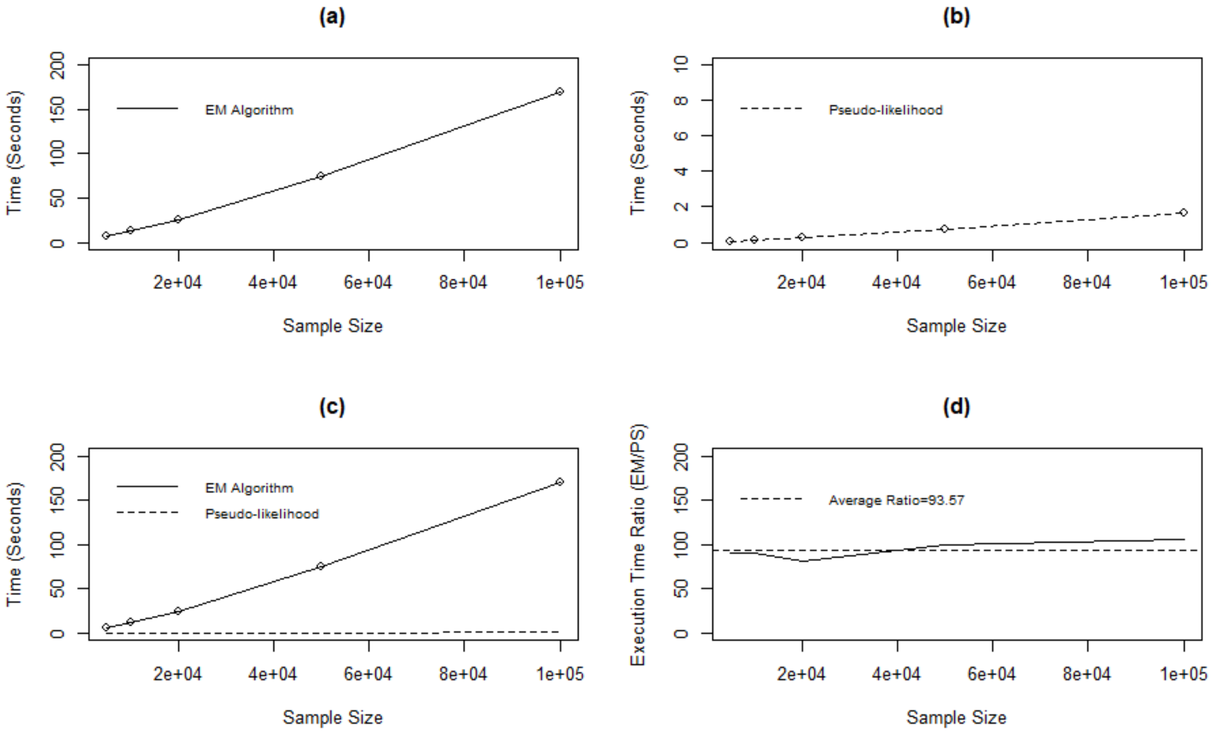

Figure 2.2: Computation time: EM versus pseudo-likelihood approach. As sample size increases: (a) represents the computational time for the EM; (b) represents the computational time of the pseudo-likelihood approach; (c) represents the computational times of the EM and pseudo-likelihood approach on the same time-scale; (d) represents the relative time of the EM versus the pseudo-likelihood approach.

				Ignore Clustering Structure					**Consider Clustering Structure**				
Number of clusters	Cause	Parameter	**Truth**	Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP
100	1	β_{01}(Intercept)	-0.4	-0.410	2.50	0.462	0.285	0.788	-0.410	2.50	0.462	0.439	0.930
		β_{11} (t)	-0.4	-0.410	2.50	0.451	0.286	0.808	-0.410	2.50	0.451	0.427	0.924
		β_{21} (z_1)	0.5	0.511	2.20	0.576	0.336	0.748	0.511	2.20	0.576	0.544	0.932
		β_{31} (z_2)	-0.5	-0.508	1.60	0.203	0.120	0.747	-0.508	1.60	0.203	0.192	0.935
	2	β_{02}(Intercept)	-0.4	-0.420	5.00	0.408	0.222	0.702	-0.420	5.00	0.408	0.393	0.949
		β_{12} (t)	-0.4	-0.409	2.25	0.382	0.226	0.771	-0.409	2.25	0.382	0.369	0.940
		β_{22} (z_1)	0.5	0.528	5.60	0.524	0.318	0.770	0.528	5.60	0.524	0.524	0.951
		β_{32} (z_2)	-0.5	-0.515	3.00	0.186	0.113	0.766	-0.515	3.00	0.186	0.186	0.939
200	1	β_{01}(Intercept)	-0.4	-0.425	6.25	0.316	0.197	0.782	-0.425	6.25	0.316	0.308	0.948
		β_{11} (t)	-0.4	-0.388	3.00	0.313	0.197	0.789	-0.388	3.00	0.313	0.299	0.940
		β_{21} (z_1)	0.5	0.526	5.20	0.393	0.232	0.751	0.526	5.20	0.393	0.383	0.941
		β_{31} (z_2)	-0.5	-0.502	0.40	0.139	0.082	0.754	-0.502	0.40	0.139	0.135	0.946
	2	β_{02}(Intercept)	-0.4	-0.408	2.00	0.273	0.153	0.742	-0.408	2.00	0.273	0.277	0.949
		β_{12} (t)	-0.4	-0.406	1.50	0.272	0.155	0.739	-0.406	1.50	0.272	0.261	0.943
		β_{22} (z_1)	0.5	0.510	2.00	0.378	0.220	0.755	0.510	2.00	0.378	0.371	0.944
		β_{32} (z_2)	-0.5	-0.510	2.00	0.133	0.077	0.761	-0.510	2.00	0.133	0.130	0.935
400	1	β_{01}(Intercept)	-0.4	-0.403	0.75	0.224	0.138	0.769	-0.403	0.75	0.224	0.219	0.947
		β_{11} (t)	-0.4	-0.400	0.00	0.213	0.138	0.814	-0.400	0.00	0.213	0.211	0.941
		β_{21} (z_1)	0.5	0.508	1.60	0.272	0.162	0.755	0.508	1.60	0.272	0.273	0.954
		β_{31} (z_2)	-0.5	-0.503	0.60	0.098	0.057	0.751	-0.503	0.60	0.098	0.095	0.937
	2	β_{02}(Intercept)	-0.4	-0.397	0.75	0.200	0.107	0.709	-0.397	0.75	0.200	0.196	0.955
		β_{12} (t)	-0.4	-0.404	1.00	0.190	0.108	0.726	-0.404	1.00	0.190	0.184	0.943
		β_{22} (z_1)	0.5	0.498	0.40	0.268	0.154	0.740	0.498	0.40	0.268	0.261	0.945
		β_{32} (z_2)	-0.5	-0.499	0.20	0.091	0.054	0.746	-0.499	0.20	0.091	0.092	0.957
800	1	β_{01}(Intercept)	-0.4	-0.394	1.50	0.157	0.097	0.770	-0.394	1.50	0.157	0.154	0.945
		β_{11} (t)	-0.4	-0.409	2.25	0.149	0.097	0.802	-0.409	2.25	0.149	0.148	0.946
		β_{21} (z_1)	0.5	0.499	0.20	0.198	0.114	0.738	0.499	0.20	0.198	0.192	0.941
		β_{31} (z_2)	-0.5	-0.506	1.20	0.068	0.040	0.750	-0.506	1.20	0.068	0.067	0.949
	2	β_{02}(Intercept)	-0.4	-0.401	0.25	0.142	0.075	0.702	-0.401	0.25	0.142	0.139	0.946
		β_{12} (t)	-0.4	-0.406	1.50	0.131	0.075	0.740	-0.406	1.50	0.131	0.129	0.943
		β_{22} (z_1)	0.5	0.509	1.80	0.192	0.109	0.732	0.509	1.80	0.192	0.185	0.945
		β_{32} (z_2)	-0.5	-0.503	0.60	0.067	0.038	0.719	-0.503	0.60	0.067	0.065	0.944

Table 2.4: Model results from using the pseudo-likelihood method when data are clustered. Cluster size was set at 50, double sampling percent was set at 20%.

		Verifiable Outcome	
Variables	Total, N=31179 [%]	No, N=24222 [%]	Yes, N=6957 [%]
Independent Variables			
Age at ART initiation			
Mean (SD)	37.4 (9.7)	37.6 (9.3)	38.4 (10.0)
Median (min - max)	36.2 (18.0 - 90.1)	36.4 (18.2 - 82.2)	37.1 (18.1 - 81.4)
Gender			
Female	19961 [64]	15958 [66]	4003 [58]
Male	11218 [36]	8264 [34]	2954 [42]
Study time in months			
Mean (SD)	14.3 (15.3)	15.1 (15.5)	11.5 (14.3)
Median (min - max)	8.4 (0 - 108.2)	9.5 (0.2 - 104.9)	5.4 (0.0 - 108.2)
CD4 count at ART initiation			
Mean (SD)	188.8 (174.6)	194.3 (175.2)	169.4 (171.2)
Median (min - max)	155 (0.0 - 3030.0)	163.0 (0.0 - 2869.0)	131.0 (0.0 - 3030.0)
Outcome Variables			
Observed Cause Of Failure			
Death	2719 [8.7]	0 [0.0]	2719 [39]
Loss to Clinic	28460 [91]	24222 [100]	4238 [61]
Confirmed Cause of Failure			
Death	3862 [12]	0 [0.0]	3862 [56]
Loss to Clinic	3095 [9.9]	0 [0.0]	3095 [44]
None (Outcome not validated)	24222 [78]	24222 [100]	0 [0.0]

Table 2.5: Characteristics of patients involved in the missclassification model of the probability of classifying patients as disengaged from care when they are in fact dead. All the patients came from the AMPATH program.

	Complete Case Analysis, N=3862				**Pseudo-likelihood Method, N=28084**			
	Estimate	SE	Z	Pr(> $\|Z\|$)	Estimate	SE	Z	Pr(> $\|Z\|$)
(Intercept)	-1.075	0.0870	-12.363	0.0000	0.656	0.0743	8.838	0.0000
Gender (Male versus Female)	-0.113	0.0724	-1.555	0.1200	-0.208	0.0635	-3.273	0.0011
Centered Age (Age minus mean of age)	0.011	0.0035	3.097	0.0020	0.006	0.0030	1.886	0.0594
$\sqrt{\text{CD4 Count}}$	0.012	0.0061	1.965	0.0495	0.016	0.0059	2.653	0.0080
Study time (months)	0.025	0.0115	2.161	0.0307	0.058	0.0087	6.629	0.0000
$I(3 \leq$ Study time $< 6)\times$ (Study time - 3)	0.016	0.0601	0.269	0.7876	-0.031	0.0358	-0.868	0.3856
$I($ $6 \leq$ Study time $< 12)\times$(Study time - 6)	-0.028	0.0379	-0.743	0.4574	-0.053	0.0232	-2.299	0.0215
I(Study time $\geq 12)\times$(Study time - 12)	-0.027	0.0155	-1.711	0.0871	-0.068	0.0107	-6.370	0.0000

Table 2.6: Misclassification model when using complete-case analysis, and the proposed pseudo-likelihood method. Complete case analysis consisted of 3862 subjects, and the pseudo-likelihood based analysis consisted of 28084 subjects, where 3862 received weight of 1, and the rest received a weight between 0 and 1.

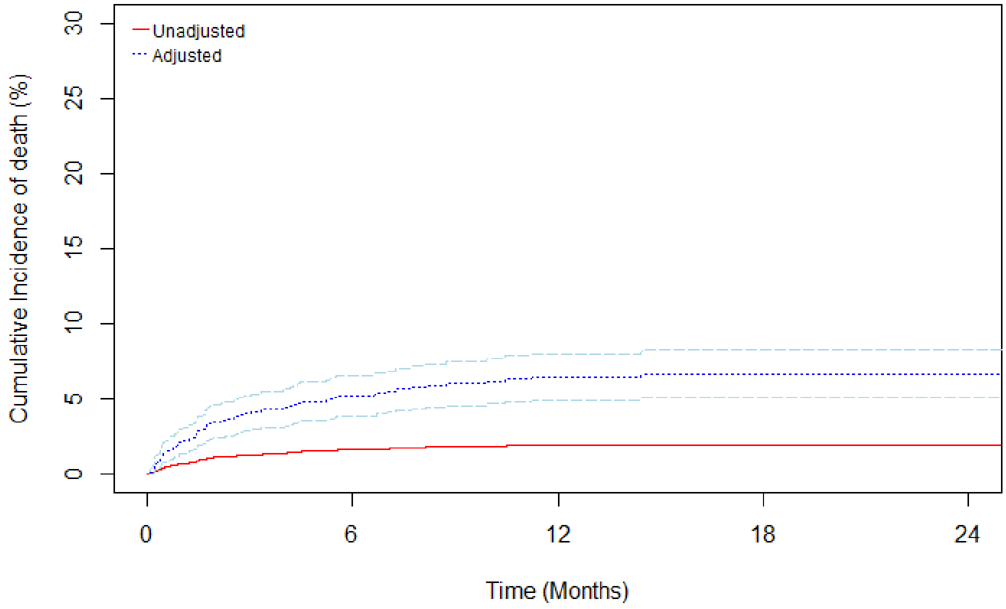

Figure 2.3: Naive (unadjusted) and misclassification-adjusted cumulative incidence functions of death at FACES. The light-blue dashed lines represent the point-wise 95% confidence-interval limits for the misclassification-adjusted CIF.

CHAPTER 3

Modeling Cause-specific Hazards While Adjusting for Externally-sourced Outcome-misclassification Information

In competing risks models, misclassified outcomes often lead to biased estimation of cause-specific hazards. Properly accounting for outcome misclassification, therefore, becomes a critical issue in the analysis. The accommodation, however, depends on the availability of the misclassification probabilities, which can be estimated either from internal or external sources. In real scientific investigations, misclassification errors can rarely be quantified from within the study; analysts, therefore, rely on estimates ascertained from external sources. In this chapter, I describe a parametric pseudo-likelihood method for estimating cause-specific hazards, under the assumption that the misclassification probabilities obtained from external sources are transferrable. I show that under such an assumption, the resulting pseudo-likelihood estimator remains consistent and asymptotically normal. I also show that the variance of the pseudo-likelihood estimator has a closed-form expression that accounts for different sources of variability. I assess the finite-sample properties of the estimator through a simulation study. To illustrate the use of the proposed method, I analyzed data generated by a real clinical investigation.

3.1 Introduction

Estimation of cause-specific hazard functions in competing-risk models requires accurate assessments of the causes of failures. For example, in analyses of survival outcomes, causes of death must be ascertained accurately to ensure the validity of analytical results. Direct and verifiable assessment of failure causes, i.e., the gold-standard cause ascertainment, is often cost inhibitive and thus is rarely feasible in large scale scientific investigations. In some instances, alternative data sources are available for determination of the causes. For example, in studies of HIV/AIDS, causes of death could be ascertained from patient medical records in lieu of autopsy reports. Such external sources, if not treated with care, could introduce errors to the cause determination. In the case of cause of death, medical records are often subject to physician errors and coding inaccuracies, and thus giving rise to misclassification (Flanders, 1992). In a competing-risk analysis, misclassified causes tend to undermine the validity of the model parameter estimation and inference.

One way to alleviate the impact of incorrectly determined failure causes is to adjust for the probabilities of misclassification. For all practical purposes, these probabilities are unlikely to be available *a priori*; estimates can be obtained by evaluating the true failure causes in a validation sample and then extrapolate to the full sample (Spiegelman, Carroll, and Kipnis 2001). In general, an internal validation based on a sub-sample is preferred as it is more likely to be representative of the original study population. Estimating the misclassification probabilities from external sources is only used when internal validation is logistically or operationally unfeasible. But in many situations, external validation does have an advantage of reduced validation cost. For example, for HIV-related studies conducted in resource-limited countries, determining the causes of death from existing external data is of considerable appeal.

Several authors have studied the impact of outcome misclassification in competing-risk models, and proposed methods for outcome misclassification adjustments. Among these, Ebrahimi (1996) proposed a Bayesian approach for accommodating outcome misclassification in competing-risks estimation (Ebrahimi 1996). Van Rompaye et al. (2012) proposed a semi-parametric method for estimating the cause-specific hazards in the presence of failure cause misclassification (Van Rompaye, Jaffar, and Goetghebeur 2012).Gravel et al. (2018) proposed a parametric full-likelihood approach that uses internal-validation as a source of misclassification information(Gravel et al. 2018). Bakoyannis and Yiannoutsos (2015) described a nonparametric method for modeling the cumulative incidence functions while accounting for non-differential outcome misclassification (Bakoyannis and Yiannoutsos 2015). Bakoyannis et al. (2018) proposed a semiparametric approach for modeling cause-specific hazards when the misclassification information came from an internal validation sample (Bakoyannis, Zhang, and Yiannoutsos 2018). Most recently, Edwards et al. (2019) extended the work of Bakoyannis and Yiannoutsos (2015) to cater to a scenario where misclassification rates differ among the subjects in a study (differential misclassification).

In this research, I present a parametric method for modeling cause-specific hazards in the presence of outcome misclassification, and I focus on the situation where the misclassification information used for estimation adjustment comes from an external source. In the proposed method, I first estimate the misclassification probabilities from external data, and then incorporate the estimates into the true likelihood. In essence, the method amounts to a pseudo-likelihood estimation. I show that under appropriate conditions, the final estimator is consistent and asymptotically normally distributed. For inference, I also derive a closed-form variance estimator that accounts for various sources of uncertainty, including the data-generating process, cause-specific hazard parameter estimation, and the estimation of misclassification probabilities based on external data.

In the following sections, I present the method in the context of a real clinical investigation.

3.2 Motivation: Real-world example

Among HIV clinics that contribute data to the International Epidemiologic Databases for the Evaluation of HIV/AIDS (IeDEA East Africa), one prominent research question is that of identifying the factors that influence the hazards of death and disengagement from care. Answering this question in a statistical fashion entails performing a competing-risks analysis by modeling the cause-specific hazards of death and disengagement from care. Outcomes are treated as competing risks since interest lies in whichever of the two outcomes comes first: Observing death precludes us from observing disengagement from care; the opposite is also true. For IeDEA, the process of using competing risks methods is complicated by the potential for misclassification among patients deemed disengaged from care: Some of these patients may actually be dead. As a result, the extent of death misclassification should be estimated before performing a competing-risks analysis.

IeDEA East Africa estimates the level of misclassification by re-ascertaining, through patient outreach, the outcomes for a sample of those deemed disengaged from care (Elvin H Geng et al. 2011; E. H. Geng et al. 2008; Egger et al. 2012). This outreach is sometimes referred to as *internal validation/double-sampling*. Using data from outreach, IeDEA can infer the death misclassification probabilities in the cohort from which the validation sample is drawn. These misclassification probabilities are then used to adjust competing-risks analyses of interest. That said, the aforementioned internal-validation scheme is not possible to perform at a large scale due the prohibitive cost of outreaching patients. As a result, not all treatment programs that contribute data to IeDEA can perform patient-outreach. In order to perform competing risks analyses that are adjusted for misclassification in settings without outcome validation, one way to proceed is to rely on misclassification information from settings that have outcome validation. This transfer of information is done under the assumption of “transportability” of misclassification probabilities across geographical areas (Justice, Covinsky, and Berlin 1999).

In this chapter, I will present an application using data from two treatment programs that contribute data to IeDEA. The first dataset will come from AMPATH(Academic Model Providing Access to Healthcare), and the second dataset will come from FACES(Family AIDS Care and Education Services). The main difference between these two datasets is that the AMPATH data has an internal-validation sample, whereas the FACES data does not. Using these datasets, we will illustrate how the death misclassification probabilities estimated in AMPATH can be used to adjust the estimation of cause-specific hazards of death and disengagement from care in FACES. A scheme of the analysis is presented in Figure 3.1.

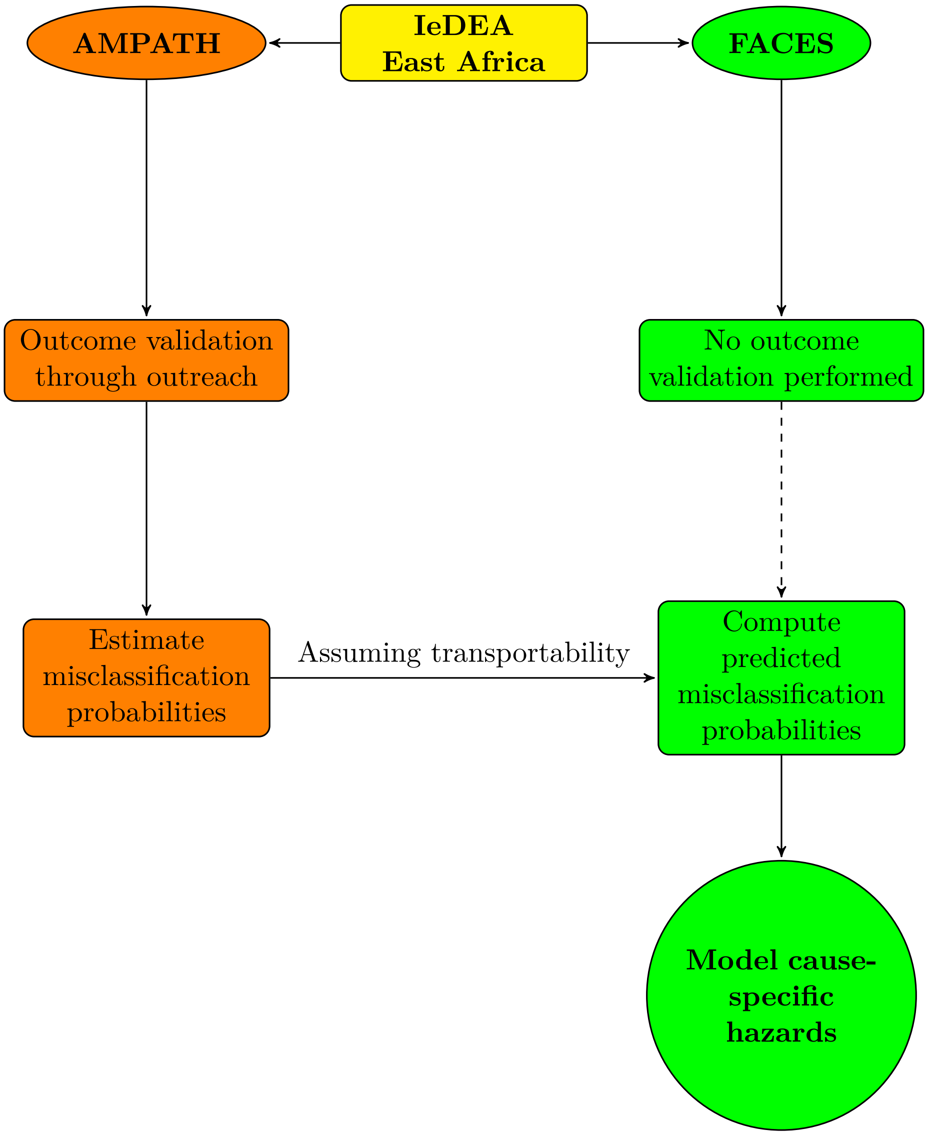
AMPATH
IeDEA
East Africa
FACES
Outcome validation through outreach
No outcome validation performed
Estimate misclassification probabilities
Assuming transportability
Compute predicted misclassification probabilities
Model cause-specific hazards

Figure 3.1: Scheme for modeling cause-specific hazards of death and disengagement from care at FACES while adjusting for death misclassification information that was derived from AMPATH.

3.3 Methods

3.3.1 Notation

Assume we have a m-cause competing-risks system wherein a subject can fail from any one of m causes. Let the true cause of failure be represented by $C \in \{1, 2, ..., m\}$. Given the potential for outcome misclassification, we may observe $C^* \in \{1, 2, ..., m\}$, where C^* may not be consistent with C. Let U be the failure time; V be the censoring time, and T be the right-censored failure time, where $T = \min(U, V)$. We will assume that U and V are independent. In addition, we will assume that censoring distribution is independent of the cause of failure. Lastly, let $\boldsymbol{Z}$ be the subject characteristics. For each of the n subjects, $i = 1, 2..., n$, were will observe $(T_i, C_i^*, \boldsymbol{Z}_i)$.

Outcomes that are ascertained using gold-standard approaches will be referred to as **true** outcomes, and those that are ascertained using error-prone approaches are referred to as **observed/misclassified** outcomes.

3.3.2 Likelihood

When there is no misclassification among the competing outcomes, the observed cause C^* is consistent with the underlying true cause C. This means for each subject i, we observe $(T_i = t_i, C_i^*, \boldsymbol{Z}_i) = (T_i = t_i, C_i, \boldsymbol{Z}_i)$. For such a scenario, the likelihood is a function of the true cause-specific hazards as shown by Equation 3.1.

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{m} \prod_{i=1}^{n} \lambda_j(t_i; \boldsymbol{\theta}_j, \boldsymbol{Z}_i)^{\delta_{ij}} \exp\left[- \int_0^{t_i} \lambda_j(u; \boldsymbol{\theta}_j, \boldsymbol{Z}_i) du\right] \tag{3.1}$$

where:

1. $\lambda_j(t; \boldsymbol{Z})$ is the true cause-specific hazard of cause j at time t conditional on covariates $\boldsymbol{Z}$, for $j \in \{1, 2, ..., m\}$;
2. $\delta_{ij} = I[C_i = j]$ is the event indicator of cause-j for subject i;
3. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, captures the association between the risk factors and the cause-specific hazards.

When there is a possibility of outcome misclassification, $(T_i = t_i, C_i^*, \boldsymbol{Z}_i)$ is not necessarily the same as $(T_i = t_i, C_i, \boldsymbol{Z}_i)$, on account of the fact that C and C^* are not necessarily the same. As a result, the likelihood with respect to the observed data changes to the form shown in Equation 3.2.

$$L(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \prod_{j=1}^{m} \prod_{i=1}^{n} \lambda_j^*(x_i; \boldsymbol{\theta}_j^*, \boldsymbol{Z}_i)^{\delta_{ij}^*} \exp\left[-\int_0^{x_i} \lambda_j(u; \boldsymbol{\theta}_j, \boldsymbol{Z}_i) du\right] \tag{3.2}$$

where:

1. $\lambda_j^*(t; \boldsymbol{Z})$ is the cause-specific hazard of <u>observed</u> cause j at time t conditional on covariates $\boldsymbol{Z}$, for $j \in \{1, 2, ..., m\}$;
2. $\delta_{ij}^* = I[C_i^* = j]$ is the event indicator of observed cause-j for subject i.

The cause-specific hazard of observed cause j at time t, $\lambda_j^*(t; \boldsymbol{Z})$, is not necessarily the same as the cause-specific hazard of interest (true cause-specific hazards), $\lambda_j(t; \boldsymbol{Z})$. In fact,

$$\lambda_j^*(t; \boldsymbol{\theta}^*, \boldsymbol{Z}) = \sum_{k=1}^{m} \lambda_k(t; \boldsymbol{\theta}_k, \boldsymbol{Z}) P\left(C^* = j | T = t, C = k, \boldsymbol{Z}, \boldsymbol{\beta}_k\right).$$

In other words, $\lambda_j^*(t)$ is a linear combination of the true cause-specific hazards, with the weights being functions of misclassification probabilities. The likelihood with respect to possibly misclassified outcomes can, therefore, be written as follows:

$$L(\boldsymbol{\theta},\boldsymbol{\beta}) = \prod_{j=1}^{m}\prod_{i=1}^{n}\left[\sum_{k=1}^{m}\lambda_k(t_i;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)\pi^*_{jk}(\boldsymbol{\beta}_k;\boldsymbol{Z}_i)\right]^{\delta^*_{ij}} \times \exp\left[-\int_0^{t_i}\lambda_j(u;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)du\right] \tag{3.3}$$

where $\pi^*_{jk}(\boldsymbol{\beta}_k;\boldsymbol{Z}) = P\left(C^* = j|T = t, C = k, \boldsymbol{Z}, \boldsymbol{\beta}_k\right)$ is the probability of observing cause j when the true cause of failure is k, conditional on $(\boldsymbol{Z}, T = t)$. Loosely, these probabilities which include *misclassification probabilities, sensitivities and specificities* will be referred to as misclassification probabilities.

In order to proceed with estimation using the likelihood in (3.3), we need to either have prior knowledge about π^*_{jk} or estimate π^*_{jk} using available data. When an internal-validation sample is available, π^*_{jk} can be estimated efficiently using a pseudo-likelihood approach presented in Mpofu et al. (2019). That being said, when an internal-validation sample is available, we can forgo the estimation of $\pi^*_{jk}(\boldsymbol{\beta}_k;\boldsymbol{Z})$. This is because when an internal validation sample is available, there exist other likelihood formulations (other than likelihood 3.3) that will result in more efficient estimates of cause-specific hazard parameters. One such formulation is presented by Bakoyannis et al. (2018) (Bakoyannis, Zhang, and Yiannoutsos 2018). When an internal-validation sample is available, Bakoyannis et al. (2018) proposed that the likelihood should be expressed as follows:

$$\begin{aligned} L(\boldsymbol{\theta},\boldsymbol{\eta}) = \prod_{j=1}^{m}\prod_{i=1}^{n}&\lambda_j(x_i;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)^{\left[R_i\times\delta_{ij}+\sum_{k=1}^{m}\delta^*_{ik}\times(1-R_i)\times p_{jk}(\boldsymbol{\eta}_k;\boldsymbol{Z}_i)\right]}\\ &\times\exp\left[-\int_0^{x_i}\lambda_j(u;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)du\right] \end{aligned} \tag{3.4}$$

where,

1. $R_i = 1$ indicates that the outcome for subject i was validated (thereby known);
2. $\delta_{ij} = I[C_i = j]$ is the event indicator of true cause-j for subject i;

3. $\delta^*_{ik} = I[C^*_i = k]$ is the event indicator of observed cause-k for subject i;
4. $p_{jk}(\boldsymbol{\eta}_k; \boldsymbol{Z}_i) = P[C_i = j | C^*_i = k, T_i = t_i, \boldsymbol{Z}_i, \boldsymbol{\eta}_k]$, is the probability that the true cause of failure is cause-j given the observed cause is cause-k(i.e., predictive values), given $(\boldsymbol{Z}, T = t)$.

In this likelihood formulation (3.3.2), Bakoyannis et al. (2018) treated unvalidated outcomes as missing values, and replaced the missing values by their predictive values.

When there is no outcome validation, neither likelihood (3.3) nor (3.3.2) is immediately applicable. The reason for this is that neither the misclassification probabilities required for likelihood (3.3) nor the predictive values required for likelihood (3.3.2) are identifiable from the sample at hand. In fact, it would be imprudent to use the method by Bakoyannis et al. (2018) since the unavailability of validation sample also means 100% missingness of the true cause of failure variable (C). That being said, we can still use likelihood (3.3) in estimation if we can borrow outcome-misclassification information from settings that have internal-validation sampling. This transfer of information is made assuming that misclassification probabilities are *transportable* across different populations. The same assumption cannot be made for predictive values, therefore likelihood (3.3.2) cannot be used in settings that do not have internal validation.

3.3.3 Transportability of misclassification probabilities

Misclassification probabilities in settings without outcome validation can be estimated by appealing to the transportability of misclassification probabilities (Lyles et al. 2011). That is, we assume that the misclassification probability models in the current study are the same as those from an external study (R. J. Carroll et al. 2006; Spiegelman 2010). According to Spiegelman (2010), a downside to the transportability assumption is that it is not empirically verifiable (Spiegelman 2010). Spiegelman also notes that the similarity in covariate distributions across different settings can support the credibility of the transportability assumption, however, it does not guarantee that the assumption truly holds.

In our motivating example that consists of East-Africa IeDEA HIV-treatment programs, the borrowing of misclassification information across different settings has both plausibility and statistical justifications. Plausibility is driven by fact that IeDEA sites are geographically proximal, and typically collect data on the same covariates. Statistically, the transfer of misclassification probabilities is favorable because misclassification probabilities are independent of the underlying prevalences of the causes of failure. The latter justification is in the same spirit as in medical diagnostic tests wherein the sensitivities and specificities of tests are invariant across settings with different disease prevalences. The same justification, however, cannot be used for predictive values as they are dependent on the underlying prevalence.

3.4 Theory

I will begin by clarifying the statistical model. Let $\zeta = (\boldsymbol{\theta}, \boldsymbol{\beta}) \in R^{d_1+d_2}$ be the full-parameter space. In the problem, $\boldsymbol{\theta}$ defines the cause-specific hazard model and is the parameter of interest(i.e, the structural parameter), and $\boldsymbol{\beta}$ defines the misclassification model and is the nuisance parameter. In addition, let's define the random variable $\boldsymbol{Y} = (T, C^*, \boldsymbol{Z})$. Assuming that probability density of $\boldsymbol{Y}$ is $p(\boldsymbol{y}; \zeta)$, the statistical model is family of densities, $\mathcal{P}_\zeta = \{p(\boldsymbol{y}; \zeta) : \zeta \in \boldsymbol{\zeta} \subset R^{d_1+d_2}, d_1, \, d_2 \in \mathbb{N}\}$. Given independent and identically distributed realizations, $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n$ from density $g(\boldsymbol{y})$, in general, the estimation goal is find $\zeta_0 \in \boldsymbol{\zeta}$ such that $p(\boldsymbol{y}; \zeta_0) = g(\boldsymbol{y})$ (Grace 2016).

3.4.1 Pseudo-likelihood estimation

In our estimation problem, the parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_m)$, with $\boldsymbol{\theta}_k$ being the parameter associated with cause-k. The misclassification parameter $\boldsymbol{\beta}$ in the likelihood is replaced by its estimate $\hat{\boldsymbol{\beta}}$, thereby reducing the parameter-space to R^{d_1}. Consequently, the goal of maximum pseudo-likelihood estimation is to find $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^{d_1}$ that maximizes the log-pseudo-likelihood, that is:

$$\dot{l}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}_{n_e}) = \sum_{i=1}^{n} \dot{l}_i(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) = 0 \tag{3.5}$$

where $\hat{\boldsymbol{\beta}}_{n_e} \in \mathbb{R}^{d_2}$ is misclassification parameter estimate borrowed from an external setting, that has sample size n_e.

3.4.2 Asymptotic properties

The asymptotic properties of the maximum pseudo-likelihood estimator were studied under mild regularity conditions, similar to those in Gong and Samniego (Gong and Samaniego 1981). The conditions were as follows:

1. $p(\boldsymbol{y}; \zeta) \neq p(\boldsymbol{y}; \zeta^*) \implies \zeta \neq \zeta^*$
2. The support, $S = \{\boldsymbol{y} : p(\boldsymbol{y}; \zeta) > 0\}$ does not depend on $\zeta = (\boldsymbol{\theta}, \boldsymbol{\beta})$.
3. ζ is an interior point in $\boldsymbol{\zeta}$.
4. For all $\boldsymbol{y}$, $(\boldsymbol{\theta}, \boldsymbol{\beta})$, the partial derivatives $\dot{l}_{\boldsymbol{\theta}}$, $\ddot{l}_{\boldsymbol{\theta\theta}}$, $\dddot{l}_{\boldsymbol{\theta\theta\theta}}$, $\dot{l}_{\boldsymbol{\beta}}$,$\ddot{l}_{\boldsymbol{\theta\beta}}$, $\dddot{l}_{\boldsymbol{\theta\theta\beta}}$, $\dddot{l}_{\boldsymbol{\theta\beta\beta}}$ exist.

5. The third partial derivatives are bounded by an integrable function, that is

$$|\dddot{l}_{\boldsymbol{\theta\theta\theta}};\ \dddot{l}_{\boldsymbol{\theta\theta\beta}};\ \dddot{l}_{\boldsymbol{\theta\beta\beta}}| \leq M(\boldsymbol{y})$$

for all $(\boldsymbol{\theta}, \boldsymbol{\beta})$ and $\boldsymbol{y}$, where $M(\boldsymbol{y})$ is an integrable function.

6. For all $(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \boldsymbol{\zeta}$ and for all $\boldsymbol{y}$,
$$\left| \frac{d}{d\boldsymbol{\beta}} \log \frac{p(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\beta})}{p(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\beta}_0)} \right| \leq M(\boldsymbol{y}, \boldsymbol{\theta}),$$
where $E[M(\boldsymbol{y}, \boldsymbol{\theta})] < \infty$, $\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $M(\boldsymbol{y})$ is an integrable function.

7. The estimator of the misclassification parameter is consistent, that is, $\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| \xrightarrow{p} 0$.

Theorem 1: Under conditions 1-4 and 7, the maximum pseudo-likelihood estimator $\boldsymbol{\theta}_n$ is consistent, that is,

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \xrightarrow{p} 0$$

The proof of consistency is presented in Section 3.10.2.

Theorem 2: Under conditions 4 through 7, the maximum pseudo-likelihood estimator is asympotically normal, that is:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{\Omega}\right).$$

where,

a. $\boldsymbol{\Omega} = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + q.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$,

b. $\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is a $d_1 \times d_2$ matrix $((d_1 \times d_1) \times (d_1 \times d_2))$,

c. $q = \frac{n}{n_e}$ as $n \to \infty$ is the limiting ratio of the current study sample size (n) and the external study sample size(n_e).

This derivation is consistent with one from Ogden and Tarpey (2006) (Ogden and Tarpey 2006). The proof for asympotic normality is presented in Section 3.10.3.

Remark 1: With the sample at hand, $\Omega = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + q.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$, can be estimated by replacing the parameters with their consistent estimators. That is:

$$\hat{\Omega} = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n; \hat{\boldsymbol{\beta}}_{n_e}) + q.\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_{n_e})\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})^T.$$

Empirically, the variance of maximum pseudo-likelihood estimate, $\hat{\boldsymbol{\theta}}_n$, can be estimated as follows:

$$\begin{aligned}\hat{V}(\hat{\boldsymbol{\theta}}_n) &= \frac{1}{n}\left[\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) + q.\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_{n_e})\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})^T\right] \\ &= \frac{1}{n}.\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) + \frac{q}{n}.\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_{n_e})\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})^T \\ &= \frac{1}{n}.\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) + \frac{1}{n_e}.\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_{n_e})\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})^T.\end{aligned}$$

Remark 2: The variance estimator accounts for the different sources, namely: the variance due to estimating $\boldsymbol{\theta}_0$ as captured by $\frac{1}{n}.\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})$, and the variance due to estimating $\boldsymbol{\beta}_0$ in a external setting as captured by $\frac{1}{n_e}.\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}_{n_e})\mathbf{W}(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})^T$.

Remark 3: In a situation when an internal-validation sample is available, the asympototic variance in Theorem 2 changes to: $\boldsymbol{\Omega} = E[\tilde{\Psi}_i(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\tilde{\Psi}_i(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T]$ where,

$$\tilde{\Psi}_i(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0) + \mathbf{W}(\boldsymbol{\theta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\beta}_0)$$

The new variance estimator, $\hat{\boldsymbol{\Omega}}$, is formed by replacing the parameters $(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)$ with their consistent estimators $(\hat{\boldsymbol{\theta}}_n; \hat{\boldsymbol{\beta}}_n)$. The estimators have the same sample size index, n, since the cause-specific hazard and misclassification parameters are estimated using exactly the same sample. Empirically, $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\Psi}_i(\hat{\boldsymbol{\theta}}_n; \hat{\boldsymbol{\beta}}_n)\tilde{\Psi}_i(\hat{\boldsymbol{\theta}}_n; \hat{\boldsymbol{\beta}}_n)^T$.

Remark 4: If the misclassification-parameter estimator, $\hat{\boldsymbol{\beta}}_{n_e}$, is also a pseudo-likelihood estimator as in Mpofu et al. (2019), the variance estimator of $\boldsymbol{\Omega}$, can be partitioned into components that clearly capture the sources of variance when one creates a pseudo-likelihood estimator using another pseudo-likelihood estimator as a plug-in.

In Mpofu et al. (2019), it was shown that, in the presence of an internal-validation sample, the misclassification parameter, $\boldsymbol{\beta}$, could be estimated efficiently using a pseudo-likehood approach that uses as a plug-in, the estimate of the predictive-value parameter, $\boldsymbol{\gamma}$. In general, asymptotic variance of the misclassfication parameter $\hat{\boldsymbol{\beta}}_{n_e}$ can be partitioned as follows:

$$\boldsymbol{\Sigma}_{borrowed} = \boldsymbol{\Sigma}_1 + s.\boldsymbol{\Sigma}_2$$

where $\mathbf{\Sigma}_1$ is the variance associated with estimating the misclassification paramter when there is 100% double-sampling, $\mathbf{\Sigma}_2$ is the variance associated with estimating the predictive value parameter using the internal-validation sample, and $s = \frac{n_e}{n_v}$ is ratio of the main-study sample size, n_e, and the internal-validation sample size, n_v, as $n_e \to \infty$.

$\mathbf{\Sigma}_{borrowed}$ represents $\mathbf{I}^{-1}(\boldsymbol{\beta}_0)$ in

$$\mathbf{\Omega} = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + q.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$$

It then follows that:

$$\mathbf{\Omega} = \mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + q.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{\Sigma}_1\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T + c.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{\Sigma}_2\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$$

where:

i. $\mathbf{I}^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)$ is variance associated with cause-specific hazards parameter estimation in the current study assuming the misclassification parameter $\boldsymbol{\beta}$ is known.

ii. $q.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{\Sigma}_1\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$ is the additional variance associated with borrowing an misclassification parameter from an external study with 100% double-sampling among the non-censored (that is, an external study where all event data have been validated).

iii. $c.\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\mathbf{\Sigma}_2\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)^T$ is the additional variance due to using a pseudo-likelihood approach to estimate the misclassification parameter in an external setting.

iv. $c = \frac{n}{n_v}$ is the ratio of current study sample size n and the size of the internal-validaton sample in an external study as $n \to \infty$.

The estimator, $\hat{\boldsymbol{\theta}}_n$, depends on an external misclassification estimator, $\hat{\boldsymbol{\beta}}_{n_e}$. When an internal-validation sample is available in an external setting, the misclassification parameters can be estimated using a pseudo-likelihood approach as shown in Mpofu et al. (2019). The asymptotic variance $\boldsymbol{\Omega}$ can be partitioned into three components that capture the process within which the pseudo-likelihood estimator $\hat{\boldsymbol{\theta}}_n$ was created

3.5 Example: Parametric estimation under a two-cause system

Without losing generality, I will focus on a competing risks system with two causes of failure. Let the baseline cause-specific hazard for cause $j \in \{1, 2\}$ be $\lambda_{0j}(t; \boldsymbol{\phi}_j)$.

Assuming proportional hazards, the cause-specific hazard for cause $j \in \{1, 2\}$ is

$$\lambda_j(t; \boldsymbol{\theta}_j, \boldsymbol{\phi}_j) = \lambda_{0j}(t; \boldsymbol{\phi}_j) \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_j\right), \text{ where}$$

$$\lambda_{0j}(t; \boldsymbol{\phi}_j) = \begin{cases} \phi_j : \phi_j \in \mathbb{R}, \, \phi_j > 0 & \text{exponential shape,} \\ \\ \phi_j t^{\phi_j - 1} : \phi_j \in \mathbb{R}, \, \phi_j > 0 & \text{Weibull shape,} \\ \\ g(t; \boldsymbol{\phi}_j : \boldsymbol{\phi}_j \in \mathbb{R}^{l_+}, l_+ \in \mathbb{N}, l_+ \geq 2) & \text{general parametric shape} \end{cases}$$

- $\boldsymbol{Z}$ is design matrix consisting of the risk factors (covariates);
- $\boldsymbol{\theta}_j$ for $j \in \{1, 2\}$ captures the association between the risk factors and the cause-specific hazards.

Under any of the parametric cause-specific hazard formulations given above, the pseudo-log-likelihood is:

$$\begin{aligned} l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2) &= \sum_{i=1}^{n} \left\{ \delta_{i1}^{*} \log \left[\sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i) \right] - \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_1\right) \int_0^{t_i} \lambda_{01}(u; \boldsymbol{\phi}_1) du \right\} \\ &+ \sum_{i=1}^{n} \left\{ \delta_{i2}^{*} \log \left[\sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i) \right] - \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_2\right) \int_0^{t_i} \lambda_{02}(u; \boldsymbol{\phi}_2) du \right\} \\ &= \sum_{i=1}^{n} \left\{ \delta_{i1}^{*} \log \left[\sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i) \right] - \Lambda_{01}(t_i; \boldsymbol{\phi}_1) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_1\right) \right\} \\ &+ \sum_{i=1}^{n} \left\{ \delta_{i2}^{*} \log \left[\sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i) \right] - \Lambda_{02}(t_i; \boldsymbol{\phi}_2) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_2\right) \right\} \end{aligned}$$

where $(\boldsymbol{\theta}_j, \boldsymbol{\phi}_j)$ are the parameters associated with cause j, for $j \in \{1, 2\}$.

The score function is given by Equation 3.5.

$$\nabla l(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = \begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^{*} \pi_{11}^{*}(\hat{\beta}_1; \boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*} \pi_{21}^{*}(\hat{\beta}_1; \boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) . \lambda_{01}(t_i; \boldsymbol{\phi}_1) - \Lambda_{01}(t_i; \boldsymbol{\phi}_1) \right\} \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_1\right) \\ \sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^{*} \pi_{12}^{*}(\hat{\beta}_2; \boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*} \pi_{22}^{*}(\hat{\beta}_2; \boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) \lambda_{02}(t_i; \boldsymbol{\phi}_2) - \Lambda_{02}(t_i; \boldsymbol{\phi}_2) \right\} \exp\left(\boldsymbol{Z}_i \theta_2\right) \\ \sum_{i=1}^{n} \left\{ \left(\frac{\delta_{i1}^{*} \pi_{11}^{*}(\hat{\beta}_1; \boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*} \pi_{21}^{*}(\hat{\beta}_1; \boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) . \frac{d\lambda_{01}(t_i; \boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1} - \frac{d\Lambda_{01}(t_i; \boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1} \right\} \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_1\right) \\ \sum_{i=1}^{n} \left\{ \left(\frac{\delta_{i1}^{*} \pi_{12}^{*}(\hat{\beta}_2; \boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*} \pi_{22}^{*}(\hat{\beta}_2; \boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) . \frac{d\lambda_{02}(t_i; \boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2} - \frac{d\Lambda_{02}(t_i; \boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2} \right\} \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_2\right) \end{bmatrix}$$

where,

1. $p_1^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i) = \sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$;
2. $p_2^{*}(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i) = \sum_{k=1}^{2} \lambda_{0k}(t_i; \boldsymbol{\phi}_k) \exp\left(\boldsymbol{Z}_i \boldsymbol{\theta}_k\right) \pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$.

Based on the Equation 3.5, it is clear that part of task when setting up the score function is computing the gradients of baseline cause-specific hazards and baseline cumulative cause-specific hazards with respect to pertinent parameters.

3.5.1 Example: Exponential-shaped baseline cause-specific hazards

Assume that $\lambda_1(t; \boldsymbol{Z}) = \exp(\boldsymbol{Z}\boldsymbol{\theta}_1)$, and $\lambda_2(t; \boldsymbol{Z}) = \exp(\boldsymbol{Z}\boldsymbol{\theta}_2)$,

where $\boldsymbol{Z} = \left[\boldsymbol{Z}_1^T, \boldsymbol{Z}_2^T, ..., \boldsymbol{Z}_n^T\right]$ is an $n \times d_1$ matrix, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are $d_1 \times 1$ matrices.

The respective baseline and cumulative cause-specific hazards for cause 1 and 2 are:

1. $\lambda_{01}(t; \alpha_1) = 1$, $\Lambda_{01}(t; \alpha_1) = t$;
2. $\lambda_{02}(t; \alpha_2) = 1$, $\Lambda_{01}(t; \alpha_2) = t$.

It then follows that:

1. $p_1^*(\boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i) = \sum_{k=1}^{2} \exp\left(\boldsymbol{Z}_i\boldsymbol{\theta}_k\right)\pi_{1k}^*(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$;
2. $p_2^*(\boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i) = \sum_{k=1}^{2} \exp\left(\boldsymbol{Z}_i\boldsymbol{\theta}_k\right)\pi_{2k}^*(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$.

Based on Equation 3.5, the score function is:

$$\nabla l = \begin{bmatrix} \sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^* \pi_{11}^*(\hat{\boldsymbol{\beta}}_1; \boldsymbol{Z}_i)}{p_1^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^* \pi_{21}^*(\hat{\boldsymbol{\beta}}_1; \boldsymbol{Z}_i)}{p_2^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) - t_i \right\} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_1\right) \\ \sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^* \pi_{12}^*(\hat{\boldsymbol{\beta}}_2; \boldsymbol{Z}_i)}{p_1^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} + \frac{\delta_{i2}^* \pi_{22}^*(\hat{\boldsymbol{\beta}}_2; \boldsymbol{Z}_i)}{p_2^*(\boldsymbol{\theta}, \boldsymbol{\phi}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}_i)} \right) - t_i \right\} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_2\right) \end{bmatrix}$$

3.5.2 Example: Weibull-shaped baseline cause-specific hazards

Assume that the respective cause-specific hazards for cause 1 and cause 2 are:

1. $\lambda_1(t; \boldsymbol{Z}) = \alpha_1 t^{\alpha_1 - 1} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_1\right)$
2. $\lambda_2(t; \boldsymbol{Z}) = \alpha_2 t^{\alpha_2 - 1} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_2\right)$.

In order to proceed with estimation, plug components a to j below into score Equation 3.5:

a) $\lambda_{01}(t;\alpha_1) = \alpha_1 t^{\alpha_1 - 1}$

b) $\lambda_{02}(t;\alpha_2) = \alpha_2 t^{\alpha_2 - 1}$

c) $\Lambda_{01}(t;\alpha_1) = t^{\alpha_1}$

d) $\Lambda_{02}(t;\alpha_2) = t^{\alpha_2}$

e) $\frac{d\lambda_{01}(t;\alpha_1)}{d\alpha_1} = t^{\alpha_1-1} + \alpha_1 t^{\alpha_1-1}\log t$

f) $\frac{d\lambda_{02}(t;\alpha_2)}{d\alpha_2} = t^{\alpha_2-1} + \alpha_2 t^{\alpha_2-1}\log t$

g) $\frac{d\Lambda_{01}(t;\alpha_1)}{d\alpha_1} = t^{\alpha_1}\log t$

h) $\frac{d\Lambda_{02}(t;\alpha_2)}{d\alpha_2} = t^{\alpha_2}\log t$

i) $p_1^*(\boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}) = \sum_{k=1}^{2} \alpha_k t^{\alpha_k - 1} \exp(\boldsymbol{Z}\theta_k)\pi_{1k}^*(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$

j) $p_2^*(\boldsymbol{\theta}, \boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}, \boldsymbol{Z}) = \sum_{k=1}^{2} \alpha_k t^{\alpha_k - 1} \exp(\boldsymbol{Z}\theta_k)\pi_{2k}^*(\hat{\boldsymbol{\beta}}_k; \boldsymbol{Z}_i)$.

3.6 Simulation Study

I studied the finite-sample properties of the pseudo-likelihood estimator using a simulation study. Each simulation iteration involved two datasets: one for the external study, and another one for the current study. The latter is the dataset of interest from which cause-specific hazards are modeled, and the former is used for estimating misclassification probabilities. That being said, the current and external-study datasets were simulated using the same procedure.

3.6.1 Generating the true cause of failure

I considered a study with two competing causes for failure, cause 1 and cause 2. Let $C \in \{1, 2\}$ represent the true cause of failure: The "true cause" of failure being one that is ascertained correctly. In the study, a subject/participant was followed until he/she failed from cause 1 or 2, or was censored. Letting, U represent the time-to-event, and V represent the censoring time, the time contributed to study by a subject was $T = \min(U, V)$. Censoring time was assumed to be independent of both the time-to-event and cause of failure. For each subject, $i = 1, 2, ..., n$, we observed $(T_i, C_i, \boldsymbol{Z}_i)$, with $C_i = 0$ indicating that subject i had been censored, and $\boldsymbol{Z}_i$ representing the covariates belonging to subject i.

Assuming the baseline cause-specific hazards were of the Weibull-form, and assuming proportional hazards, the cause-specific hazard for cause $k \in \{1, 2\}$ took the form:

$$\lambda_k(t|\boldsymbol{Z}) = \alpha_k t^{\alpha_k - 1} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_k\right),$$

where, $\alpha_k t^{\alpha_k - 1}$ is the baseline cause-specific hazard at time t associated with cause $k \in \{1, 2\}$; and, $\boldsymbol{\theta}_k$ captures the multiplicative dependence of cause-specific hazard with covariates $\mathbf{Z}$. In order to simplify the illustration, we assumed that cause-specific hazards for cause 1 and 2 depended on the same $\mathbf{Z}$.

Competing risks data were simulated using the method described by Beyersmann et al. (2009) (Beyersmann et al. 2009). For the two-cause system described above, this began with simulating the time-to-event by solving:

$$t^{\alpha_1} \exp\left(\mathbf{Z}\boldsymbol{\theta}_1\right) + t^{\alpha_2} \exp\left(\mathbf{Z}\boldsymbol{\theta}_2\right) + \log(1 - Q) = 0, \ t \geq 0$$

where $Q \sim U(0, 1)$.

For subject $i = 1, 2, ..., n$, the process of generating the time-to-event and the cause of failure proceeded as follows:

1. The time-to-event, u_i, was generated
2. For given u_i, the conditional probability of failing due to cause 1 at time $T = u_i$ was computed using the formula

$$P[C_i = 1|\mathbf{Z}_i, T = u_i] = \frac{\lambda_1(t = u_i|\mathbf{Z}_i)}{\lambda_1(t = u_i|\mathbf{Z}_i) + \lambda_2(t = u_i|\mathbf{Z}_i)} \tag{3.6}$$

3. A Bernoulli random variable, D_i, was generated with probability of success $P[C_i = 1|\mathbf{Z}_i, T = u_i])$. If $D_i = 1$, then true cause of failure was cause 1 ($C_i = 1$), otherwise the true cause of failure was 2 ($C_i = 2$).
4. The censoring time v_i was generated from $V_i \sim Exp(\eta)$.
5. The time contribution to the study was $t_i = min(u_i, v_i)$
6. If was $t_i = v_i$, then the subject was censored, that is, $C_i = 0$, otherwise $C_i = 1 \times (D_i = 1) + 2 \times (D_i = 2)$.

3.6.2 Generating the observed/misclassified cause of failure

Assuming that the outcome-detection method was subject to error, we observed C^*, where C^* was not necessarily the same as C. If $C^* \neq C$, the subject(study-unit) was said to be misclassified. In addition, those who were censored were assumed to be correctly classified, that is $C^* = 0 \iff C = 0$.

The observed/misclassified cause of failure was generated as follows:

a. Given $C_i \in 1, 2$, the probability of misclassification for subject i was defined as follows:

$$P[C_i^* = j|C_i = k, \mathbf{W}_i, \boldsymbol{\beta}_k] = \pi_{jk}(\boldsymbol{\beta}_k; \mathbf{W}_i) = \frac{\exp(\mathbf{W}_i\boldsymbol{\beta}_k)}{1 + \exp(\mathbf{W}_i\boldsymbol{\beta}_k)} \tag{3.7}$$

where $\mathbf{W}$ represented the covariates that were associated with misclassification.

b. The misclassification indicator was generated as follows:

$$M_i|C_i, \mathbf{W}_i \ \sim Ber\left[I(C_i = 1) \times \pi_{21}(\boldsymbol{\beta}_1; \mathbf{W}_i) + I(C_i = 2) \times \pi_{12}(\boldsymbol{\beta}_2; \mathbf{W}_i)\right]$$

where $M_i = 1$ indicated that outcome was misclassifieed (that is, the observed outcome was not the same as the true outcome).

c. The observed/misclassified cause of failure was generated as follows:

$$C_i^* = \begin{cases} C_i & \text{if } M_i = 0 \\ 1 \times (C_i = 2) + 2 \times (C_i = 1) & \text{if } M_i = 1. \end{cases}$$

3.6.3 Simulation dataset generation

In each simulation, two datasets were generated: one to act as an external-study dataset, and another to act as the current-study dataset. The datasets were generated as described in Subsections 3.6.1 and 3.6.2. For a fixed misclassification model, as defined by Equation 3.7, the datasets were generated based on the characteristics presented in Table 3.1.

		Study	
Cause		**External**	**Current**
	Data settings		
	Sample Size	5000	1000
	Double-sampling (%)	i) 20; ii) 50	NA
	Covariate 1, Z_1	$Z_1 \sim N(0,1)$	$Z_1 \sim N(2,1)$
	Covariate 2, Z_2	$Z_2 \sim Beta(1,1)$	$Z_2 \sim Beta(4,1)$
	Cause-specific hazard parameters		
1	α_1 (Shape)	2.00	1.50
1	$\theta_{12}\,(z_1)$	-0.60	-0.50
1	$\theta_{13}\,(z_2)$	0.80	0.80
2	α_2 (Shape)	2.00	2.50
2	$\theta_{22}\,(z_1)$	-0.50	-0.30
2	$\theta_{23}\,(z_2)$	0.70	0.60
	Censoring distribution	$V \sim Exp(0.75)$	$V \sim Exp(0.6)$
	Misclassification parameters		
1	β_{10} (Intercept)	$\log \frac{m_1}{1-m_1}$	$\log \frac{m_1}{1-m_1}$, $m_1 \in \{0.2, 0, 4\}$
1	$\beta_{11}\,(z_1)$	0.10	0.10
1	$\beta_{12}\,(z_2)$	0.70	0.70
1	$\beta_{13}\,(t)$	-1.00	-1.00
2	β_{20} (Intercept)	$\log \frac{m_2}{1-m_2}$	$\log \frac{m_2}{1-m_2}$, $m_2 \in \{0.1, 0, 3\}$
2	$\beta_{21}\,(z_1)$	0.15	0.15
2	$\beta_{22}\,(z_2)$	0.80	0.80
2	$\beta_{23}\,(t)$	-0.90	-0.90

Table 3.1: Data characteristics for external and current settings.

3.6.4 Experimental considerations

For a fixed sample size as shown in Table 3.1, I explored the impact, on estimation, of the level of misclassification and the level of double sampling. Simulations were performed while setting double sampling in the external sample at 20% and 50%. Misclassification was set at low and moderate levels. Low misclassification coincided with setting $(m_1 = 0.2, m_2 = 0.1)$, and moderate misclassification coincided with setting $(m_1 = 0.4, m_2 = 0.3)$.

3.6.5 Parameter estimation and performance

In each iteration of the simulation study, parameter estimation proceeded as follows:

1. Using the external sample, the misclassification paramaters, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, were estimated using the pseudo-likelihood approach described in Mpofu et al. (2019).
2. Under the assumption of transportability, I used the misclassification estimates from the external setting to estimate misclassification probabilities, $\pi_{jk}(\hat{\boldsymbol{\beta}}_k; \mathbf{X}_i)$, for the current sample of interest.
3. Using the current study-sample, I estimated $\boldsymbol{\tau} = (\alpha_1, \alpha_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ using the proposed pseudo-likelihood approach for modeling cause-specific hazards while adjusting for misclassification probabilities derived from external studies.
4. In addition, for the current sample, I estimated $\boldsymbol{\tau} = (\alpha_1, \alpha_2, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ using a naive approach that ignores outcome misclassification.

The steps described in 3.6.1, 3.6.2,3.6.3 and 3.6.5 were repeated for 2000 times. With the 2000 estimates, the following summary statistics were computed: average estimates: $\tilde{\boldsymbol{\tau}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\tau}}_i$; the absolute percent bias: %Bias $= \left|100 \times \frac{\tilde{\boldsymbol{\tau}} - \boldsymbol{\tau}_{true}}{\boldsymbol{\tau}_{true}}\right|$; the asympotic standard error: ASE$(\tilde{\boldsymbol{\tau}}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{SE}(\hat{\boldsymbol{\tau}}_i)$; the Monte-Calo standard deviation: $SD\left(\tilde{\boldsymbol{\tau}}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\boldsymbol{\tau}}_i - \tilde{\boldsymbol{\tau}}\right)^2$; and the 95 % coverage probability.

3.7 Simulation Results

I began the simulation study by considering the low misclassification setting. For the 2000 datasets used, on average, 24.4% of those who failed from cause 1 were classified as failing from cause 2, and 13.3% of those who failed from cause 2 were classified as failing from cause 1. With 20% double sampling in the external sample, the proposed estimation resulted in estimates with small bias and close to nominal 95% coverage. The variability of the pseudo-likelihood estimates was correctly estimated as shown by the closeness of the asymptotic standard errors (ASE) and Monte-Carlo standard deviations (MCSD). The same good performance was observed when there was 50% double sampling in the external study. At 50% double sampling, there was an attenuation in the variability of the estimates.

At moderate misclassification, on average, 45.7% of those who failed from cause 1 were classified as failing from cause 2, and 36.4% of those who failed from cause 2 were classified as failing from cause 1. The proposed method continued to show some of the good finite sample properties as seen under low misclassification. The proposed method resulted in point estimates with small bias. That said, the proposed method did not correctly compute the standard errors when double sampling was set at 20%: This was evidenced by the discrepancies between the ASE and MCSD. When double sampling was increased to 50%, the discrepancies between the ASE and MCSD between smaller than when double sampling was set at 20%.

In all the simulation scenarios considered in this study, naïve estimation was found to results in biased estimation. The results for all the simulations performed are presented in Table 3.2.

				Naïve					**Proposed**				
Study	**Cause**	**Misclass. (%)**	**Truth**	Estimate	% Bias	ASE	MCSD	CP	Estimate	% Bias	ASE	MCSD	CP
Double-samp. %			$\alpha_1 = 1.5\,$(Shape)	1.661	10.7	0.073	0.076	0.417	1.509	0.6	0.094	0.090	0.962
20	1	24.4	$\theta_{11} = -0.5$	-0.471	5.8	0.054	0.054	0.920	-0.508	1.6	0.094	0.084	0.966
Sample Size			$\theta_{21} = 0.8$	0.656	18.0	0.141	0.141	0.826	0.804	0.5	0.183	0.176	0.959
External=5000													
Current= 1000			$\alpha_2 = 2.5\,$(Shape)	2.186	12.6	0.078	0.084	0.040	2.514	0.6	0.201	0.180	0.964
Covariate Dist.	2	13.3	$\theta_{12} = -0.3$	-0.318	6.0	0.047	0.045	0.948	-0.301	0.3	0.074	0.069	0.967
Different			$\theta_{22} = 0.6$	0.731	21.8	0.132	0.127	0.835	0.599	0.2	0.201	0.190	0.968
Double-samp. %			$\alpha_1 = 1.5\,$(Shape)	1.661	10.7	0.073	0.076	0.417	1.508	0.5	0.084	0.081	0.958
50	1	24.4	$\theta_{11} = -0.5$	-0.471	5.8	0.054	0.054	0.920	-0.505	1.0	0.079	0.075	0.964
Sample Size			$\theta_{21} = 0.8$	0.656	18.0	0.141	0.141	0.826	0.803	0.4	0.172	0.170	0.958
External=5000													
Current= 1000			$\alpha_2 = 2.5\,$(Shape)	2.186	12.6	0.078	0.084	0.040	2.513	0.5	0.166	0.157	0.958
Covariate Dist.	2	13.3	$\theta_{12} = -0.3$	-0.318	6.0	0.047	0.045	0.948	-0.301	0.3	0.067	0.065	0.968
Different			$\theta_{22} = 0.6$	0.731	21.8	0.132	0.127	0.835	0.600	0.0	0.187	0.179	0.961
Double-samp. %			$\alpha_1 = 1.5\,$(Shape)	1.873	24.9	0.078	0.081	0.001	1.523	1.5	0.166	0.168	0.965
20	1	45.7	$\theta_{11} = -0.5$	-0.409	18.2	0.053	0.052	0.576	-0.505	1.0	0.148	0.128	0.950
Sample Size			$\theta_{21} = 0.8$	0.619	22.6	0.142	0.142	0.746	0.752	6.0	0.408	0.359	0.956
External=5000													
Current= 1000			$\alpha_2 = 2.5\,$(Shape)	1.983	20.7	0.074	0.081	0.000	2.518	0.7	0.328	0.275	0.950
Covariate Dist.	2	36.4	$\theta_{12} = -0.3$	-0.357	19.0	0.048	0.045	0.806	-0.311	3.7	0.107	0.094	0.960
Different			$\theta_{22} = 0.6$	0.755	25.8	0.131	0.126	0.796	0.599	0.2	0.411	0.357	0.952
Double-samp. %			$\alpha_1 = 1.5\,$(Shape)	1.873	24.9	0.078	0.081	0.001	1.518	1.2	0.153	0.159	0.956
50	1	45.7	$\theta_{11} = -0.5$	-0.409	18.2	0.053	0.052	0.576	-0.508	1.6	0.134	0.126	0.938
Sample Size			$\theta_{21} = 0.8$	0.619	22.6	0.142	0.142	0.746	0.762	4.8	0.358	0.353	0.942
External=5000													
Current= 1000			$\alpha_2 = 2.5\,$(Shape)	1.983	20.7	0.074	0.081	0.000	2.529	1.2	0.278	0.274	0.943
Covariate Dist.	2	36.4	$\theta_{12} = -0.3$	-0.357	19.0	0.048	0.045	0.806	-0.308	2.7	0.100	0.096	0.952
Different			$\theta_{22} = 0.6$	0.755	25.8	0.131	0.126	0.796	0.594	1.0	0.361	0.349	0.938

Table 3.2: The simulation results from modeling cause-specific hazards using the naive and proposed approaches.

3.8 Application

As stated in Section 3.1, I performed a data analysis in order to illustrate the use of the proposed method of modeling cause-specific hazards while adjusting for externally-sourced misclassification information. Particularly, I modelled the cause-specific hazards of death and disengagement from care among people living with HIV/AIDS (PLWH) that contributed data to IeDEA to East Africa. Data used in modeling were collected at two IeDEA programs: AMPATH(Academic Model Providing Access to Healthcare) and FACES (Family AIDS Care & Education Services). The data were collected between year 2001 and year 2011. In this period, AMPATH contributed 63,890 patients, and FACES contributed 3,886 patients. Study participants were followed from anti-retroviral (ART) initiation until death, disengagement from care or censoring. A patient was considered to be disengaged from care if he/she did not report for his/her next scheduled clinic visit and did not report for care within 60 days after next-scheduled visit date. The problem with such an approach is that some patients are mistakenly identified as disengaged from care when in fact they are dead. Given this possibility of death under-reporting, the outcomes for a sub-sample of patients deemed to be disengaged from care by the program workers were re-ascertained at AMPATH. The re-ascertainment entailed tracing patients within their communities and ascertaining the correct vital status. The same outcome validation, however, was not performed at FACES. As result, the extent of death misclassification was not identifiable at FACES. The time to any event (disengagement or death) at AMPATH and FACES was summarized as shown in Figure 3.2. Patient characteristics were further summarized as shown in Table 3.3.

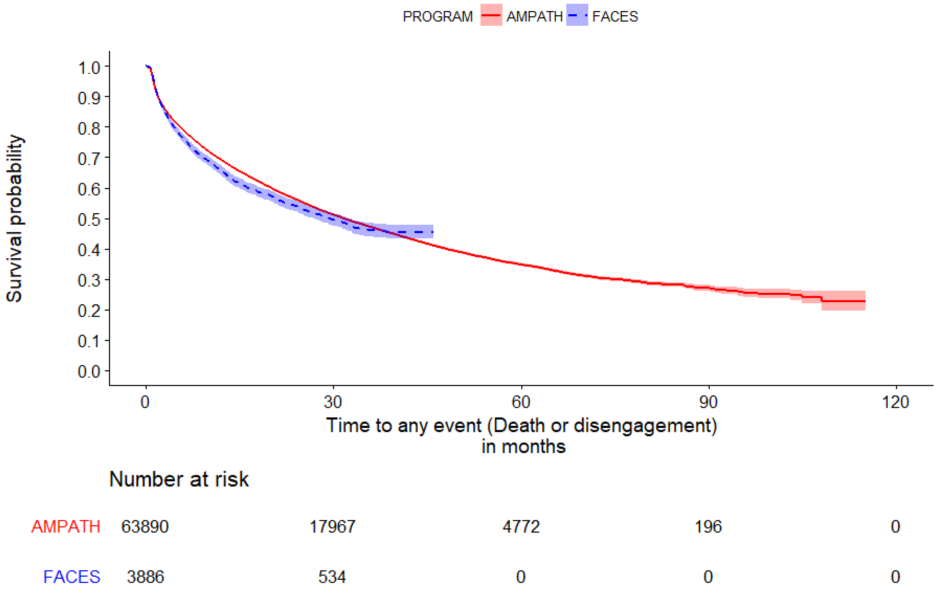

Figure 3.2: Summary of time to any event at AMPATH and FACES.

	East Africa IeDEA Program		
Variable	AMPATH, N=63890	FACES, N=3886	P Value
Gender, n(%)			0.722
Female	41944 (66)	2562 (66)	
Male	21946 (34)	1324 (34)	
Age at ART initiation			<.001
Mean [SD]	38.3 [9.2]	34.0 [9.7]	
Median (min - max)	37.3 (18.2 - 81.1)	32.2 (18.0 - 77.6)	
Time contributed to the study			<.001
Mean [SD]	31.6 [25.0]	14.7 [11.7]	
Median (min - max)	30.1 (0.0 - 115.3)	11.3 (0.1 - 45.9)	
CD4 count			<.001
Mean [SD]	172.5 [155.4]	202.8 [163.7]	
Median (min - max)	145.0 (0.0 - 2379.0)	182.0 (1.0 - 2811.0)	
Observed Cause Of Failure, n(%)			<.001
Censored	32711 (51)	2272 (58)	
Observed Death	2719 (4.3)	73 (1.9)	
Observed Loss to Clinic	28460 (45)	1541 (40)	

Table 3.3: Patient Characteristics at AMPATH and FACES

Except for the variable gender, there was evidence of differences of covariate distributions at AMPATH and FACES. The covariate distributions were summarized as shown in Figure 3.3.

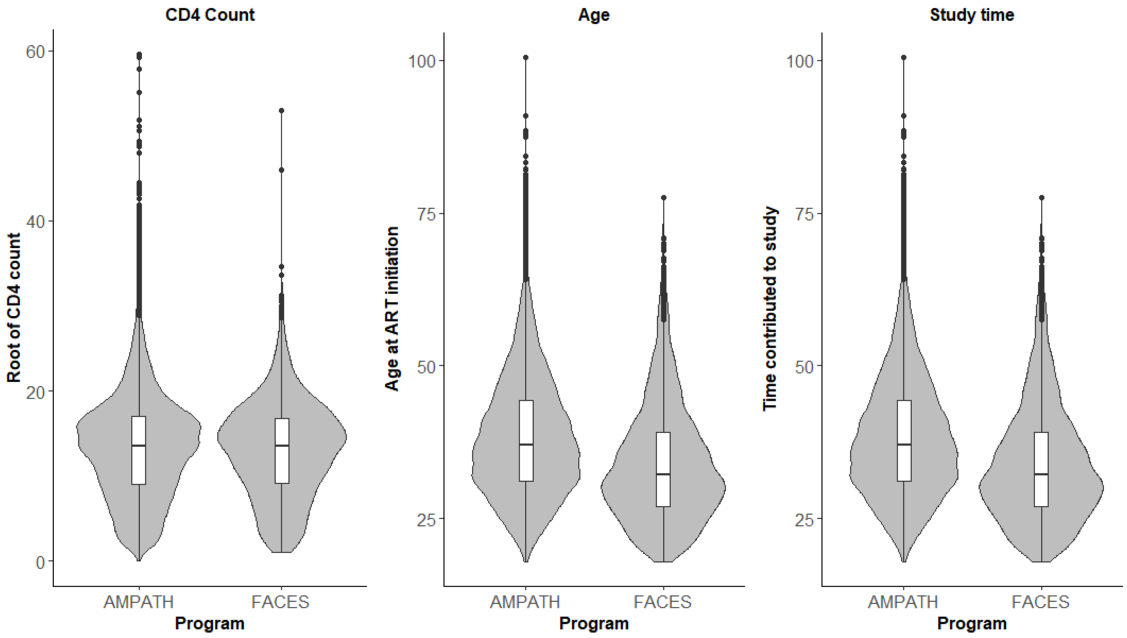

Figure 3.3: Covariate distributions at AMPATH and FACES.

3.8.1 Death misclassification model

Using data from AMPATH, I modeled the probability of being classified as disengaged from care when in fact dead (death misclassification). Among the 28,460 patients deemed to be disengaged from care by the healthcare workers, outcome validation was performed on 4238(14.9%). Among these cases, 1143(27%) were found to be actually deceased. As a result, the number of deaths increased from 2719 to 3862, meaning that 29.6%(1143/3862) were misclassified. The misclassification probabilities were modeled using the pseudo-likelihood approach presented by Mpofu et al. (2019). Death misclassification was modeled conditional on gender(male versus female), age at ART initiation, CD4 count at ART initiation (in square-root form), and time contributed to the study (in months) (in piece-wise linear form). The model for the log-odds of death misclassification at AMPATH was as shown in Table 3.4.

	Estimate	**SE**	**Z**	**Pr(>|Z|)**
(Intercept)	0.656	0.0743	8.838	0.0000
Gender (Male versus Female)	-0.208	0.0635	-3.273	0.0011
Centered Age (Age minus mean of age)	0.006	0.0030	1.886	0.0594
$\sqrt{\text{CD4}}$	0.016	0.0059	2.653	0.0080
Study time (months)	0.058	0.0087	6.629	0.0000
$I(3 \leq$ Study time $< 6)\times$ (Study time - 3)	-0.031	0.0358	-0.868	0.3856
$I($ $6 \leq$ Study time $< 12)\times$(Study time - 6)	-0.053	0.0232	-2.299	0.0215
I(Study time $\geq 12)\times$(Study time - 12)	-0.068	0.0107	-6.370	0.0000

Table 3.4: The model for the log-odds of death misclassification model at AMPATH.

The conditional death misclassification model at FACES was assumed to be same as that at AMPATH, assuming the transportability of misclassification across the two programs. The misclassification model as shown in Table 3.4 was used to compute predicted death misclassification probabilities at FACES. The resulting probabilities were adjust the models for the cause-specific hazards of death and disengagement from care at FACES. I assumed that the baseline cause-specific hazards took on the Weibull form, and also assumed multiplicative dependence between the hazards and the covariates. The cause-specific hazard model results at FACES were as presented in Table 3.5.

		Unadjusted				**Adjusted**			
				95% CI				95% CI	
Outcome	**Parameter**	Estimate	SE	Lower	Upper	Estimate	SE	Lower	Upper
Death	Shape (α_1)	0.446	0.048	0.351	0.541	0.621	0.039	0.545	0.696
	Scale (θ_{01})	-4.131	0.318	-4.753	-3.508	-2.920	0.298	-3.504	-2.336
	Male vs. Female (θ_{11})	0.529	0.243	0.052	1.005	0.060	0.226	-0.382	0.502
	$\sqrt{\text{CD4}}\,(\theta_{21})$	-0.102	0.022	-0.145	-0.058	-0.119	0.023	-0.163	-0.075
	Centered Age (θ_{31})	-0.001	0.013	-0.026	0.024	0.027	0.011	0.004	0.049
Disengagement	Shape (α_2)	0.781	0.017	0.747	0.815	0.796	0.020	0.757	0.834
	Scale (θ_{02})	-2.641	0.086	-2.809	-2.472	-3.040	0.127	-3.288	-2.792
	Male vs. Female (θ_{12})	0.079	0.056	-0.031	0.189	0.110	0.076	-0.039	0.259
	$\sqrt{\text{CD4}}\,(\theta_{22})$	-0.028	0.005	-0.038	-0.019	-0.014	0.007	-0.027	-0.001
	Centered Age (θ_{32})	-0.025	0.003	-0.031	-0.019	-0.037	0.005	-0.047	-0.027

Table 3.5: Models for cause-specific hazards of death and disengagement from care. In the first instance, there is no adjustment for possible misclassification, whereas, in the second instance, there is adjustment for possible misclassification.

The results showed several differences in the misclassification-unadjusted and -adjusted models. First, the adjusted model suggests a higher instantaneous risk of death, compared to the unadjusted model. The unadjusted model suggest that risk of death decreases with age, although the effect is not statistically significant at the 0.05 alpha level. The adjusted model, on the other hand, suggests that the risk of death increases with age, and the effect is statistically significant at the 0.05 alpha level. For disengagement from care, the adjusted model results were consistent with those of the unadjusted model, although the point estimates were different.

3.9 Discussion

Event cause misclassification represents a critical challenge in survival analysis with competing events. As many have shown, misattribution of the events could lead to bias in parameter estimation and undermine the validity of statistical inference. Appropriate estimating the misclassification probabilities and accounting for them help to alleviate the problem. In this research, I present a parametric method based on an external validation sample. I showed that the cause-specific hazards could be estimated using a pseudo-likelihood method, when the misclassification probabilities are estimated from an external validation sample under the assumption of transportability. The resultant hazard estimates remain consistent and asymptotically normally distributed. With a close-formed estimator for the variance, the proposed method provides a theoretical basis for large sample inference. An extensive simulation study further confirmed that the proposed model has a good finite-sample performance under various parameter settings.

Using the proposed method, I modeled the cause-specific hazards of death and care disengagement among people living with HIV/AIDS (PLWH) enrolled in the IeDEA network in East Africa. Taking advantage of the two IeDEA programs, namely AMPATH and FACES, I estimated the misclassification probabilities from the AMPATH sample, which had an internal validation subsample, transported the estimates to the FACES sample, which did not have a validation subsample, and successfully estimated the parameters of interest in a Weibull model. The example has clearly demonstrated the practical utility of the method.

A fundamental assumption used in this research is the transportability of the misclassification rates. Admittedly, short of a validation study, it is difficult to directly assess the validity of this assumption in a given application. An indirect verification of the assumption, however, can be ascertained from a goodness-of-fit test. We are currently examining the viability of such an indirect validation. Notwithstanding this limitation, we put forward a practical competing risk model that accounts for the misclassification errors in failure cause determination.

3.10 Appendix

3.10.1 Estimation objective

The estimating Equation 3.5 can also be written as follows:

$$\begin{aligned}
0 &= \sum_{i=1}^{n} \dot{l}_i(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \dot{l}_i(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) \\
&= \mathbb{P}_n \psi(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) \\
&= \Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e})
\end{aligned}$$

3.10.2 Proof of consistency

I will prove *Theorem 1* from Section 3.4.2 by showing that $\sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| \xrightarrow{p} 0$. That is, by showing that class of functions indexed by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\{\psi(\boldsymbol{\theta}, \boldsymbol{\beta}) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is P-Glivenko-Cantelli.

First, recognize that:

$$\begin{aligned}
\sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| &= \sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) + \Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| \\
&\leq \sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| \\
&+ \sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\|, \text{ by countable sub-addivity of norms}
\end{aligned}$$

Secondly:

a. $\sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| \xrightarrow{as*} 0$, by the strong law of large numbers.

b. Through a Taylor series expansion at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$,

$$\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) \approx \Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\dot{\Psi}_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) + o_p\left(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right)$$

It then follows that, $\sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\|$ is approximately equal to

$$\sup_{\theta\in\Theta} \|(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\dot{\Psi}_n(\boldsymbol{\theta}, \boldsymbol{\beta}_0) + o_p\left(\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|\right)\|.$$

$$\begin{aligned}\sup_{\theta\in\Theta} \|(\hat{\beta}_n - \beta_0)\dot{\Psi}_n(\boldsymbol{\theta}, \beta_0) + o_p\left(\|\hat{\beta}_n - \beta_0\|\right)\| &= \sup_{\theta\in\Theta} \|(\hat{\beta}_n - \beta_0)\left[\dot{\Psi}_n(\boldsymbol{\theta}, \beta_0) - \dot{\Psi}(\boldsymbol{\theta}, \beta_0) + \dot{\Psi}(\boldsymbol{\theta}, \beta_0)\right] + o_p\left(\|\hat{\beta}_n - \beta_0\|\right)\| \\ &\leq \sup_{\theta\in\Theta} \|(\hat{\beta}_n - \beta_0)\left[\dot{\Psi}_n(\boldsymbol{\theta}, \beta_0) - \dot{\Psi}(\boldsymbol{\theta}, \beta_0)\right]\| \\ &+ \sup_{\theta\in\Theta} \|(\hat{\beta}_n - \beta_0)\dot{\Psi}(\boldsymbol{\theta}, \beta_0)\| + o_p(1) \\ &= o_p(1)\end{aligned}$$

Given the results in a and b, one can conclude that $\sup_{\theta\in\Theta} \|\Psi_n(\boldsymbol{\theta}, \hat{\boldsymbol{\beta}}_n) - \Psi(\boldsymbol{\theta}, \boldsymbol{\beta}_0)\| \xrightarrow{p} 0$, hence $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\| \xrightarrow{p} 0$.

3.10.3 Proof of asymptotic normality

In order to prove *Theorem 2* from Section 3.4.2, begin by recognizing that,

$$\begin{aligned}\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) &= \Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) + \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) - \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) \\ &= \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) + \left[\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) - \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0)\right].\end{aligned}$$

By Taylor expansion at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) = \Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|) \tag{3.8}$$

And by Taylor series expansion at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$

$$\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) = \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) + \frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(|\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0|) \tag{3.9}$$

Plugging Equations 3.8 and 3.9 into Equation 3.10.3, it would follow that:

$$\begin{aligned}
\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) &= \Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) + \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) - \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) \\
&= \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) + \left[\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) - \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0)\right] \\
&= \Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|) \\
&+ \left[\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0) + \frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(|\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0|) - \Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta}_0)\right] \\
&= \Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|) \\
&+ \left[\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(|\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0|)\right],
\end{aligned}$$

where $\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$ is a $d_1 \times d_2$ matrix.

Assume that $\frac{n}{n_e} \to q$, $q > 0$, as the sample size goes to ∞. That is, the study and external sample will grow towards infinity at the same rate.

Pre- and post-multiply the left- and right-hand sides of Equation 3.10.3 by $\sqrt{n}$.

$$
\begin{aligned}
\sqrt{n}\Psi_n(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{\beta}}_{n_e}) &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{n}\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \sqrt{n}o_p(|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|) \\
&+ \sqrt{n}\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + \sqrt{n}o_p(|\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0|) \\
0 &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{n}\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(\sqrt{n}|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0|) \\
&+ \sqrt{n}\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + \sqrt{q}.o_p(\sqrt{n_e}|\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0|) \\
0 &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{n}\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(O_p(n^{-1/2})) \\
&+ \sqrt{n}\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + \sqrt{q}.o_p(O_p({n_e}^{-1/2})) \\
0 &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{n}\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&+ \sqrt{n}\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1).
\end{aligned}
$$

Based on Equation 3.10.3 it follows that:

$$
\begin{aligned}
0 &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) \\
&+ \sqrt{n}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0)\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + o_p(1) \\
&= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) \\
&+ \sqrt{n}\left(\dot{\Psi}_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) - I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0) + I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)\right)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&+ \sqrt{n}\left(\frac{d}{d\boldsymbol{\beta}}\Psi_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} - \frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} + \frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right)(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1) \\
&= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) \\
&- \sqrt{n}I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(1) \\
&+ \sqrt{q}\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\sqrt{n_e}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1) + o_p(1) \\
&= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) \\
&- \sqrt{n}I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&+ \sqrt{q}\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\sqrt{n_e}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1)
\end{aligned}
$$

From above, one can deduce that:

$$
\begin{aligned}
\sqrt{n}I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \sqrt{n}\Psi_n(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{q}\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\beta=\beta_0}\sqrt{n_e}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1) \\
\sqrt{n}I(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{l}_i(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) + \sqrt{q}\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\beta=\beta_0}\sqrt{n_e}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1) \\
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}I^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0) + \sqrt{q}I^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\beta=\beta_0}\sqrt{n_e}(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1) \\
\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}I^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0; \boldsymbol{\beta}_0) + \sqrt{q}\sqrt{n_e}\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{n_e} - \boldsymbol{\beta}_0) + o_p(1),
\end{aligned}
$$

where, $\mathbf{W}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0) = I^{-1}(\boldsymbol{\theta}_0, \boldsymbol{\beta}_0)\frac{d}{d\boldsymbol{\beta}}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\beta})|_{\beta=\beta_0}$ is a $d_1 \times d_2$ matrix $((d_1 \times d_1) \times (d_1 \times d_2))$.

By the central limit theorem:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0},\mathbf{I}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\right),$$

and,

$$\sqrt{n_e}.\sqrt{q}\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{n_e}-\boldsymbol{\beta}_0) \xrightarrow{d} N\left(\mathbf{0},q.\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)^T\right), \text{ for fixed } q.$$

Since the estimation of $\boldsymbol{\beta}$ is performed external to the sample of interest, one can assume the independence of the above two components. One can, therefore, deduce that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) + \sqrt{n_e}.\sqrt{q}\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{n_e}-\boldsymbol{\beta}_0) + o_p(1)$$

implies that:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0},\mathbf{I}^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)+q.\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\mathbf{I}^{-1}(\boldsymbol{\beta}_0)\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)^T\right).$$

When there is an internal-validation in the current study, there is no need to use misclassification probabilities from an external study since misclassification probabilities can now be estimated within the current study. In such a scenario, the independence between $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}_n$ that was assumed above no longer holds, since the same study sample is used in calculating both $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\theta}}_n$. For this scenario:

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) + \sqrt{n}\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)I^{-1}(\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\beta}_0) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) + \mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)I^{-1}(\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\beta}_0)\right] + o_p(1) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\Psi}_i(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0) + o_p(1)
\end{aligned}$$

where $\tilde{\Psi}_i(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0) = I^{-1}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\theta}_0;\boldsymbol{\beta}_0) + \mathbf{W}(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)I^{-1}(\boldsymbol{\beta}_0)\dot{l}_i(\boldsymbol{\beta}_0)$.

By the central-limit theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0},E\left[\tilde{\Psi}_i(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)\tilde{\Psi}_i(\boldsymbol{\theta}_0,\boldsymbol{\beta}_0)^T\right]\right)$$

3.10.4 Derivation of the general score function

The derivation of the general socre function as shown in Equation 3.5 is presented below:

$$\begin{aligned}
\nabla l(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\phi}_1,\boldsymbol{\phi}_2) &= \begin{bmatrix} \frac{dl(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\phi}_1,\boldsymbol{\phi}_2)}{d\boldsymbol{\theta}_1} \\ \frac{dl(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\phi}_1,\boldsymbol{\phi}_2)}{d\boldsymbol{\theta}_2} \\ \frac{dl(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\phi}_1,\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_1} \\ \frac{dl(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\boldsymbol{\phi}_1,\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2} \end{bmatrix} \\
&= \begin{bmatrix}
\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \frac{\delta_{i1}^{*}\pi_{11}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)\lambda_{01}(t_i;\phi_1)\exp(\boldsymbol{Z}\boldsymbol{\theta}_1)}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{21}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)\lambda_{01}(t_i;\phi_1)\exp(\boldsymbol{Z}\boldsymbol{\theta}_1)}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} - \Lambda_{01}(t_i;\phi_1)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_1) \right\} \\
\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \frac{\delta_{i1}^{*}\pi_{12}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)\lambda_{02}(t_i;\boldsymbol{\phi}_2)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_2)}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{22}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)\lambda_{02}(t_i;\boldsymbol{\phi}_2)\exp(\boldsymbol{Z}\boldsymbol{\theta}_2)}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} - \Lambda_{02}(t_i;\boldsymbol{\phi}_2)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_2) \right\} \\
\sum_{i=1}^{n} \left\{ \frac{\delta_{i1}^{*}\pi_{11}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_1)\frac{d\lambda_{01}(t_i;\boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1}}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{21}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)\exp(\boldsymbol{Z}\boldsymbol{\theta}_1)\frac{d\lambda_{01}(t_i;\boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1}}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} - \frac{d\Lambda_{01}(t_i;\boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1}\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_1) \right\} \\
\sum_{i=1}^{n} \left\{ \frac{\delta_{i1}^{*}\pi_{12}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)\exp(\boldsymbol{Z}\boldsymbol{\theta}_2)\frac{d\lambda_{02}(t_i;\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2}}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{1k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{22}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)\exp(\boldsymbol{Z}\boldsymbol{\theta}_2)\frac{d\lambda_{02}(t_i;\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2}}{\sum_{k=1}^{2}\lambda_{0k}(t_i;\boldsymbol{\phi}_k)\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_k)\pi_{2k}^{*}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{Z}_i)} - \frac{d\Lambda_{02}(t_i;\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2}\exp(\boldsymbol{Z}_i\boldsymbol{\theta}_2) \right\}
\end{bmatrix} \\
&= \begin{bmatrix}
\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^{*}\pi_{11}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{21}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} \right) .\lambda_{01}(t_i;\boldsymbol{\phi}_1) - \Lambda_{01}(t_i;\boldsymbol{\phi}_1) \right\} \exp(\boldsymbol{Z}_i\boldsymbol{\theta}_1) \\
\sum_{i=1}^{n} \boldsymbol{Z}_i \left\{ \left(\frac{\delta_{i1}^{*}\pi_{12}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{22}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} \right) \lambda_{02}(t_i;\boldsymbol{\phi}_2) - \Lambda_{02}(t_i;\boldsymbol{\phi}_2) \right\} \exp(\boldsymbol{Z}_i\theta_2) \\
\sum_{i=1}^{n} \left\{ \left(\frac{\delta_{i1}^{*}\pi_{11}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{21}^{*}(\hat{\boldsymbol{\beta}}_1;\boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} \right) . \frac{d\lambda_{01}(t_i;\boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1} - \frac{d\Lambda_{01}(t_i;\boldsymbol{\phi}_1)}{d\boldsymbol{\phi}_1} \right\} \exp(\boldsymbol{Z}_i\boldsymbol{\theta}_1) \\
\sum_{i=1}^{n} \left\{ \left(\frac{\delta_{i1}^{*}\pi_{12}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)}{p_1^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} + \frac{\delta_{i2}^{*}\pi_{22}^{*}(\hat{\boldsymbol{\beta}}_2;\boldsymbol{Z}_i)}{p_2^{*}(\boldsymbol{\theta},\boldsymbol{\phi},\hat{\boldsymbol{\beta}},\boldsymbol{Z}_i)} \right) . \frac{d\lambda_{02}(t_i;\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2} - \frac{d\Lambda_{02}(t_i;\boldsymbol{\phi}_2)}{d\boldsymbol{\phi}_2} \right\} \exp(\boldsymbol{Z}_i\boldsymbol{\theta}_2)
\end{bmatrix}
\end{aligned}$$

CHAPTER 4

Showcasing of Validation-sampling Remedies when Modeling Cause-specific Hazards in the Presence of Outcome Misclassification

In the previous chapters, I spent time to state the problem of outcome misclassification in competing risks analysis. I proposed statistical solutions for adjusting for misclassification when modeling cause-specific hazards. In this chapter, I will recapitulate all the ideas presented in Chapters **??**, 2 and 3. With an epidemiology audience in mind, I illustrate internal- and external-validation sampling remedies for dealing with outcome misclassification in studies with competing risks. Specifically, I will present pseudo-likelihood-based methods for correctly modeling cause-specific hazards, depending on the availability of an outcome validation sample. I will highlight the statistical methods by modeling of cause-specific hazards of death and disengagement from care, among people living with HIV/AIDS who contribute data to IeDEA East Africa: Standard competing risks methods are not suitable for such a task because of death misclassification, with some patients being classified as disengaged from care when they are truly dead.

4.1 Introduction

Time-to-event studies with competing events may be susceptible to event misclassification. For example, in a mortality study considering cardiovascular and non-cardiovascular causes of death, it is possible for some of those who die from cardiovascular diseases to be classified as having died from non-cardiovascular diseases, and vice versa. Such an error in classification can lead to bias when modeling competing-risks quantities such as cause-specific hazards and cumulative incidence functions. Estimation bias may be remedied by either using gold-standard approaches to repeat data collection or by using statistical methods that adjust estimators of interest for misclassification. The latter approach is cost-effective, hence more favorable than the former remedy. That said, statistical remedies are only suitable when there exist information about the extent of outcome misclassification.

One way to identify the extent of outcome misclassification is using internal-validation or double sampling. This involves re-ascertaining outcomes on a sub-sample of the main-study sample using a gold-standard approach, that is more accurate, and is usually more expensive than the initial outcome-ascertainment approach (Tenenbein 1970; R. J. Carroll et al. 2006). Such internal validation is used in studies of vital status among people living with HIV/AIDS (PLWH) who contribute data to the International Epidemiologic Databases for the Evaluation of HIV/AIDS in East Africa (IeDEA-EA). In this context, the need for outcome validation arises because some patients are mistakenly identified as disengaged from care when, in fact, they are dead. This error in observation can bias competing-risks analyses wherein death and disengagement from care are the outcomes of interest (Bakoyannis and Yiannoutsos 2015). That being said, such bias may be be avoided by adjusting estimators using the death-misclassification information resulting from internal-validation sampling. For IeDEA, validation entails re-ascertaining vital-status on sub-samples of those who are initially deemed to be disengaged from care. However, due to financial constraints, it is not feasible to perform the aforementioned internal validation at all the treatment programs that contribute data to IeDEA-EA. Consequently, not all the treatment programs have misclassification information required to adjust competing-risks estimators for possible death misclassification.

The challenge for treatment programs that do not have outcome validation is two-fold: First, competing-risks quantities estimated using the observed data are likely to be biased due to possible misclassification; Secondly, misclassification adjustment is not immediately available based on the study sample at hand. The aforementioned challenges may be solved by using misclassification probabilities modeled in an external setting that has outcome validation, assuming the *transportability* of misclassification probabilities across different settings. In other words, we assume that for a fixed set of patient characteristics, an individual from a setting without validation has the same propensity to be misclassified as an individual from a setting with outcome validation. In epidemiology, such a use of external information is refered to as external validation (Lyles and Lin 2010).

With an epidemiologic audience in mind, I will illustrate internal- and external-validation sampling solutions when modeling the cause-specific hazards of death and disengagement from care as described in the motivating example. I will follow the pseudo-likelihood approach for modeling cause-specific hazards as presented in Chapter 3. Under this approach, the additional parameters in the likelihood are related to outcome misclassification, and are replaced by their estimates. The advantage of this approach is that it is not only statistically principled, but it is also easy to understand and implement in readily available statistical software such as `R`. Moreover, as shown in Chapter 3, the approach has nice large- and finite-sample properties assuming the *transportability* assumption holds. The presentation will proceed as follows: In Section 4.2 I review some basics of competing risks survival analysis, under the ideal scenario with no misclassification, and when there is misclassification. In Section 4.3, I describe the data analysis methodology. In Section 4.4, I describe the results, and provide a brief discussion of the findings in Section 4.5.

4.2 Review of competing risks survival analysis

Survival analysis involving competing events is called competing risks. Events are called competing risks, within the study context or by nature, if observing one event precludes us from observing the other events. In the motivating study, death and disengagement are considered to be competing risks because interest lies in whichever event comes first. Observing disengagement from care, within the study context, precludes us from observing death, and vice versa.

4.2.1 Notation

In keeping with the two-cause system similar to that of our motivating study, let $C \in \{1, 2\}$ represent the true cause of failure, and $C^* \in \{1, 2\}$ represent the observed cause cause of failure. The observed cause is subject to observation error, therefore C^* is not necessarily the same as C. Censoring is represented by $C = C^* = 0$. The right-censoring time to failure is defined as $T = \min(U, V)$, where U is the time to event, and V is the time to censoring. The time to censoring is assumed to be independent of both the time to event and the cause of failure. Lastly, let $\boldsymbol{Z}$ represent the independent variables.

4.2.2 Definitions of common quantities

In competing risks analysis, the quantities of interest usually include: cause-specific hazards and cumulative incidence functions. Assuming a world where people can fail from any cause $j \in \{1, 2\}$:

1. The cause-specific hazard, $\lambda_j(t)$, is defined as the instantaneous rate of failing to due to

cause j at time t, conditional of on surviving to at least beyond t, in a world where one can fail from the other cause (Kalbfleisch and Prentice 2011). In biomedical studies, models of cause-specific hazards are typically used for identifying the risk factors for causes of failure (Hinchliffe, Abrams, and Lambert 2013; Austin, Lee, and Fine 2016).

2. The cumulative incidence function for cause-j, $F_j(t) = P[T \leq t, J = j]$, measures the absolute risk of failure due to cause-j by a certain time point, say t (Kalbfleisch and Prentice 2011). In biomedical studies, cumulative incidence functions are used to make predictions (Hinchliffe, Abrams, and Lambert 2013; Austin, Lee, and Fine 2016). The cumulative incidence for cause-j can be also written as follows: $F_j(t) = \int_0^t \lambda_j(u) \exp\left[-\int_0^u \sum_{j=1}^2 \lambda_j(s)ds\right] du$. This definition highlights the dual use of cause-specific hazards in competing risks. That is, cause-specific hazard models can be used for studying relationships and, after transformation, for making predictions.

Similar of Chapter 3, I will restrict our focus to cause-specific hazards. In particular, I model cause-specific hazards in order to identify the risk factors of death and disengagement from care in a cohort of PLWH who contribute data to IeDEA East Africa.

4.2.3 Cause-specfic hazards in the presence of outcome misclassification

In a world without outcome misclassification, that is, $C = C^*$, Kalbfliesh and Prentice (2011) (Kalbfleisch and Prentice 2011) showed that the modeling cause-specific hazards is fairly simple, since the likelihood of interest is a function of cause-specific hazards. The log-likelihood for a two-cause system with no outcome misclassification is as shown in Equation 4.1 (Kalbfleisch and Prentice 2011).

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{2}\sum_{i=1}^{n}\left\{\delta_{ij}\log\lambda_j(t_i;\boldsymbol{Z}_i,\boldsymbol{\theta}_j) - \int_0^{t_i}\lambda_j(u;\boldsymbol{Z}_i,\boldsymbol{\theta}_j)du\right\} \tag{4.1}$$

where

1. $\lambda_j(t;Z)$ is the cause-specific hazard of cause j at time t conditional on covariates $\boldsymbol{Z}$, for $j \in \{1,2\}$;
2. $\delta_{ij} = I[C_i = j]$ is the event indicator of cause-j for subject i.
3. $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is the parameter of interest: It captures the association between the cause-specific hazards and the independent variables.

The log-likelihood presentation shown in Equation 4.1 allows for modeling cause-specific hazards as regular marginal hazards. That is, without losing generality, when modeling cause-specific hazard for cause 1, events attributed to cause 2 are treated as censored outcomes.

In the presence of outcome misclassification, it can be shown that the log-likelihood can be written in as shown by Equation 4.2.

$$\begin{aligned} l^*(\boldsymbol{\theta},\boldsymbol{\beta}) = &\sum_{j=1}^{2}\sum_{i=1}^{n}\delta^*_{ij}\log\left[\sum_{k=1}^{2}\lambda_k(t_i;\boldsymbol{Z}_i,\boldsymbol{\theta}_k)\pi^*_{jk}(\boldsymbol{\beta}_k;\boldsymbol{X}_i)\right] \\ &-\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t_i}\lambda_j(u;\boldsymbol{Z}_i,\boldsymbol{\theta}_j)du, \end{aligned} \tag{4.2}$$

1. $\delta^*_{ij} = I[C^* = j]$ is the event indicator of observed cause-j for subject i
2. For $j,k \in \{1,2\}$, $\pi^*_{jk}(\boldsymbol{X}_i;\boldsymbol{\beta}_k) = P\left(C^*_i = j | C_i = k, \boldsymbol{X}_i = (\boldsymbol{Z}_i, T = t), \boldsymbol{\beta}_k\right)$ is the proba-

bility of observing cause j when the true cause of failure is k, conditional on subject characteristics $\boldsymbol{X}_i$ (misclassification probability).

In the log-likelihood 4.2, $\boldsymbol{\theta}$ is the parameter of interest, and $\boldsymbol{\beta}$ is a nuisance parameter. Under this likelihood formulation, we can no longer model the cause-specific hazards individually as when using log-likelihood 4.1.

The likelihood formulation given in Equation 4.2 is suitable in both situations where there exists or there does not exist an internal-validation sample. The former is a case of internal validation, and the latter is case of external validation. It is, however, not efficient to use likelihood formulation 4.2 when there is an internal-validation sample. Reason being that likelihood 4.2 does not directly incorporate the validated outcomes into the likelihood and, therefore, a waste of data. A likelihood form suitable when there is internal-validation sampling is presented in Bakoyannis et al. (2019), and can be written as shown by log-likelihood 4.3 (Bakoyannis, Zhang, and Yiannoutsos 2018).

$$\begin{aligned} l(\boldsymbol{\theta},\boldsymbol{\eta}) = &\sum_{j=1}^{2}\sum_{i=1}^{n}\left[R_i\delta_{ij} + \sum_{k=1}^{2}\delta_{ik}^{*}\times(1-R_i)\times p_{jk}(\boldsymbol{\eta}_k;\boldsymbol{Z}_i)\right]\log\lambda_j(t_i;\boldsymbol{\theta}_j,\boldsymbol{Z}_i) \\ &-\sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t_i}\lambda_j(u;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)du \end{aligned} \tag{4.3}$$

where,

1. $R_i = 1$ indicates that the outcome for subject i was validated (thereby known);
2. $\delta_{ij} = I[C_i = j]$ is the event indicator of true cause-j for subject i;
3. $\delta_{ik}^{*} = I[C_i^{*} = k]$ is the event indicator of observed cause-k for subject i;
4. $p_{jk}(\boldsymbol{\eta}_k;\boldsymbol{Z}_i) = P[C_i = j|C_i^{*} = k, T_i = t_i, \boldsymbol{Z}_i, \boldsymbol{\eta}_k]$, is the probability that the true cause

of failure is cause-j given the observed cause is cause-k(i.e., predictive values), given $(\boldsymbol{Z}, T = t)$.

Log-likelihood 4.3 represents a re-expression of a misclassification problem as a missing-data problem. When using log-likelihood 4.3, outcomes for those whose observed outcomes were not validated are treated as missing data. In this case, the missing values are replaced by the conditional expectations of the true outcome values given the observed data; that is, the predictive values. Such a substitution is justified by the linearity between the log-likelihood and the true outcomes as shown by log-likelihood 4.1. Moreover, missing outcome values are assumed to be missing at random (MAR). That is, missingness can be explained by the observed data, and not the unobserved outcomes (Rubin 1976). Under the notation defined above, the MAR assumption is formally defined as follows:

$$P[R = 0|C = 1, C^* \neq 0, \boldsymbol{Z}] = P[R = 0|C = 2, C^* \neq 0, \boldsymbol{Z}] = P[R = 0|C^* \neq 0, \boldsymbol{Z}]$$

Without losing generality, this implies that,

$$P[C = 1|R = 0, C^* \neq 0, \boldsymbol{Z}] = P[C = 1|R = 1, C^* \neq 0, \boldsymbol{Z}] = P[C = 1|C^* \neq 0, \boldsymbol{Z}]$$

Colloquially this means that the predictive value model is independent of whether or not observed outcome was validated. Therefore, one can use the predictive value model for those whose outcomes were validated to inform the predictive values for those whose outcomes were not validated.

4.3 Methods

4.3.1 Study design and setting

This study is a retrospective-cohort study consisting of PLWH who initiated anti-retroviral therapy (ART) at treatment centers that contributed data to the East Africa IeDEA consortium. The treatment programs included AMPATH (Academic Model Providing Access to Healthcare), and FACES (Family AIDS Care & Education Services). AMPATH contributed 63,890(84.14%) patients, and FACES contributed 12,043(15.86%) patients. Data from AMPATH were collected between 2001 and 2011, and data from FACES were collected between 2007 and 2014. The study was restricted to patients who initiated ART at age 18 or older. Patients were followed from ART initiation until death, disengagement from care, or administrative censoring.

4.3.2 Data collection and management

The data used in this study were collected during the routine care of patients at 31 treament centers belonging to AMPATH, and 8 FACES treatment centers. Data abstraction from electronic medical records, data quality control, and data preparation for analysis were managed by the Data management team at IeDEA East Africa.

4.3.3 Ethics Statement

Data were used in this manuscript with permission from both AMPATH and FACES. Institutional Reviews Boards (IRB) associated with AMPATH, FACES and Indiana University approved this study.

4.3.4 Outcomes and outcome validation

The outcomes of interest in the study were death and disengagement while in HIV care. A patient was considered to be disengaged from care if he/she missed the next scheduled clinic visit on the medical chart, and did not report for care within the 60 days following the next scheduled visit. Death and disengagement from care were considered to be competing risks, because time-to-event was based on whichever event came first. For the purpose of this analysis, let $C = 1$ and $C = 2$ represent the true death and disengagement events respectively. In addition, let $C^* = 1$ and $C^* = 2$ represent observed death and disengagement respectively. Once again, I reiterate to the reader that "observed" events are those ascertained using an error-prone method, and "true" events are those ascertained using a gold-standard methods.

In our study, "observed" deaths were always correctly ascertained. The same, however, could not be said for "observed" disengagers because of death under-reporting: Some of those observed as disengaged were actually dead. Given this death under-reporting, both AMPATH and FACES traced some of the patients who had been absent from care for 90 at least days. This community outreach, although done to improve patient retention in HIV care, had a useful side-effect: It validated the vital-status data from those who were initially classified as disengaged from care. AMPATH validated the vital-status data of about 15.3% of those who were initially classified as disengaged. FACES, on the other hand, validated about 2.3% of the vital-status data of those who were initially classified as disengaged from care. The validation data from FACES was considered too small, and was not used in further analyses. I relied on misclassification information from AMPATH to adjust cause-specific hazard models at FACES.

4.3.5 Identifying misclassified deaths among the "observed" disengagers at AMPATH

Misclassified deaths could be observed/identified within the validated portion of the study sample at AMPATH. In the temporal-frame of study, death misclassification or lackthereof could be divided into the following four cases:

1. *Case 1*: Patient is initially classified as disengaged from care, and found to be alive during outreach. The initial decision to classify the patient as disengaged from care is correct. This case is illustrated in Figure 4.1.

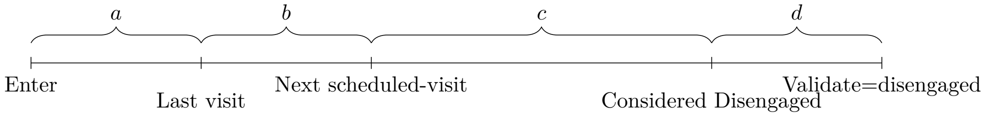

Figure 4.1: Case 1: Dealing with observed disengagement

In case 1, the observed outcome was $C^* = 2$; the true outcome was $C = 2$, and the time-to-event was $t = a + b + c$.

2. *Case 2*: Patient is initially classified as disengaged from care, and found to be dead during outreach, however the death occurred after date of disengagement. The initial decision to classify the patient as disengaged from care is correct. This case is illustrated in Figure 4.2.

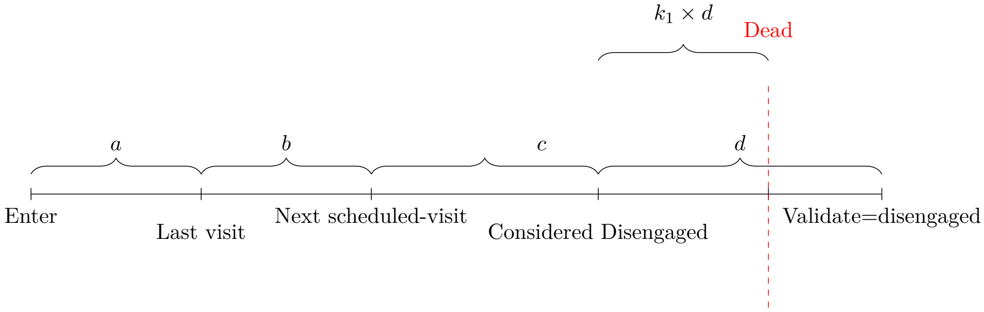

Figure 4.2: Dealing with observed disengagement

In case 2, the observed outcome was $C^* = 2$; the true outcome was $C = 2$, and the time-to-event was $t = a + b + c$.

3. *Case 3*: Patient is initially classified as disengaged from care, and found to be dead during outreach, and the death occurred before date of disengagement. In fact, death occurred before the planned next-scheduled visit date. The initial decision to classify the patient as disengaged from care is incorrect. This case is illustrated in Figure 4.3.

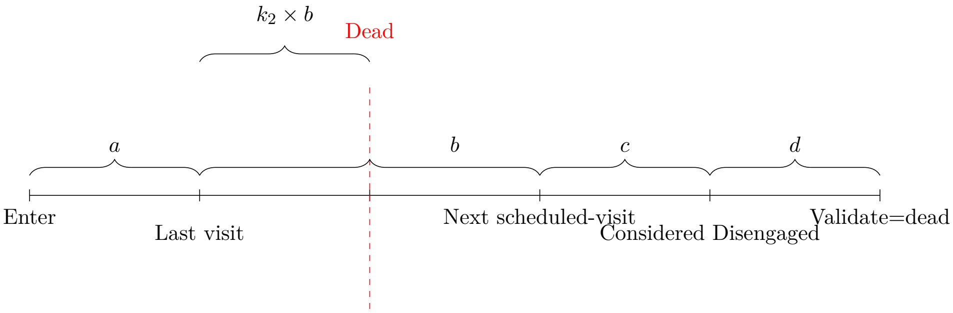

Figure 4.3: Dealing with observed disengagement

For case 3, the observed outcome was $C^* = 2$; the true outcome was $C = 1$, the time-to-event was $t = a + b + c$.

4. *Case 4*: Patient is initially classified as disengaged from care, and found to be dead during outreach, and the death occurred before date of disengagement. In fact, death occurred between the next-scheduled visit date and the supposed disengagement date. The initial decision to classify the patient as disengaged from care is inccorrect. This case is illustrated in Figure 4.4.

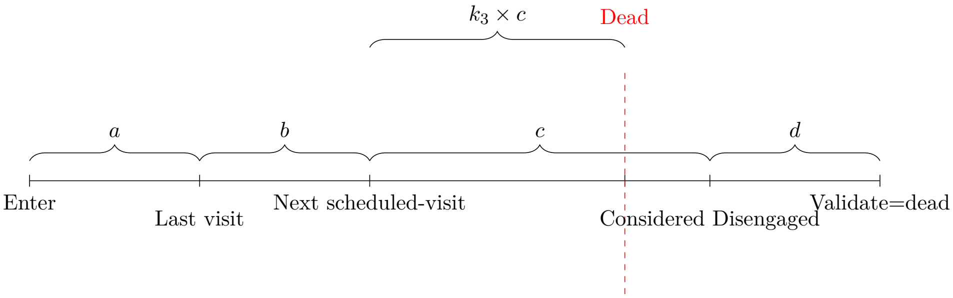

Figure 4.4: Dealing with observed disengagement

In case 4, the observed outcome was $C^* = 2$; the true outcome was $C = 1$, and the time-to-event was $t = a + b + c$.

For all the four cases considered, the time-to-event was ascertained at the date when the patient was initially considered to be disengaged from care. For case 3 and 4, time-to-event is over-estimated by at most 90 days.

4.3.6 Independent variables

The covariates considered in this study included: Sex(Male versus female), age at ART initiation, CD4 count at ART initiation per (μL), WHO Stage at ART initiation (1,2,3,4), BMI at ART initiation, care facility type(clinic, hospital) and the setting of care facility (urban, rural). The time contributed to the study, although part of the survival outcome, was also considered to be an independent variable when modeling misclassification probabilities and predictive values.

4.3.7 Statistical Analysis

Data were summarized by treatment program. Continuous variables were summarized using mean, standard deviation, median, and inter-quartile range (IQR). Categorical variables were summarized using frequencies. As appropriate, independent variables were compared by treatment program using one-way ANOVA, Pearson Chi-squared or Fisher's exact test.

The dataset used consisted of 66934(88.15%) complete cases, and 8999(11.85%) subjects with at least one missing value. At AMPATH, about 10.29% (6572/63890) of subjects had a missing value, and 20.15% (2427/12043) of the subjects at FACES had a missing value. With data stratified by treatment program (AMPATH, FACES), the missing values for CD4 count, weight, height and WHO stage at ART initiation were multiply imputed 100 times using the fully conditional specification (FCS) (**???**).

At AMPATH, the cause-specific hazards of death and disengagement were modelled using two approaches. First using likelihood 4.2, and secondly using likelihood 4.3. The analysis was performed as depicted in Figure 4.5.

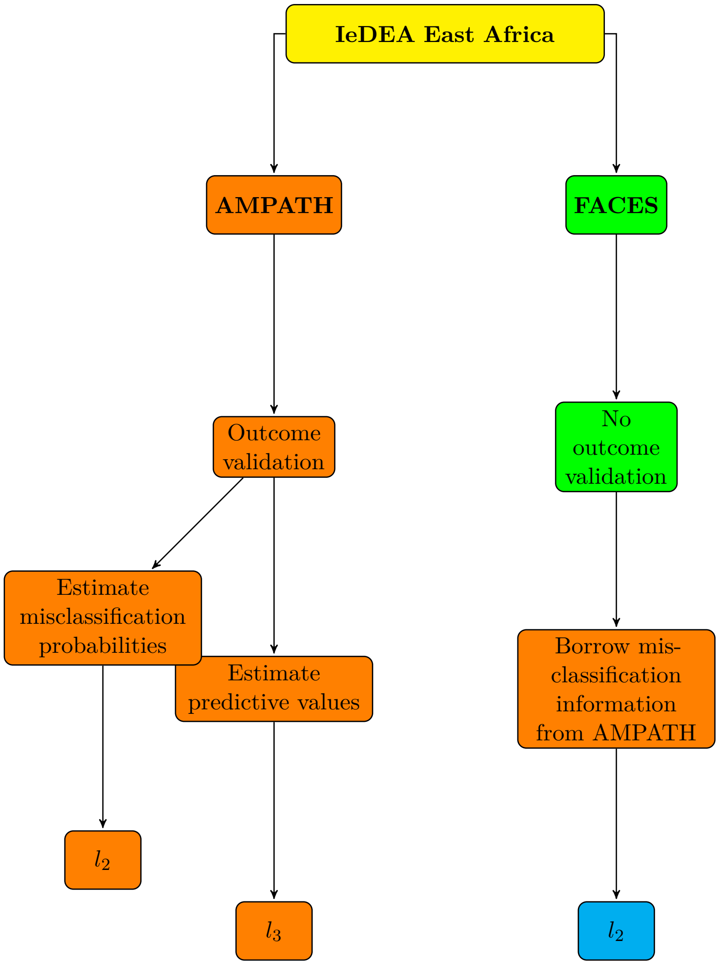
IeDEA East Africa
AMPATH
FACES
Outcome validation
No outcome validation
Estimate misclassification probabilities
Estimate predictive values
Borrow misclassification information from AMPATH
l_2
l_3
l_2

Figure 4.5: Scheme for modeling cause-specific hazards of death and disengagement from care at AMPATH and FACES while adjusting for death misclassification.

4.3.7.1 Predictive value model of death among "observed" disengagers at AMPATH

The probability that a patient was truly dead given that he/she was observed to be disengaged was modelled using logistic regression. Specifically, I modeled the log-odds of $P[C = 1|C^* = 2, \boldsymbol{X} = (\boldsymbol{Z}, T = t)]$. The model included the independent variables noted in Section 4.3.6, and the time to the observed disengagement. The goodness-of-fit for the overall model was assessed using the Supremum goodness-of-fit test at the 0.05 alpha level (D. Y. Lin, Wei, and Ying 2002).

4.3.7.2 Death misclassification model at AMPATH

The probability of being classified as disengaged from care when, in fact, dead was modeled using the logistic pseudo-likelihood approach presented in Chapter 2. Under this approach, true outcomes for those who were not validated were assumed to be missing at random (MAR). This assumption allowed us to replace the missing true outcome values with their predictive values, as modeled in Section 4.3.7.1. This approach for modeling misclassification probabilities was used because it uses both the validated and the unvalidated sample, thereby leading to significant efficiency gains over an approach that only uses the validated sample(complete-case analysis). The covariates included in the misclassification model were the same as those included in the predictive value model in Section 4.3.7.1.

4.3.7.3 Cause-specific hazards

I modeled the cause-specific hazards of death and disengagement from care at AMPATH and FACES parametrically using the schematic presented in Figure 4.5. In addition to assuming proportional hazards, I assumed that the baseline cause-specific hazards took on the Weibull form. Generically, the cause-specific hazard model could be represented as follows:

$$\lambda_j(t|\boldsymbol{Z}) = \alpha_j \rho_j t^{\alpha_j - 1} \exp\left(\boldsymbol{Z}\boldsymbol{\theta}_j\right)$$

for $j \in \{1, 2\}$, with α_j being the shape parameter, ρ_j being the scale parameter, $\boldsymbol{Z}$ being the matrix of covariates noted in Section 4.3.6, and $\boldsymbol{\theta}_j$ being the logarithm of the multiplicative dependence between the covariates and the cause-specific hazard for cause-j.

Cause-specific hazards were modeled using either internal- or external-validation approaches as outlined in Table 4.1. The analysis at AMPATH was based on internal validation, and the analysis at FACES was based on external validation since misclassification information was borrowed from AMPATH. Death misclassification probabilities were used to adjust cause-specific hazards models as described by Mpofu et al. (2019) in Chapter 3, and depicted by Figure 4.6. Predictive values were used as described by Bakoyannis et al. (2019), and shown in Figure Figure 4.7.

Treament Program	Validation	Problem treated as	Remedy relies on	Data source	Likehood form used
AMPATH	Internal	Missing-data problem	Predictive values	AMPATH	l_3
		Misclassification problem	Misclassification probabilities	AMPATH	l_2
FACES	External	Misclassification problem	Misclassification probabilities	AMPATH and FACES	l_2

Table 4.1: Statistical methods considered for dealing with death misclassification when modeling the cause-specific hazards of death and disengagement from care.

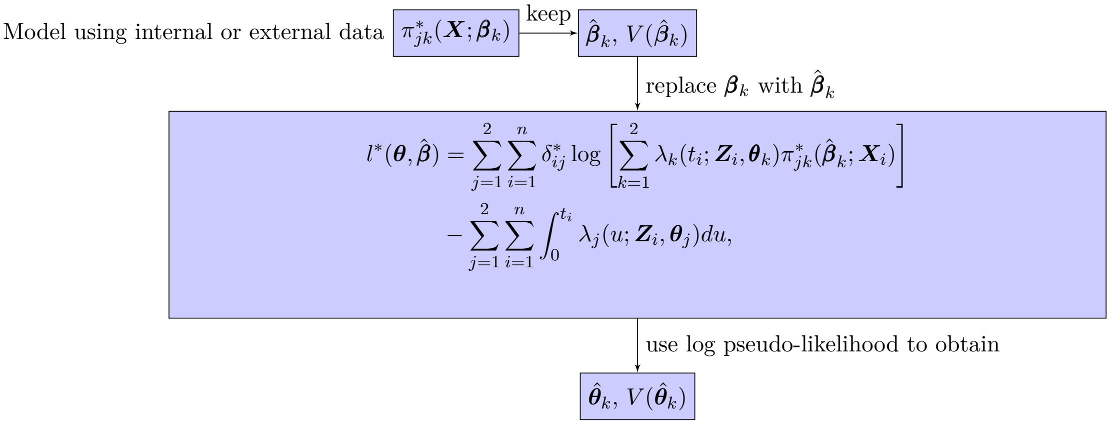
Model using internal or external data
$\pi^*_{jk}(\boldsymbol{X};\boldsymbol{\beta}_k)$
keep
$\hat{\boldsymbol{\beta}}_k,\, V(\hat{\boldsymbol{\beta}}_k)$
replace $\boldsymbol{\beta}_k$ with $\hat{\boldsymbol{\beta}}_k$
$l^*(\boldsymbol{\theta},\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{2}\sum_{i=1}^{n}\delta^*_{ij}\log\left[\sum_{k=1}^{2}\lambda_k(t_i;\boldsymbol{Z}_i,\boldsymbol{\theta}_k)\pi^*_{jk}(\hat{\boldsymbol{\beta}}_k;\boldsymbol{X}_i)\right] - \sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t_i}\lambda_j(u;\boldsymbol{Z}_i,\boldsymbol{\theta}_j)du,$
use log pseudo-likelihood to obtain
$\hat{\boldsymbol{\theta}}_k,\, V(\hat{\boldsymbol{\theta}}_k)$

Figure 4.6: Misclassification-probability based pseudo-likelihood approach for modeling cause-specific hazards.

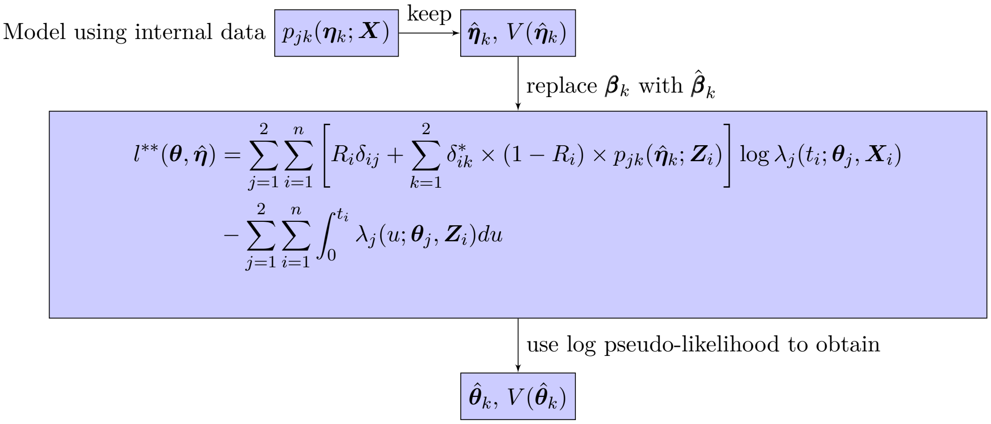
Model using internal data
$p_{jk}(\boldsymbol{\eta}_k;\boldsymbol{X})$
keep
$\hat{\boldsymbol{\eta}}_k,\, V(\hat{\boldsymbol{\eta}}_k)$
replace $\boldsymbol{\beta}_k$ with $\hat{\boldsymbol{\beta}}_k$
$l^{**}(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}) = \sum_{j=1}^{2}\sum_{i=1}^{n}\left[R_i\delta_{ij} + \sum_{k=1}^{2}\delta^*_{ik}\times(1-R_i)\times p_{jk}(\hat{\boldsymbol{\eta}}_k;\boldsymbol{Z}_i)\right]\log\lambda_j(t_i;\boldsymbol{\theta}_j,\boldsymbol{X}_i) - \sum_{j=1}^{2}\sum_{i=1}^{n}\int_0^{t_i}\lambda_j(u;\boldsymbol{\theta}_j,\boldsymbol{Z}_i)du$
use log pseudo-likelihood to obtain
$\hat{\boldsymbol{\theta}}_k,\, V(\hat{\boldsymbol{\theta}}_k)$

Figure 4.7: Predictive-value based pseudo-likelihood approach for modeling cause-specific hazards.

Lastly, the cause-specific hazards of death and disengagement were modelled under the assumption that there was no death misclassification. These analyses were referred to as unadjusted analyses. Hypothesis tests were performed using two-sided tests at the 0.05 alpha-level. Statistical analyses were performed using R version 3.4.1 and SAS version 9.4.

4.4 Results

The characteristics patients who initiated anti-retroviral therapy at AMPATH and FACES were summarized as shown in Table 4.2. AMPATH contributed 63,890 patients between 2001 to 2011. Of these, 32711(51.2%) were administratively censored, 3493 (5.5%) died, and 27686 (43.3%) were deemed to be disengaged from care. FACES contributed 12,043 patients in the period spanning from 2007 to 2014. Of the 12,043 patients, 6483 (53.8%)were administratively censored (that is, were alive and in care when the study ended), 303 (2.5) died, and 5257(43.7%) were deemed to be disengaged from care. With the competing events, death and disengagement, pooled into a a composite event, the median survival time at AMPATH was 32.3months (95% C.I., 31.7-32.8 months), and the median survival time at FACES was 19.0 months(95% C.I., 18.2-19.7 months).

Among the 27686 patients initially classified as disengaged at AMPATH, 15.3% (4238 of 27686) were double-sampled. Through double-sampling an additional 1143 death cases were discovered at AMPATH. In other words, 24.7% of the deaths were initially classified as disengaged from care. FACES, on the other hand, double-sampled 2.3% (122 of 5257) of those who were initially deemed to be disengaged from care. The validation sample at FACES was considered too small, and therefore not used for misclassification-remedial purposes. In order to model cause-specific hazards, at FACES, while adjusting for death misclassification, I relied on death misclassification information from AMPATH.

		East Africa IeDEA Program		
Variable	Total, N=75933(%)	AMPATH, N=63890(%)	FACES, N=12043(%)	P Value
Gender				0.003
Female	50016 (65.9)	41944 (65.7)	8072 (67.0)	
Male	25917 (34.1)	21946 (34.3)	3971 (33.0)	
Age at ART initiation				<.001
Mean (SD)	37.4 (10.0)	38.2 (9.8)	32.9 (9.7)	
Median (IQR)	36.1 (30.1 – 43.5)	37.0(31.1 – 44.3)	30.9 (25.9 – 38.0)	
Time contributed to the study				<.001
Mean (SD)	21.6 (19.8)	23.1 (20.7)	13.7 (11.8)	
Median (IQR)	15.1 (5.6 – 31.6)	16.5 (6.0 – 34.3)	9.7 (4.0 – 20.1)	
CD4 at ART initiation				<.001
Mean (SD)	218.7 (172.2)	200.3 (149.5)	316.5 (238.7)	
Median (IQR)	188.0(105.0 – 282.1)	177.0 (100.3 – 259.4)	285.0 (158.0 – 417.0)	
Weight at ART initiation				<.001
Mean (SD)	56.4 (10.7)	56.0 (10.6)	58.7 (10.9)	
Median (IQR)	55.5 (49.0 - 62.0)	55.0 (49.0 – 62.0)	58.0 (51.5 – 65.0)	
Height at ART initiation				<.001
Mean (SD)	165.9 (8.3)	165.8 (8.3)	166.5 (8.3)	
Median (IQR)	165.0 (160.0 – 171.5)	165.0 (160.0- 171.5)	166.0 (160.6 - 172.0)	
WHO Stage at ART initiation				<.001
1	20999 (27.7)	16662 (26.1)	4337 (36.0)	
2	17992 (23.7)	14086 (22.0)	3906 (32.4)	
3	29852 (39.3)	26697 (41.8)	3155 (26.2)	
4	7090 (9.34)	6445 (10.1)	645 (5.4)	
Observed Cause Of Failure				<.001
Censored	39194 (51.6)	32711 (51.2)	6483 (53.8)	
Observed Death	3796 (5.0)	3493 (5.5)	303 (2.52)	
Observed Loss to Clinic	32943 (43.4)	27686 (43.3)	5257 (43.7)	
True Cause of Failure				<.001
Censored	39194 (51.6)	32711 (51.2)	6483 (53.8)	
Confirmed Death	4987 (6.6)	4636 (7.3)	351 (2.9)	
Confirmed Lost to Clinic	3169 (4.2)	3095 (4.8)	74 (0.6)	
Missing	28583 (37.6)	23448 (36.7)	5135 (42.6)	
Care Facility				<.001
Hospital	53328 (70.2)	46444 (72.7)	6884 (57.2)	
Clinic	22605 (29.8)	17446 (27.3)	5159 (42.8)	
Setting				<.001
Rural	31267 (41.2)	27798 (43.5)	3469 (28.8)	
Urban	44666 (58.8)	36092 (56.5)	8574 (71.2)	

Table 4.2: Characteristics of patients at AMPATH and FACES.

4.4.1 Predictive value and misclassification models at AMPATH

Using the internal-validation sample from AMPATH, a predictive model of death was fit using data from those who were initially observed as disengaged from care. The minimal goal in this modeling exercise was to find a predictive value model that fit the data well, without regard to model interpretability. The first model identified to fit the data well was summarized as shown in Table 4.3. The p-value for the Supremum goodness-of-fit test associated with this model was 0.0594. This model was then used in two seperate tasks. First, to model the cause-specific hazards of death and disengagement from care at AMPATH, while treating as missing, the true outcome data among the unvalidated disengers. Secondly, to model the death misclassification probabilities at AMPATH using the pseudo-likelihood approach presented in Chapter 2. In both modeling instances, the missing true outcomes among the unvalidated disengagers were replaced with estimated predictive values, assuming outcome data were missing at random (MAR). The results of modeling the log-odds of death misclassfication given the subject characteristics were presented in Table 4.4.

	Term	Estimate	SE	Z	$Pr[>\|Z\|]$
1	(Intercept)	11.884	0.859	13.827	0.000
2	$I[\text{Male} = 1]$	-0.419	0.220	-1.905	0.057
3	Centered Age	0.039	0.004	9.552	0.000
4	$\sqrt{\text{CD4}}$	-0.220	0.028	-7.823	0.000
5	LogTime	-1.612	0.142	-11.371	0.000
6	LogBMI	-2.836	0.256	-11.075	0.000
7	WHO stage 2 vs 1	0.449	0.146	3.082	0.002
8	WHO stage 3 vs 1	0.599	0.132	4.540	0.000
9	WHO stage 4 vs 1	1.268	0.154	8.248	0.000
10	Clinic vs Hospital	-0.778	0.119	-6.523	0.000
11	Urban vs Rural	-0.431	0.101	-4.281	0.000
12	$I[\text{Male} = 1] \times \sqrt{\text{CD4}}$	0.033	0.018	1.892	0.058
13	LogTime × $\sqrt{\text{CD4}}$	0.065	0.011	5.776	0.000

Table 4.3: Predictive value model of true event being death given disengagement from care is observed

	Term	Estimate	SE	Z	$Pr[>\|Z\|]$
1	(Intercept)	0.069	0.396	0.173	0.862
2	Male vs Female	-0.174	0.060	-2.913	0.004
3	Centered Age	0.005	0.003	1.902	0.057
4	$\sqrt{\text{CD4}}$	0.026	0.007	3.941	0.000
5	$\sqrt{\text{BMI}}$	0.037	0.087	0.429	0.668
6	$\sqrt{\text{Study time}}$	0.042	0.020	2.072	0.038
7	WHO stage 2 vs 1	0.182	0.129	1.408	0.159
8	WHO stage 3 vs 1	0.012	0.117	0.100	0.921
9	WHO stage 4 vs 1	0.040	0.125	0.320	0.749
10	Clinic vs Hospital	-0.342	0.083	-4.102	0.000
11	Urban vs Rural	0.367	0.071	5.203	0.000

Table 4.4: Model for the probability of observing disengagement from care when individual is actually dead.

At the 0.05 alpha level, the misclassification model suggested that males had lower odds of death misclassification than females, after adjusting for other model covariates. Model also suggested that those who were treated at clinics (health centers) had lower odds of death misclassification than those treated at larger hospitals, holding constant the other model covariates. In addition, the model suggested that those with higher CD4 count, contributed more study time or based in urban versus rural areas had higher odds of being classified as disengaged when, in fact, dead. Lastly, there was not sufficient evidence at 0.05 alpha level to support an association between death misclassification and the covariates BMI and WHO stage.

The death misclassification model as shown in Table 4.4 was then used to compute misclassification probabilities for the study sample. Since there was no misclassification among those who were initially observed as dead, the probability of observing death among true disengagers was zero; that is, $P[C^* = 1|C = 2, \boldsymbol{Z}] = 0$.

4.4.2 Cause-specific hazards at AMPATH

The cause-specific hazards of death and disengagement from care at AMPATH were modeled in three ways. In the first approach, I ignored misclassification; in the second approach, I adjusted for death misclassification using the predictive values calculated from the model in Table 4.3, and in third approach, I adjusted for misclassification using misclassification probabilities calculated from the model in Table 4.4. The second and third model approached relied on maximization of pseudo-log-likelihood from log-likelihood 4.2 and 4.3 respectively. The model results were as shown in Table 4.5. The scale, shape and hazard-ratio estimates from the three modeling approaches were also compared visually as shown in Figure 4.8,4.9, and 4.10 and 4.11 respectively.

After adjusting for death misclassification, the shape parameter of the baseline cause-specific hazard of death increased from 0.584 (95% C.I., 0.567-0.601) regardless of whether I adjusted using predictive values or misclassification probabilities. When using a predictive-value adjustment, the shape-parameter estimate was 0.668(95% C.I., 0.656 - 0.680), and when using misclassification probabilities to adjust, the shape-parameter estimate was 0.731(95% C.I., 0.721-0.742). Although the shape-parameter estimates from the two adjustments were different, they both suggested that the hazard of death was decreasing a rate lower than when do not consider death misclassification. In contrast, the two adjustments resulted in shape parameters with different interpretations when modeling the cause-specific hazards of disengagement. The shape-parameter estimate after predictive-value adjustment suggested that the risk of disengagement increased with time. The opposite was found after adjusting using misclassification probabilities. That said, as shown in Table 4.5, the log-hazard ratios of death and disengagement after either adjustment were not very different in magnitude.

At the 0.05 alpha level, the models for the cause-specific hazard of death after adjusting for death misclassification, suggested that the hazard of death was higher in males, among the older, and among those who had a higher WHO stage at ART initiation. Models also suggested that the risk of death was lower in those who had higher CD4 count, higher BMI, received care at clinics versus hospitals, and received care in urban areas. I also found that the hazard of disengagement was higher among males, those with higher CD4 count, those with higher BMI, those who received care at clinics, and those who received care in urban settings. The models also suggested that older patients had a lower hazard of disengagement than younger patients. WHO stage did not have a statistically significant association with the cause-specific hazard of disengagement.

		Naïve				Predictive Value Adj.				Misclassification Probabilty Adj.			
				95% CI				95% CI				95% CI	
Event	**Covariate**	Estimate	SE	Lower	Upper	Estimate	SE	Lower	Upper	Estimate	SE	Lower	Upper
Death	Shape (alpha1)	0.584	0.009	0.567	0.601	0.668	0.006	0.656	0.680	0.731	0.005	0.721	0.742
	Intercept(Scale)	-1.200	0.142	-1.477	-0.923	-0.673	0.155	-0.977	-0.369	-0.699	0.119	-0.932	-0.466
	Sex(Male vs Female)	0.251	0.035	0.182	0.319	0.145	0.036	0.075	0.215	0.140	0.029	0.084	0.197
	Centered Age	0.005	0.002	0.001	0.008	0.009	0.002	0.006	0.013	0.012	0.002	0.009	0.015
	Root cd4	-0.090	0.004	-0.098	-0.083	-0.071	0.004	-0.079	-0.063	-0.077	0.003	-0.083	-0.071
	BMI	-0.142	0.006	-0.154	-0.130	-0.138	0.007	-0.152	-0.125	-0.151	0.005	-0.161	-0.140
	WHO Stage												
	2 vs 1	0.290	0.070	0.152	0.428	0.387	0.078	0.234	0.541	0.441	0.056	0.332	0.549
	3 vs 1	0.630	0.062	0.509	0.751	0.611	0.070	0.473	0.749	0.661	0.049	0.565	0.757
	4 vs 1	1.150	0.069	1.015	1.285	1.171	0.077	1.021	1.321	1.141	0.056	1.032	1.251
	Clinic vs Hospital	-0.110	0.040	-0.188	-0.032	-0.334	0.049	-0.429	-0.238	-0.240	0.042	-0.322	-0.157
	Urban vs Rural	-0.460	0.037	-0.532	-0.389	-0.214	0.043	-0.298	-0.131	-0.152	0.036	-0.222	-0.081
Disengagement	Shape (alpha2)	0.931	0.005	0.922	0.940	1.035	0.007	1.022	1.048	0.973	0.004	0.965	0.981
	Intercept (Scale)	-3.440	0.049	-3.535	-3.345	-5.052	0.083	-5.214	-4.889	-4.844	0.071	-4.983	-4.704
	Sex(Male vs Female)	0.089	0.013	0.064	0.115	0.105	0.023	0.059	0.151	0.110	0.019	0.073	0.148
	Centered Age	-0.016	0.001	-0.018	-0.015	-0.027	0.001	-0.029	-0.024	-0.028	0.001	-0.030	-0.026
	Root cd4	-0.011	0.001	-0.013	-0.008	0.006	0.002	0.002	0.011	0.007	0.002	0.004	0.011
	BMI	-0.017	0.002	-0.021	-0.013	0.017	0.003	0.011	0.023	0.019	0.003	0.014	0.024
	WHO Stage												
	2 vs 1	-0.020	0.018	-0.056	0.016	-0.056	0.028	-0.112	-0.001	-0.068	0.023	-0.114	-0.023
	3 vs 1	0.070	0.016	0.039	0.102	-0.011	0.025	-0.060	0.038	-0.024	0.020	-0.064	0.016
	4 vs 1	0.269	0.022	0.225	0.313	-0.061	0.045	-0.150	0.028	0.001	0.034	-0.065	0.068
	Clinic vs Hospital	0.043	0.015	0.014	0.072	0.213	0.027	0.159	0.267	0.160	0.024	0.114	0.206
	Urban vs Rural	0.148	0.013	0.122	0.174	0.229	0.027	0.177	0.281	0.191	0.022	0.149	0.234

Table 4.5: Cause-specific hazard models at AMPATH. In first instance, I ignored death misclassification; in the second instance, I treated the problem as a missing data problem, and in the third instance, I treated the problem as a misclassification problem.

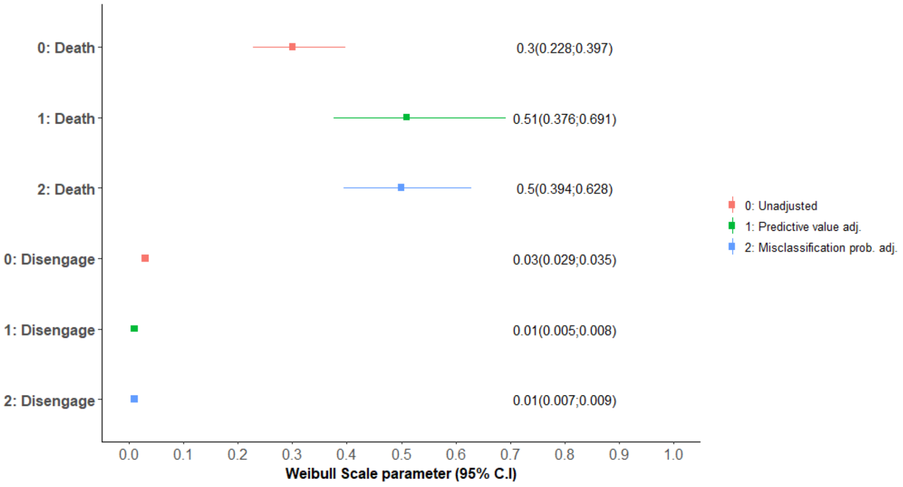

Figure 4.8: Scale parameters associated with the baseline cause-specific hazards at AMPATH.

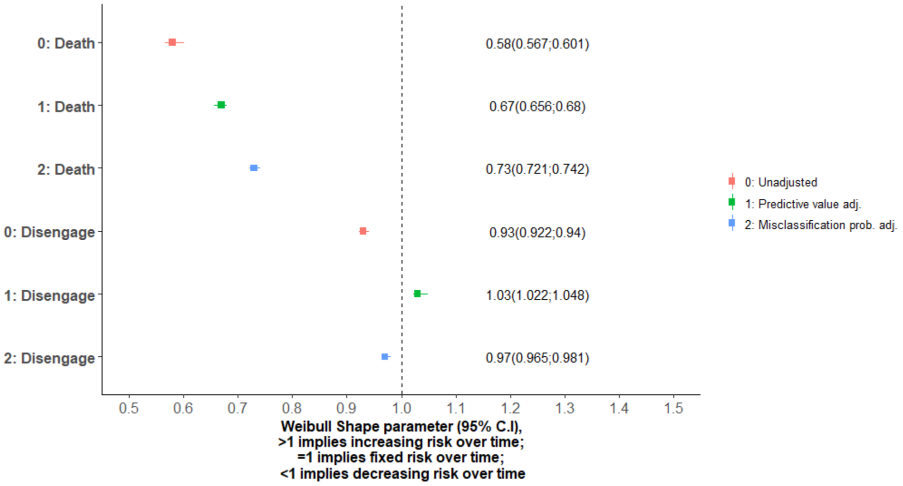

Figure 4.9: Shape parameters associated with the baseline cause-specific hazards at AMPATH.

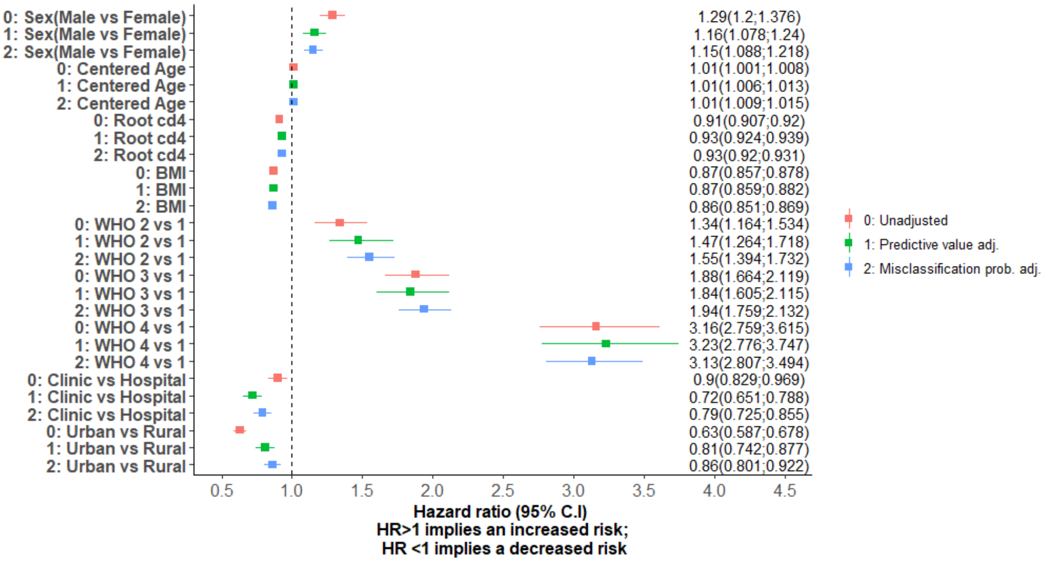

Figure 4.10: Hazard ratios representing the effects of covariates on the cause-specific hazard of death at AMPATH. The forest plot captures results from three scenarios. In scenario 1 (pink), cause-specific hazard of death was modeled without considering death misclassification. Scenerio 2 (green) adjusted for misclassification using predictive values. Scenario 3 (blue) adjusted for death misclassification using misclassification probabilities.

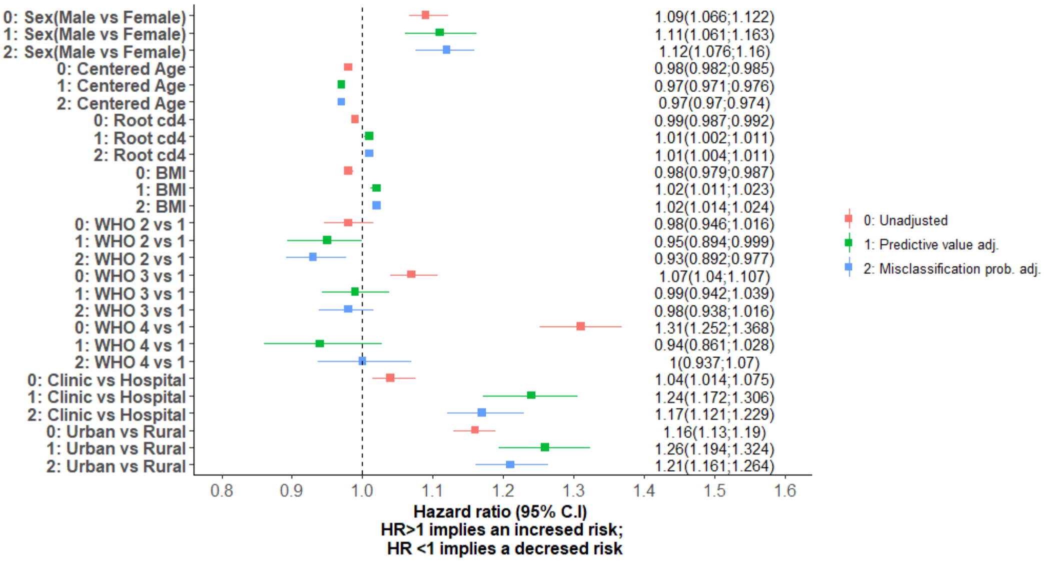

Figure 4.11: Hazard ratios representing the effects of covariates on the cause-specific hazard of disengagement from care at AMPATH.

4.4.3 Cause specific hazards at FACES

Under assumptions similar to AMPATH, I modeled the cause-specific hazards of death and disengagement from care FACES. However, unlike AMPATH I did not have sufficient validation data at FACES to quantify the extent of misclassification. As a result, the cause-specific hazards models were adjusted for misclassification using misclassification probabalities from AMPATH. The borrowing of information from AMPATH was done assuming the *transportability* of misclassification. Cause-specific hazards were also modeled ignoring death misclassification. The model results for aforementioned scenarios are presented in Table 4.6. The model results were also presented in forest plots as shown by Figures 4.12, 4.13 and 4.14.

		Unadjusted				**Adjusted**			
				95% CI				95% CI	
Event	**Covariate**	Estimate	SE	Lower	Upper	Estimate	SE	Lower	Upper
Death	Shape (alpha1)	0.567	0.029	0.510	0.623	0.860	0.031	0.800	0.920
	Intercept(Scale)	-1.340	0.473	-2.266	-0.414	-0.969	0.409	-1.770	-0.168
	Sex(Male vs Female)	0.190	0.120	-0.045	0.426	0.140	0.096	-0.049	0.329
	Centered Age	0.015	0.006	0.004	0.027	0.020	0.005	0.011	0.030
	Root cd4	-0.082	0.010	-0.102	-0.062	-0.054	0.009	-0.072	-0.036
	BMI	-0.179	0.021	-0.221	-0.137	-0.177	0.018	-0.213	-0.142
	WHO Stage								
	2 vs 1	0.512	0.230	0.061	0.963	0.600	0.223	0.162	1.037
	3 vs 1	1.199	0.218	0.773	1.625	1.199	0.239	0.730	1.668
	4 vs 1	1.799	0.242	1.325	2.273	1.590	0.261	1.079	2.101
	Clinic vs Hospital	0.060	0.138	-0.210	0.331	0.060	0.113	-0.161	0.281
	Urban vs Rural	-0.090	0.156	-0.396	0.215	-0.190	0.125	-0.434	0.055
Disengagement	Shape (alpha2)	1.091	0.012	1.068	1.115	1.123	0.015	1.094	1.151
	Intercept (Scale)	-3.011	0.106	-3.220	-2.802	-3.980	0.197	-4.366	-3.594
	Sex(Male vs Female)	-0.045	0.032	-0.108	0.017	-0.070	0.047	-0.161	0.022
	Centered Age	-0.019	0.002	-0.022	-0.015	-0.030	0.003	-0.036	-0.024
	Root cd4	-0.011	0.002	-0.015	-0.006	-0.002	0.003	-0.008	0.005
	BMI	-0.020	0.004	-0.029	-0.012	0.008	0.007	-0.005	0.022
	WHO Stage								
	2 vs 1	0.118	0.036	0.048	0.188	0.100	0.045	0.012	0.188
	3 vs 1	0.296	0.037	0.223	0.369	0.150	0.060	0.033	0.267
	4 vs 1	0.402	0.062	0.281	0.522	0.110	0.112	-0.110	0.330
	Clinic vs Hospital	-0.191	0.034	-0.258	-0.124	-0.261	0.049	-0.356	-0.166
	Urban vs Rural	-0.209	0.037	-0.281	-0.137	-0.201	0.049	-0.296	-0.105

Table 4.6: Cause-specific hazards models at FACES. First model is not adjusted for death misclassification, and the second model is adjusted for death misclassification.

When I adjusted for death misclassification, the shape-parameter of the baseline cause-specific hazard of death increased from 0.57 (95% C.I.: 0.51-0.623) to 0.86 (95% C.I.: 0.80-0.92). Both the adjusted and unadjusted shape parameters suggested that the risk of death decreased over time. That said, the rate of decline was slower after adjusting for death misclassification. The shape parameter for the baseline cause-specific hazard of disengagement did not change by much from adjusting for death misclassification: The estimate changed from 1.091(95% C.I., 1.068-1.115) to 1.123 (95% C.I., 1.094-1.151). These results suggested that the risk of disengagement increased with time.

For all the covariates considered, as shown in Table 4.6, the hazard ratios of death did not change by much after adjusting for death misclassification. This observation held both in terms of the magnitude and the interpretability of the coefficients. A comparison of the hazard-ratio estimates before and after adjusting for misclassification is presented in Figure 4.12. At the 0.05 alpha level, the misclassification-adjusted model for the cause-specific hazard of death suggested a higher risk of death among older patients, and those with higher WHO stage classification at ART initiation. Model also suggested a lower risk of death with increasing CD4 count, and BMI at ART initiation.

Adjusting for death-misclassification also did not change log-hazard ratios of disengagement by very much as shown in Table 4.6. That being said, adjusting for death-misclassification did alter some statistical relationships between the covariates and the cause-specific hazard of disengagement. For example, at the 0.05 alpha level, the unadjusted model suggested a negative relationship between CD4 count and the hazard of disengagement from care. The relationship remained negative after adjusting for death misclassification, however, there was insufficent evidence to support this relationship at the 0.05 alpha level.

A visual comparison of the changes in hazards ratios of disengagement from care, before and after adjusting for death misclassification, is presented in Figure 4.12.

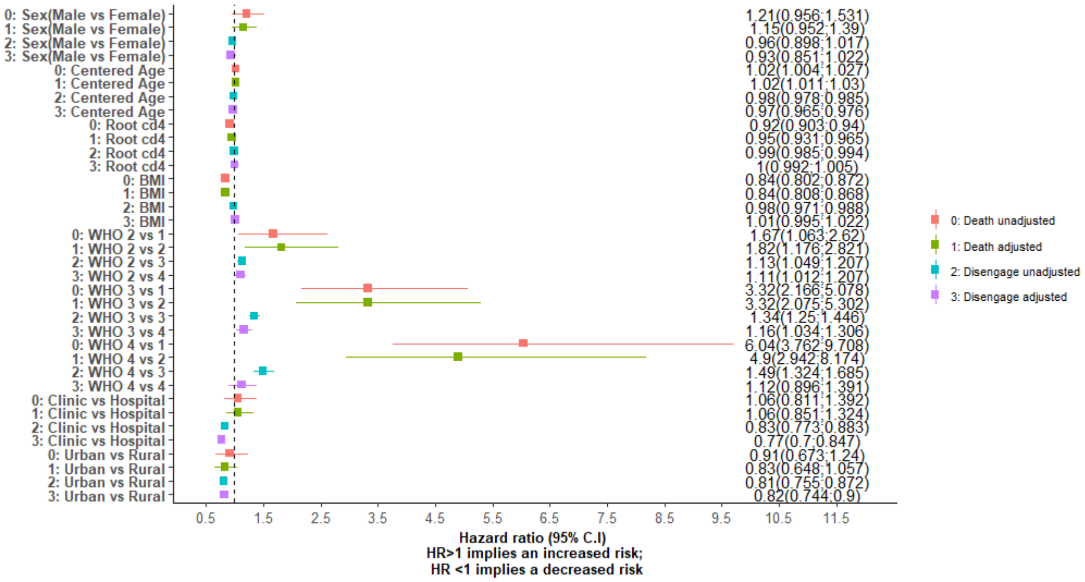

Figure 4.12: Hazard ratios representing the effects of covariates on the cause-specific hazard of death and disengagement from care at FACES.

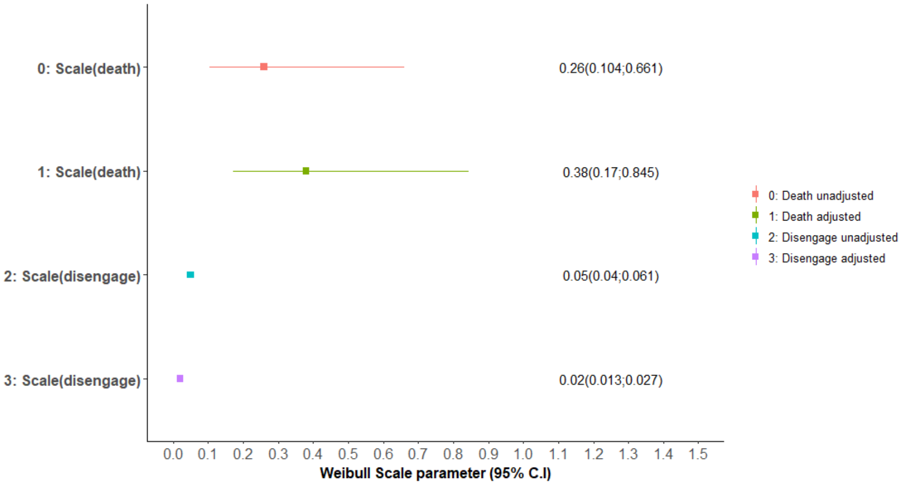

Figure 4.13: Scale parameters associated with baseline cause-specific hazards at FACES.

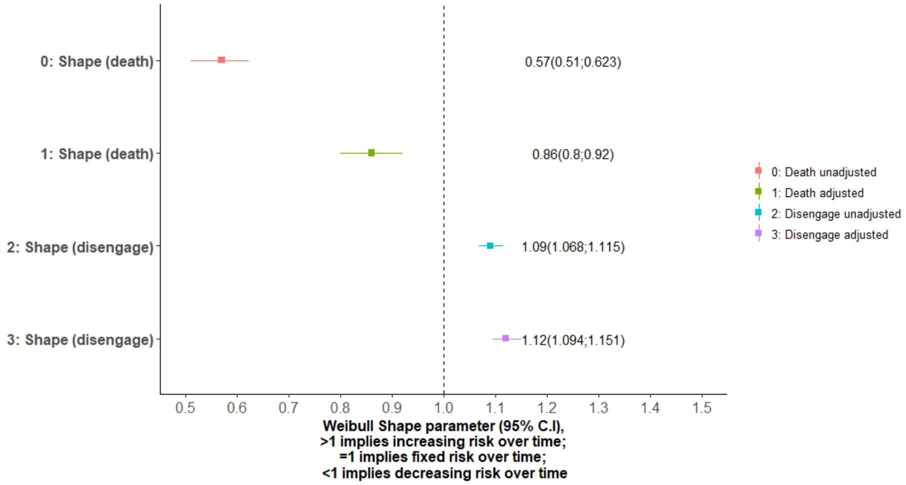

Figure 4.14: Shape parameters associated with baseline cause-specific hazards at FACES.

4.5 Discussion

This chapter revisits the problem of outcome misclassification in studies with competing risks. It is well known that such misclassification can result in biased estimation and statistical inference. That being said, the examples of how to deal with this misclassification are sparse in epidemiologic literature. This is especially true when it comes to parametric methods. In this chapter, I set out to provide a gentle and yet comprehensive overview of how one may remedy for outcome misclassification when modeling cause-specific hazards in a parametric fashion. Specifically, I presented remedies that rely on internal-and external-validation sampling to identify the extent of misclassification. I presented validation sampling as a viable solution because validation sampling provides misclassification information, which in turn can be used to adjust cause-specific hazard models. In this exploration, I noted that the misclassification information is usually packaged as predictive values or misclassification probabilities (that is complements of sensitivities and specifities), which also influences the statistical methods used when modeling cause-specific hazards. I also noted that predictive values are used when one desires to use missing data methods, and misclassification probabilities are used when one wants to treat the problem as a misclassification problem. Whichever method one chooses, I presented pseudo-likelihoods that be used to perform parameter estimation.

Pseudo-likelihood based estimation is appealing because it is intuitive and simple to implement. In fact, implementation resembles maximum likelihood estimation, with the major difference being in the variance estimation. When using pseudo-likelihood estimation, one needs to also contend with the variability from estimating the components that are plugged into the true likelihood. In modeling cause-specific hazards using pseudo-likelihod approaches that I present, one has to contend with the additional variability from modeling predictive values or misclassification probabilities. Mpofu et al. (2019) derived the closed-variance estimator for a situation where the misclassification probabilities used in forming the pseudo-likelihood are derived from an external setting or study. In the case where the misclassification probabilities or predictive values used are derived internal to the study, the derivation of a closed-form variance estimator is also fairly simple. Moreover, the whole process of obtaining pseudo-likelihood estimates and their variance estimates is simple to perform within the `R` software.

As an illustration, I modeled the cause-specific hazards of death and disengagement from care among PLWH who contributed data to two treatment programs at IeDEA East Africa, namely AMPATH and FACES. I assumed that the baseline cause-specific hazards were of a Weibull form, and also assumed proportional hazards relationships between the covariates and the hazard functions. The challenge in the motivating example was that some of the patients who were observed as disengaged from care were actually dead. I, therefore, had to contend with death misclassification when modeling the cause-specific hazards. AMPATH had a sufficiently large validation sample among those who disengaged from care, so misclassification information from AMPATH was used to adjust the cause-specific hazards models at both AMPATH and FACES. The misclassification probabilities from AMPATH were used in FACES assuming the *transportability* of misclassification. A comparision of the cause-specific hazard models at AMPATH and FACES is illustrated in Figures 4.15, 4.16, 4.18 and 4.17. Of note is that the hazard ratio estimates for death were fairly consisent across the two programs. There was, however, more variability in the hazards ratio estimates for disengagement. At both AMPATH and FACES, the (baseline cause-specific hazard) shape-parameter estimates suggested a decline in the risk of death over time, with the rate of decline being slower at FACES. On the other hand, the shape parameters associated with disengagement suggested an increase in the risk of disengagement, over time, at FACES, and a decreased risk, over time, at AMPATH. Also discernable in the model results from FACES and AMPATH is that the standard errors for point estimates at AMPATH were smaller than those at FACES. This observation is explained by the fact that AMPATH had a larger sample size, and that the misclassification adjustment at FACES relied on the transfer of misclassification information from AMPATH. If the observed differences the cause-specific hazard models at AMPATH and FACES are true, they can serve as a cautionary note to practitioners, who may try to transport/transfer predictive values across different settings. That said, I should concede that the model results from FACES must be viewed with some skepticism as they depend on the *transportability* of misclassification probabilities. The validity of this assumption is not empirically testable (Spiegelman 2010). As a result, when presenting the cause-specific hazard models to stakeholders I recommend that the analyst present both results from before and after adjusting for misclassification.

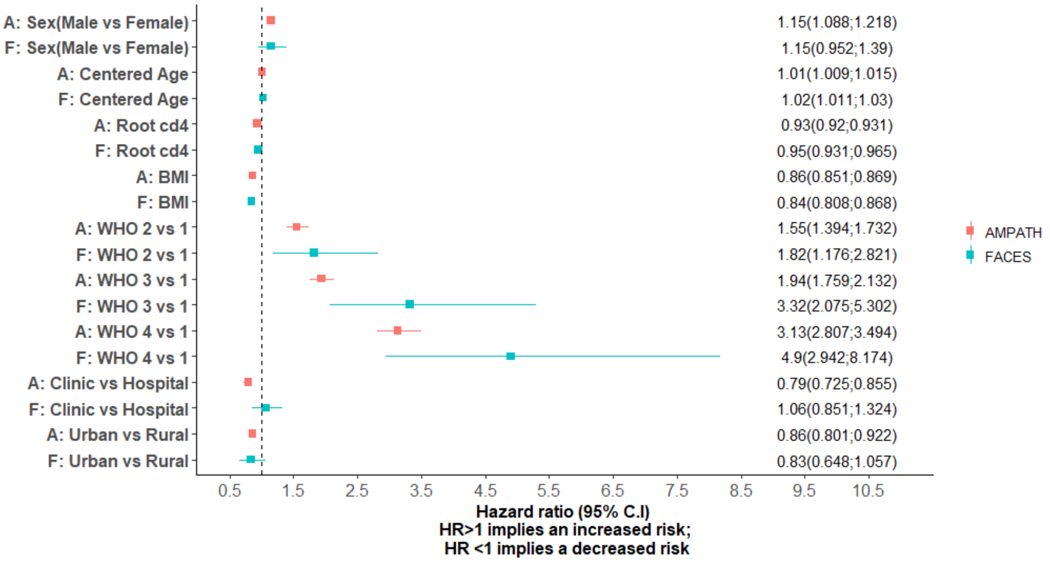

Figure 4.15: Comparing the hazard ratios of death at AMPATH and FACES, after adjusting for death misclassification.

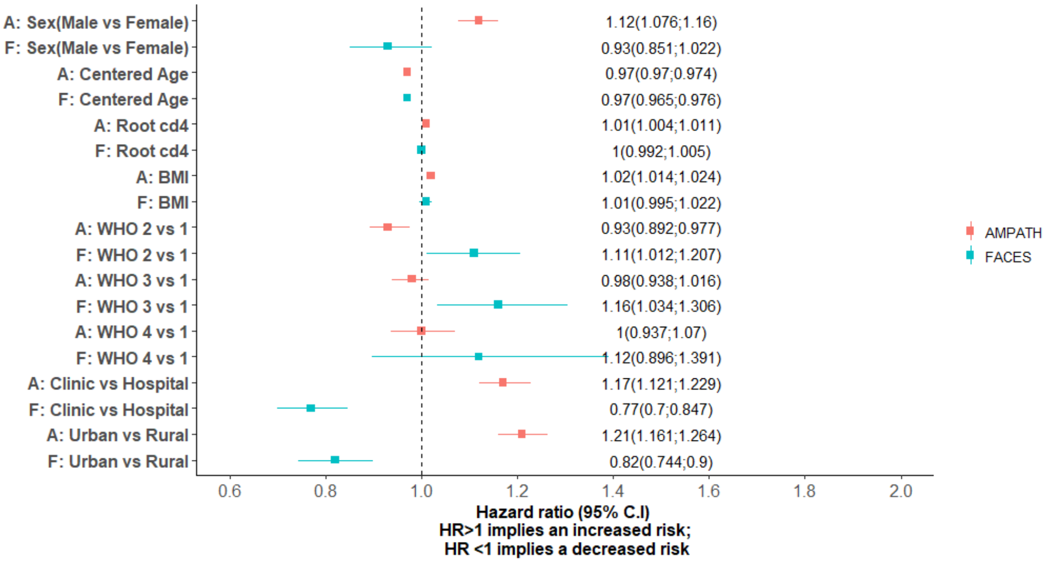

Figure 4.16: Comparing the hazard ratios of disengagement at AMPATH and FACES, after adjusting for death misclassification.

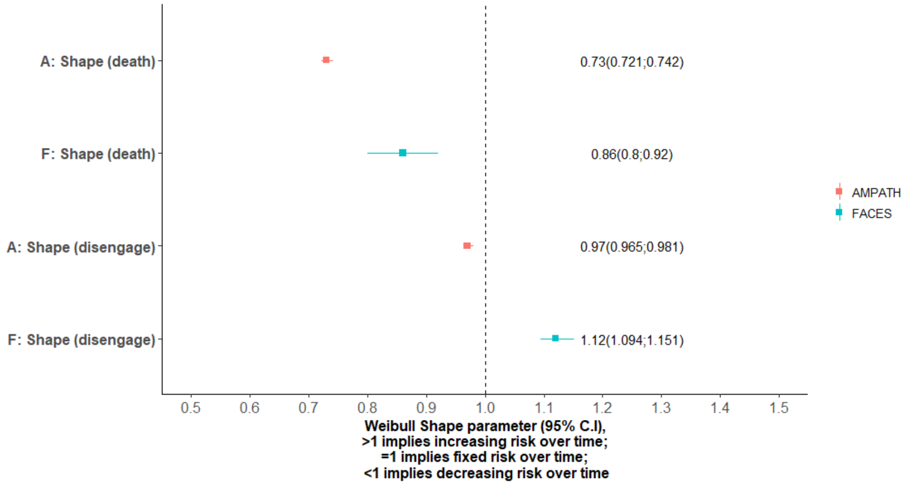

Figure 4.17: Shape parameters associated with baseline cause-specific hazards at AMPATH and FACES.

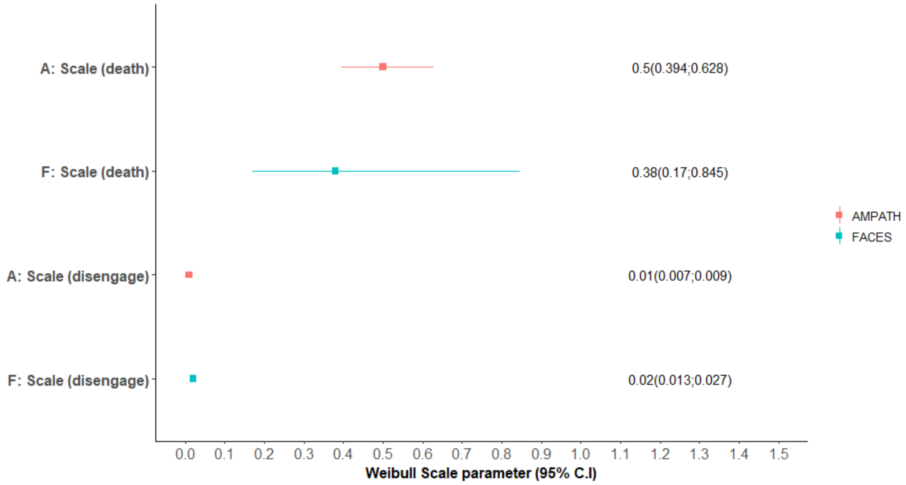

Figure 4.18: Scale parameters associated with baseline cause-specific hazards at AMPATH and FACES.

The work of developing parametric methods for modeling cause-specific hazards, in the presence of outcome misclassification, is far from over. The next step in this work may be to develop model goodness-of-fit methods which are crucial in parametric modeling. At the moment, I am exploring visual assessments with semi-parametric versions of the models used as the references. Setting aside the future plans, thus far, I am hopeful that I have presented an accessible exposition to validation sampling-based methods for dealing with outcome misclassification when modeling cause-specific hazards in a parametric fashion.

REFERENCES

Aalen, Odd, Ornulf Borgan, and Hakon Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. Book. Springer Science & Business Media.

An, M. W., C. E. Frangakis, B. S. Musick, and C. T. Yiannoutsos. 2009. "The Need for Double-Sampling Designs in Survival Studies: An Application to Monitor Pepfar." Journal Article. *Biometrics* 65 (1): 301–6.

Austin, P. C., D. S. Lee, and J. P. Fine. 2016. "Introduction to the Analysis of Survival Data in the Presence of Competing Risks." Journal Article. *Circulation* 133 (6): 601–9.

Bakoyannis, G., and C. T. Yiannoutsos. 2015. "Impact of and Correction for Outcome Misclassification in Cumulative Incidence Estimation." Journal Article. *PLoS One* 10 (9).

Bakoyannis, G., F. Siannis, and G. Touloumi. 2010. "Modelling Competing Risks Data with Missing Cause of Failure." Journal Article. *Stat Med* 29 (30): 3172–85.

Bakoyannis, G., Y. Zhang, and C. T. Yiannoutsos. 2018. "Semiparametric Analysis of Competing Risks Data Under Missing Cause of Failure." Journal Article. *arXiv*.

Barron, B. A. 1977. "The Effects of Misclassification on the Estimation of Relative Risk." Journal Article. *Biometrics* 33 (2): 414–8.

Beyersmann, J., A. Latouche, A. Buchholz, and M. Schumacher. 2009. "Simulating Competing Risks Data in Survival Analysis." Journal Article. *Stat Med* 28 (6): 956–71.

Brinkhof, M. W., B. D. Spycher, C. Yiannoutsos, R. Weigel, R. Wood, E. Messou, A. Boulle, M. Egger, J. A. Sterne, and Aids International epidemiological Database to Evaluate. 2010. "Adjusting Mortality for Loss to Follow-up: Analysis of Five Art Programmes in Sub-Saharan Africa." Journal Article. *PLoS One* 5 (11).

Bross, Irwin. 1954. "Misclassification in 2 X 2 Tables." Journal Article. *Biometrics* 10 (4): 478–86.

Carpenter, James, and Michael Kenward. 2012. *Multiple Imputation and Its Application.* Book. John Wiley & Sons.

Carroll, Raymond J, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* Book. CRC press.

Chen, Yi-Hau. 2000. "Miscellanea. a Robust Imputation Method for Surrogate Outcome Data." Journal Article. *Biometrika* 87 (3): 711–16.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin. 1977. "Maximum Likelihood from Incomplete Data via the Em Algorithm." Journal Article. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Ebrahimi, N. 1996. "The Effects of Misclassification of the Actual Cause of Death in Competing Risks Analysis." Journal Article. *Stat Med* 15 (14): 1557–66.

Edwards, J. K., S. R. Cole, M. A. Troester, and D. B. Richardson. 2013. "Accounting for Misclassified Outcomes in Binary Regression Models Using Multiple Imputation with Internal Validation Data." Journal Article. *Am J Epidemiol* 177 (9): 904–12.

Egger, M., D. K. Ekouevi, C. Williams, R. E. Lyamuya, H. Mukumbi, P. Braitstein, T. Hartwell, et al. 2012. "Cohort Profile: The International Epidemiological Databases to Evaluate Aids (Iedea) in Sub-Saharan Africa." Journal Article. *Int J Epidemiol* 41 (5): 1256–64.

Egger, M., B. D. Spycher, J. Sidle, R. Weigel, E. H. Geng, M. P. Fox, P. MacPhail, et al. 2011. "Correcting Mortality for Loss to Follow-up: A Nomogram Applied to Antiretroviral Treatment Programmes in Sub-Saharan Africa." Journal Article. *PLoS Med* 8 (1).

Geng, E. H., N. Emenyonu, M. B. Bwana, D. V. Glidden, and J. N. Martin. 2008. "Sampling-Based Approach to Determining Outcomes of Patients Lost to Follow-up in Antiretroviral Therapy Scale-up Programs in Africa." Journal Article. *JAMA* 300 (5): 506–7.

Geng, Elvin H, David V Glidden, Mwebesa Bosco Bwana, Nicolas Musinguzi, Nneka Emenyonu, Winnie Muyindike, Katerina A Christopoulos, Torsten B Neilands, Constantin T Yiannoutsos, and Steven G Deeks. 2011. "Retention in Care and Connection to Care Among Hiv-Infected Patients on Antiretroviral Therapy in Africa: Estimation via a Sampling-Based Approach." Journal Article 6 (7).

Gong, Gail, and Francisco J Samaniego. 1981. "Pseudo Maximum Likelihood Estimation: Theory and Applications." Journal Article. *The Annals of Statistics*, 861–69.

Grace, Y Yi. 2016. *Statistical Analysis with Measurement Error or Misclassification*. Book. Springer.

Gravel, C. A., A. Dewanji, P. J. Farrell, and D. Krewski. 2018. "A Validation Sampling Approach for Consistent Estimation of Adverse Drug Reaction Risk with Misclassified Right-Censored Survival Data." Journal Article. *Stat Med* 37 (27): 3887–3903.

Greenland, S. 1988. "Variance Estimation for Epidemiologic Effect Estimates Under Misclassification." Journal Article. *Stat Med* 7 (7): 745–57.

Ha, J., and A. Tsodikov. 2012. "Isotonic Estimation of Survival Under a Misattribution of Cause of Death." Journal Article. *Lifetime Data Anal* 18 (1): 58–79.

Hardt, J., M. Herke, and R. Leonhart. 2012. "Auxiliary Variables in Multiple Imputation in Regression with Missing X: A Warning Against Including Too Many in Small Sample Research." Journal Article. *BMC Med Res Methodol* 12 (1): 184.

Hinchliffe, S. R., K. R. Abrams, and P. C. Lambert. 2013. "The Impact of Under and over Recording of Cancer on Death Certificates in a Competing Risks Analysis: A Simulation Study." Journal Article. *Cancer Epidemiol* 37 (1): 11–19.

Justice, A. C., K. E. Covinsky, and J. A. Berlin. 1999. "Assessing the Generalizability of Prognostic Information." Journal Article. *Ann Intern Med* 130 (6): 515–24.

Kalbfleisch, John D, and Ross L Prentice. 2011. *The Statistical Analysis of Failure Time Data*. Book. Vol. 360. John Wiley & Sons.

Lin, D. Y., L. J. Wei, and Z. Ying. 2002. "Model-Checking Techniques Based on Cumulative Residuals." Journal Article. *Biometrics* 58 (1): 1–12.

Little, Roderick JA, and Donald B Rubin. 2014. *Statistical Analysis with Missing Data*. Book. Vol. 333. John Wiley & Sons.

Lu, K., and A. A. Tsiatis. 2001. "Multiple Imputation Methods for Estimating Regression Coefficients in the Competing Risks Model with Missing Cause of Failure." Journal Article. *Biometrics* 57 (4): 1191–7.

Lyles, R. H., and J. Lin. 2010. "Sensitivity Analysis for Misclassification in Logistic Regression via Likelihood Methods and Predictive Value Weighting." Journal Article. *Stat Med* 29 (22): 2297–2309.

Lyles, R. H., L. Tang, H. M. Superak, C. C. King, D. D. Celentano, Y. Lo, and J. D. Sobel. 2011. "Validation Data-Based Adjustments for Outcome Misclassification in Logistic Regression: An Illustration." Journal Article. *Epidemiology* 22 (4): 589–97.

Magder, L. S., and J. P. Hughes. 1997. "Logistic Regression When the Outcome Is Measured with Uncertainty." Journal Article. *American Journal of Epidemiology* 146 (2): 195–203.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." Journal Article. *Statistical Science*, 538–58.

Nardi, A., and M. Schemper. 2003. "Comparing Cox and Parametric Models in Clinical Studies." Journal Article. *Stat Med* 22 (23): 3597–3610.

Neuhaus, John M. 1999. "Bias and Efficiency Loss Due to Misclassified Responses in Binary Regression." Journal Article. *Biometrika* 86 (4): 843–55.

Ogden, R. T., and T. Tarpey. 2006. "Estimation in Regression Models with Externally Estimated Parameters." Journal Article. *Biostatistics* 7 (1): 115–29.

Parke, William R. 1986. "Pseudo Maximum Likelihood Estimation: The Asymptotic Distribution." Journal Article. *The Annals of Statistics*, 355–57.

Pepe, Margaret Sullivan. 1992. "Inference Using Surrogate Outcome Data and a Validation Sample." Journal Article. *Biometrika* 79 (2): 355–65.

Robins, James M, and Naisyin Wang. 2000. "Inference for Imputation Estimators." Journal Article. *Biometrika* 87 (1): 113–24.

Rosner, B., D. Spiegelman, and W. C. Willett. 1990. "Correction of Logistic Regression Relative Risk Estimates and Confidence Intervals for Measurement Error: The Case of Multiple Covariates Measured with Error." Journal Article. *Am J Epidemiol* 132 (4): 734–45.

Rubin, Donald B. 1976. "Inference and Missing Data." Journal Article. *Biometrika* 63 (3): 581–92.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." Journal Article. *Psychological Methods* 7 (2): 147.

Spiegelman, D. 2010. "Approaches to Uncertainty in Exposure Assessment in Environmental Epidemiology." Journal Article. *Annu Rev Public Health* 31 (1): 149–63.

Spiegelman, D., R. J. Carroll, and V. Kipnis. 2001. "Efficient Regression Calibration for Logistic Regression in Main Study/Internal Validation Study Designs with an Imperfect Reference Instrument." Journal Article. *Stat Med* 20 (1): 139–60.

Tang, L., R. H. Lyles, C. C. King, D. D. Celentano, and Y. Lo. 2015. "Binary Regression with Differentially Misclassified Response and Exposure Variables." Journal Article. *Stat Med* 34 (9): 1605–20.

Tenenbein, Aaron. 1970. "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications." Journal Article. *Journal of the American Statistical Association* 65 (331): 1350–61.

Tilling, Kate, Elizabeth J Williamson, Michael Spratt, Jonathan AC Sterne, and James R Carpenter. 2016. "Appropriate Inclusion of Interactions Was Needed to Avoid Bias in Multiple Imputation." Journal Article. *Journal of Clinical Epidemiology* 80: 107–15.

Touloumis, Anestis. 2016. "Simulating Correlated Binary and Multinomial Responses Under Marginal Model Specification: The Simcormultres Package." Journal Article. *The R Journal*.

Van Rompaye, B., S. Jaffar, and E. Goetghebeur. 2012. "Estimation with Cox Models: Cause-Specific Survival Analysis with Misclassified Cause of Failure." Journal Article. *Epidemiology* 23 (2): 194–202.

Wacholder, Sholom. 1996. "The Case-Control Study as Data Missing by Design: Estimating Risk Differences." Journal Article. *Epidemiology*, 144–50.

Web Page. 2019. https://www.iedea.org/regions/east-africa/.

Wu, Xiao, Danielle Braun, Marianthi-Anna Kioumourtzoglou, Christine Choirat, Qian Di, and Francesca Dominici. 2019. "Causal Inference in the Context of an Error Prone Exposure: Air Pollution and Mortality." Journal Article. *The Annals of Applied Statistics* 13 (1): 520–47.

Yiannoutsos, C. T., M. W. An, C. E. Frangakis, B. S. Musick, P. Braitstein, K. Wools-Kaloustian, D. Ochieng, et al. 2008. "Sampling-Based Approaches to Improve Estimation of Mortality Among Patient Dropouts: Experience from a Large Pepfar-Funded Program in Western Kenya." Journal Article. *PLoS One* 3 (12): e3843.

Zhang, G., and R. Little. 2009. "Extensions of the Penalized Spline of Propensity Prediction Method of Imputation." Journal Article. *Biometrics* 65 (3): 911–8.

Zhao, Y., J. F. Lawless, and D. L. McLeish. 2009. "Likelihood Methods for Regression Models with Expensive Variables Missing by Design." Journal Article. *Biom J* 51 (1): 123–36.

CURRICULUM VITAE

PHILANI BRIAN MPOFU

EDUCATION

- Ph.D. in Biostatistics, Indiana University, February 2020
- B.A. in Mathematics and Economics, Vassar College, May 2012

RELEVANT WORKING EXPERIENCE

- Quantitative Scientist, Flatiron Health, New York, New York (October 2019 - Present)
- Research Assistant, Indiana University, Department of Biostatistics, Indianapolis, Indiana (September 2013 - October 2019)
- ORISE Fellow, US Food and Drug Administration, Silver Spring, Maryland (June 2018 - August 2018)

HONORS AND AWARDS

- Best poster presentation award. FDA Biostatistics Research Day.
- Vassar Ann Cornelisen Fellowship 2011 for Summer Language study in Argentina
- Omnicron Delta Epsilon (March 2012)
- Vassar College James Ryland and Georgia A. Kendrick Fellowship (2012) for M.Sc. in Bayesian Statistics at U. of Helsinki (declined)

SKILLS

- Computer: R, SAS, SQL, WinBUGS, Stata, SPSS, Minitab, MS office, LateX, Linux, Python (Beginner)
- Language: English, Ndebele, Shona, Beginner Spanish.

SELECT PUBLICATIONS

- Allam, E., P. Mpofu, A. Ghoneima, M. Tuceryan and K. Kula (2018). "The Relationship Between Hard Tissue and Soft Tissue Dimensions of the Nose in Children: A 3D Cone Beam Computed Tomography Study." Journal of Forensic Sciences 63(6): 1652-1660.
- Banks, D. E., A. T. Rowe, P. Mpofu and T. C. Zapolski (2017). "Trends in typologies of concurrent alcohol, marijuana, and cigarette use among US adolescents: An ecological examination by sex and race/ethnicity." Drug alcohol dependence 179: 71-77.
- McAllister, J. W., R. M. Keehn, R. Rodgers, P. B. Mpofu, P. O. Monahan and T. M. Lock (2018). "Effects of a Care Coordination Intervention with Children with Neurodevelopmental Disabilities and Their Families." Journal of Developmental and Behavioral Pediatrics 39(6): 471-480.
- McHenry, M. S., C. I. McAteer, E. Oyungu, B. C. McDonald, C. B. Bosma, P. B. Mpofu, A. R. Deathe and R. C. Vreeman (2018). "Neurodevelopment in Young Children Born to HIV-Infected Mothers: A Meta-analysis." Pediatrics 141(2): e20172888.
- Miller, M. D., D. Y. Sze, S. A. Padia, R. J. Lewandowski, R. Salem, P. Mpofu, P. M. Haste and M. S. Johnson (2018). "Response and Overall Survival for Yttrium-90 Radioembolization of Hepatic Sarcoma: A Multicenter Retrospective Study." Journal of Vascular and Interventional Radiology 29(6): 867-873.
- Saito, S., P. Mpofu, E. J. Carter, L. Diero, K. K. Wools-Kaloustian, C. T. Yiannoutsos, M. S. Beverly, S. Tsiouris, G. R. Somi, J. Ssali, D. Nash and B. Elul (2016). "Imple-

mentation and Operational Research: Declining Tuberculosis Incidence Among People Receiving HIV Care and Treatment Services in East Africa, 2007-2012." Jaids-Journal of Acquired Immune Deficiency Syndromes 71(4): e96-e106.

- Mpofu, P., Bakoyannis, G., Yiannoutsos, C. (Invited revision). A Pseudo-Likelihood Method For Estimating Misclassification Probabilities When Outcome Data Are Partially Observed.
- Mpofu, P., Bakoyannis, G., Yiannoutsos, C., Tu, W. . (Awaiting submission). Estimating Cause-Specific Hazards While Adjusting For Misclassification Information Obtained Via External Validation Sampling.
- Bakoyannis, G., Mpofu, P., Dixon, B.. (Awaiting submission). Pseudo-expected estimating equations for missing data problems.
- Mpofu, P., Karuri, S. (Submitted). An Alternative to the Cox Model Estimator In Time-to-event Clinical Trials With "Treatment of Physician's Choice".
- Humphrey, J. M., P. Mpofu, A. C. Pettit, B. Musick, E. J. Carter, E. Messou, O. Marcy, B. Crabtree-Ramirez, M. Yotebieng and K. Anastos (2019). "Mortality among adults living with HIV treated for tuberculosis based on positive, negative, or no bacteriologic test results for tuberculosis: the IeDEA consortium." BioRxiv: 571000.

PRESENTATIONS

- “Clinical Trials with Physician's Choice Control: Estimation of Treatment Effect using an alternative to Cox model estimator”. Poster presentation, FDA Biostatistics Research Day, August 18, 2018.
- “A Pseudo-Likelihood Method For Estimating Misclassification Probabilities When Outcome Data Are Partially Observed ”. Invited-paper presentation, ENAR Conference, March 26, 2018.