HHS Public Access

Author manuscript

Biom J. Author manuscript; available in PMC 2021 November 01.

Published in final edited form as:

Biom J. 2020 November ; 62(7): 1747–1768. doi:10.1002/bimj.201900198.

A pseudo-likelihood method for estimating misclassification probabilities in competing-risks settings when true event data are partially observed

Philani B. Mpofu[*,1], **Giorgos Bakoyannis**[1], **Constantin T. Yiannoutsos**[1], **Ann W. Mwangi**[2], **Margaret Mburu**[3]

[1]Department of Biostatistics, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, Indiana, 46202, USA

[2]Department of Behavioral Science, School of Medicine, Moi University, Kenya

[3]Family AIDS Care and Education Services (FACES), Research Care and Training Program (RCTP)-Kenya Medical Research Institute (KEMRI), Kisumu Kenya

Abstract

Outcome misclassification occurs frequently in binary-outcome studies and can result in biased estimation of quantities such as the incidence, prevalence, cause-specific hazards, cumulative incidence functions etc. A number of remedies have been proposed to address the potential misclassification of the outcomes in such data. The majority of these remedies lie 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, we 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. We present the estimator through the lens of studies with competing-risks outcomes, though the estimator extends beyond this setting. We describe the consistency and asymptotic distributional properties of the resulting estimator, and derive a closed-form estimator of its variance. The finite-sample performance of this estimator is evaluated via simulations. Using data from a real-world study with competing risks outcomes, we illustrate how the proposed method can be used to estimate misclassification probabilities. We also show how the estimated

[*]Corresponding author: phmpofu@iupui.edu.

Conflict of Interest: *The authors have declared no conflict of interest.*

misclassification probabilities can be used in an external study to adjust for possible misclassification bias when modeling cumulative incidence functions.

Keywords

Competing risks; Internal validation; Misclassification; Missing data; Pseudo-likelihood

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 and Lin, 2010; Lyles et al., 2011; Tang et al., 2015). 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 internally-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 et al., 2001).

In this paper, we focus on estimating misclassification probabilities with the ultimate goal to utilize these probabilities for adjusting competing-risks estimators. The motivation for this work is a large study of people living with HIV/AIDS (PLWH) in sub-Saharan Africa, that receive care at various health facilities participating in the East-African International Epidemiology Databases to Evaluate AIDS (IeDEA-EA) consortium. Specifically, one of the study's main objectives is to estimate mortality and the incidence of disengagement from care among PLWH. Death and disengagement from care are important outcomes in the monitoring and evaluating the effectiveness of care programs (Brinkhof et al., 2010; Egger et al., 2011; Bakoyannis and Yiannoutsos, 2015). In these studies, death and disengagement from care are treated as competing risks because the interest lies in the time to the first occurring event (Putter et al., 2007; Bakoyannis and Touloumi, 2012). These studies, as shown by IeDEA researchers, are susceptible to misclassification bias due to death underreporting (Egger et al., 2011; Bakoyannis and Yiannoutsos, 2015). Unreported deaths are typically classified as disengagements from care, which leads to an underestimation of mortality and an overestimation of rates of disengagement from care. This estimation bias is often reduced by adjusting estimators using death-misclassification information that is

generated through internal-validation sampling or double sampling (Geng et al., 2008; Yiannoutsos et al., 2008). For IeDEA-EA, the process of internal validation involves intensive tracing, in the community, of a subset of patients considered disengaged from care, and active ascertainment of their vital status (Geng et al., 2008; Yiannoutsos et al., 2008; An et al., 2009). True vital status data are missing-by-design for patients who were not selected for tracing. Moreover, true vital-status data are missing for some patients who were selected for internal-validation but could not be successfully traced.

When an internal-validation sample is available, most authors use only the internal-validation sample to estimate the extent of event outcome misclassification. By performing such a complete-case analysis, they implicitly assume that missing true event data on the non-validation sample are missing completely at random (MCAR) (Pepe, 1992; Magder and Hughes, 1997; Chen, 2000). That is, they implicitly assume that 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. Given that a complete-case analysis is not an ideal approach, several 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, if they exist, auxiliary variables that make the MAR assumption plausible (Lu and Tsiatis, 2001). 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, we propose a pseudo-likelihood approach for estimating event misclassification (uni-directional or bi-directional) probabilities when some of the binary-outcome data are missing by design and due to non-response. Motivated by the real-world data problem in the IeDEA-EA study, we focus our exploration to setting where the binary-outcome data arise in the context of competing-risks problem. That being said, our proposed method can be generalized to different settings where binary data may arise. Our method relaxes the MCAR assumption, which is untenable in our study context because not everyone who is sampled for internal validation is available to provide data. Instead, we assume that data are missing at random (MAR), allowing missingness to be related to observed subject characteristics (Rubin, 1976). Furthermore, unlike Rubin's multiple imputation, we 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; Bakoyannis et al., 2019). Auxiliary covariates are an important consideration in IeDEA-EA because when we build misclassification models using data that is generated from internal-validation sampling, we try to avoid the inclusion of auxiliary covariates that are related to the internal-validation

study. Due to resource constraints, internal-validation of vital status cannot be performed at all the study sites, which raises the need for the transfer misclassification information from study sites with internal validation to study sites without internal validation. At IeDEA-EA, an example of a study-related auxiliary variable is the number of workers assigned to perform the internal-validation (double-sampling) work. The number of workers assigned to outreach patients that are considered to be disengaged from care influences the size of the study sample that is successfully internally-validated, and, in turn, the level of missingness in our true vital-status data. This variable is not available in settings without double-sampling designs and, therefore, cannot be included in the misclassification probability models. The proposed pseudo-likelihood approach is appealing because it allows us to use information from auxiliary covariates without including them in misclassification models. The proposed method is also 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 such as those in IeDEA-EA. Standard-error estimation can be performed using the R function provided in the supplementary material of this manuscript. Alternatively, given the computational efficiency of the proposed estimator, one can use bootstrap for standard error estimation

This paper proceeds as follows: In Section 2 we state the data assumptions and notation, and also frame our motivation for estimating misclassification probabilities within a competing risks setting. In Section 3, we present the likelihood of the misclassification parameter assuming that all the true events are available, and describe the derivation of pseudo-likelihood function to deal with missing true events. In Section 4, we derive the large-sample properties of the resulting pseudo-likelihood estimator. In Section 5, we evaluate the finite-sample properties of estimator using a simulation study. In Section 6, we present a data application to illustrate the estimation of misclassification probabilities. In Section 7, we illustrate the use of misclassification probabilities estimated in Section 6 to make adjustments for potential misclassification in external studies with no outcome validation. We conclude with a brief discussion of our findings in Section 8.

2 Framing the motivation

2.1 Notation and Assumptions

This paper focuses on binary data that occur within a competing-risks setting, as a result *event types* shall also be referred to as *causes of failure* or *causes*. In keeping with our real-world example, lets consider a study where each subject is followed until he/she fails from either cause 1 or cause 2, or is censored. Lets also assume that the method of ascertaining the cause of failure is subject to error, so that the observed and true causes of failure, among the non-censored, are not always the same: Censored events are always correctly classified. 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. As such, let $C \in \{1, 2\}$ represent the observed cause of failure, and $C \in \{1, 2\}$ represent the true cause of failure.

Among those observed to fail from either cause 1 or cause 2, that is, $C^* = 1$ or $C^* = 2$, event outcomes on a sub-sample are re-ascertained using the gold-standard approach. We let R_i be

the indicator that the true cause of failure is known, with $R_i = 1$ indicating that the subject i was successfully double-sampled or censored. The true cause of failure, C_i, is only observed if subject i is successfully double sampled, or is censored ($C_i = 0$). For each subject, $i = 1, 2, \ldots, n$, we observe $\{C_i^*, \boldsymbol{X}_i = (T_i, \boldsymbol{X}_i^*), R_i, (C_i \text{ if } R_i = 1)\}$, where,

1. R_i: is the indicator function that the cause of failure for subject i has been re-ascertained through double sampling or subject i's cause of failure is censored;
2. $C_i \in \{0, 1, 2\}$: is the true cause of failure, and is available only if $R_i = 1$;
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 the time contribution to study;
5. V_i: is the censoring time;
6. T_i: is the failure time (that is, time to cause 1 or cause 2);
7. $U_i = \min(T_i, V_i)$: is the follow-up time for subject i;
8. $\boldsymbol{X}_i$: are the observed covariates for subject i, including follow-up time;

We assume that the censoring time is independent of failure time and the cause of failure conditional on the subject characteristics, that is, $(T, C) \perp V|\boldsymbol{X}^*$. The censoring assumption is only stated because it is necessary for the ultimate analysis of interest which is a competing risks analysis. We also assume that subject characteristics $\boldsymbol{X}^*$ and the time to event, U, are measured without error.

2.2 The impact of event misclassification on competing risks estimation

As an illustration, we will present the impact of misclassification on the modeling of cumulative incidence functions. For the assumed two-cause system, the cumulative incidence function for cause-j, $j \in \{1, 2\}$, at time t is defined as follows:

$$F_j(t) = P[T \leq t, C = j] = \int_0^t \lambda_j(v) S(v-) dv = \int_0^t S(v-) d\,\Lambda_j(v)$$

where $S(.)$ is the overall survival,

$$\lambda_j(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h, C = j | T \geq t)}{h}$$

is the cause-specific hazard for cause-j, and $\Lambda_j(t) = \int_0^t \lambda_j(v) dv$ is the cumulative cause-specific hazard.

In the presence event misclassification, that is, when the true cause of failure, C, and the observed cause of failure, C^*, are not necessarily identical, the cumulative incidence function with respect to the observed cause-j is given by:

$$F_j^*(t) = P[T \leq t, C^* = j] = \int_0^t \lambda_j^*(v) S(v-) dv$$

where $\lambda_j^*(t)$ is the cause-specific hazard with respect to the observed cause-*j*.

$F_j^*(t)$ is not necessarily the same as the desired target, $F_j(t)$. Mathematically, this difference can be explained by the fact that $\lambda_j^*(t)$ is not necessarily the same as $\lambda_j(t)$. In fact, it can be shown that the cause-specific hazard with respect to the observed cause-*j* is a linear combination of the true cause-specific hazards weighted by misclassification probabilities, that is:

$$\lambda_j^*(t) = \sum_{k=1}^{2} \lambda_k(t) P(C^* = j | T = t, C = k) \quad (1)$$

where, for $j, k \in \{1, 2\}$, $P(C^* = j | T = t, C = k)$ is the probability of observing cause-*j* given that true cause of failure is cause-*k*, conditional on the failure time, $T = t$. The proof for Equation 1 is presented in supporting information.

When one has correctly measured competing risks data, the cumulative incidence function for cause-*j* can be estimated consistently using the Aalen-Johansen estimator,

$$\hat{F}_j(t) = P[T \leq t, C = j] = \int_0^t \hat{S}(v-) d\hat{\Lambda}_j(v)$$

where $\hat{S}(.)$ is the Kaplan-Meier estimate for the survival function, and $\hat{\Lambda}_j(.)$ is the Nelson-Aalen estimator for the cumulative cause-specific hazard for cause-*j*. However, when the causes of failure are misclassified, the Aalen-Johansen estimator, for the reason illustrated by Equation 1, will result in biased estimation. As a result, Bakoyannis and Yiannoutsos (2015) proposed a modified Aalen-Johansen estimator which accounts for non-differential misclassification, and Edwards et al. (2019) extended the estimator to adjust for differential misclassification. In order for one to perform unbiased estimation under either method by Bakoyannis and Yiannoutsos (2015) or Edwards et al. (2019), one needs estimates of misclassification probabilities to be used in adjustment. As such, in this paper, we present a deliberative, and statistically-principled approach for using internal-validation data to estimate misclassification probabilities that are earmarked for competing risks applications.

3 Estimating misclassification probabilities

Equation 1 gave us a general structure of the conditional misclassification probabilities required to adjust for misclassification under a two-cause competing risks systems. The general structure, under bidirectional misclassification, is as follows:

1. $P[C^* = 1 \mid C = 2, \boldsymbol{X}, \beta_2] = \pi_{12}^*(\boldsymbol{X}; \beta_2;)$

2. $P[C^* = 2 \mid C = 1, \boldsymbol{X}, \beta_1] = \pi^*_{21}(\boldsymbol{X}; \beta_1)$

where, $\boldsymbol{X} = (T = t, \boldsymbol{X}^*)$ represents the matrix of subject characteristics, and $\beta = (\beta_1, \beta_2)$ represents the association between misclassification probabilities and subject characteristics. For simplicity, we 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. The reader should also notice that we are not treating the true cause, C, as a covariate within the misclassification models. Equivalently, we could have specified a single model, that contains the main effects $\boldsymbol{X}$ and C and interactions between $\boldsymbol{X}$ and C. However, we use separate models since it is easier this way to adapt the proposed methods to settings with uni-directional misclassification (e.g. with $\pi^*_{12}(\mathbf{X}; \beta_2) = 0$), as it is the case with our motivating IeDEA-EA study.

With real-world applications in mind, we 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:

$$\log\left[\frac{\pi^*_{12}(\boldsymbol{X}; \beta_2)}{1 - \pi^*_{12}(\boldsymbol{X}; \beta_2)}\right] = \boldsymbol{X}^T \beta_2 \tag{2}$$

$$\log\left[\frac{\pi^*_{21}(\boldsymbol{X}; \beta_1)}{1 - \pi^*_{21}(\boldsymbol{X}; \beta_1)}\right] = \boldsymbol{X}^T \beta_1 \tag{3}$$

where $\beta_1, \beta_2 \in \mathbb{R}^q$, and $\boldsymbol{X}_{n \times q}$.

3.1 Ideal complete-data likelihood

Under the assumptions presented above, the log-likelihood of $\beta = (\beta_1, \beta_2)^T \in \mathbb{R}^{2q}$ based on the full data is

$$\begin{aligned} l(\beta; \boldsymbol{\Delta}) = &\sum_{i=1}^{n} \delta_{1i}\{\delta^*_{2i} \boldsymbol{X}_i^T \beta_1 - \log[1 + \exp(\boldsymbol{X}_i^T \beta_1)]\} \\ &+ \sum_{i=1}^{n} \delta_{2i}\{\delta^*_{1i} \boldsymbol{X}_i^T \beta_2 - \log[1 + \exp(\boldsymbol{X}_i^T \beta_2)]\} \end{aligned} \tag{4}$$

where $\delta_{1i} = I[C_i = 1]$ and $\delta_{2i} = I[C_i = 2]$ are true-event indicators, $\delta^*_{1i} = I[C^*_i = 1]$ and $\delta^*_{2i} = I[C^*_i = 2]$ are observed-event indicators, and $\boldsymbol{\Delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \boldsymbol{\delta}^*_1, \boldsymbol{\delta}^*_2)$ is a composite of true- and observed-event vectors. For censored subject, i, $(\delta_{1i} = 0, \delta_{2i} = 0)$ if and only if $(\delta^*_{1i} = 0, \delta^*_{2i} = 0)$. The derivation of the full-likelihood from which log-likelihood (4) is obtained can be found in supporting information.

In the above full log-likelihood (4), we should notice that δ_{1i} and δ_{2i} are only observable among those who were successfully double-sampled or censored, that is, some subset of $\{i = 1, 2, \ldots, n\}$. As a result, maximum likelihood estimation is not straightforward. We can proceed with maximum likelihood estimation by setting up an EM algorithm (Dempster et al., 1977; Magder and Hughes, 1997). 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 our motivating HIV study. To overcome these shortcomings, we proceed by first formulating the objective function as a pseudo/estimated likelihood.

3.2 Setting up the pseudo-likelihood

We begin by recognizing that the true-event indicators, δ_{1i} and δ_{2i}, are linear in the log-likelihood as shown in Equation 4. As a result of this linearity, we can still perform consistent estimation by replacing missing true-event indicator values using their conditional expectations given the observed data. That is, among those missing true outcome values, δ_{ji} is replaced by $\mathrm{E}[\delta_{ji} \mid \delta^*_{ki}, \boldsymbol{Z}_i] = p_{jk}(\boldsymbol{Z}_i; \gamma_k)$ for $j, k \in \{1, 2\}$, where $\boldsymbol{Z}$ is a matrix of subject characteristics. 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 and 3.

In the context of a real data analysis, estimation proceeds by replacing δ_{1i} and δ_{2i}, in the full log-likelihood 4 by $\tilde{\delta}_{1i}$ and $\tilde{\delta}_{2i}$ respectively, where:

$$\tilde{\delta}_{1i}(R_i, \delta_{1i}, \boldsymbol{Z}_i; \hat{\gamma}_{n_v}) = R_i \times \delta_{1i} + (1 - R_i) \times \left[p_{11}(\boldsymbol{Z}_i; \hat{\gamma}_{1, n_v})^{\delta^*_{1i}} p_{12}(\boldsymbol{Z}_i; \hat{\gamma}_{2, n_v})^{1 - \delta^*_{1i}} \right] \quad (5)$$

and

$$\tilde{\delta}_{2i}(R_i, \delta_{2i}, \boldsymbol{Z}_i; \hat{\gamma}_{n_v}) = R_i \times \delta_{2i} + (1 - R_i) \times \left[p_{21}(\boldsymbol{Z}_i; \hat{\gamma}_{1, n_v})^{\delta^*_{1i}} p_{22}(\boldsymbol{Z}_i; \hat{\gamma}_{2, n_v})^{1 - \delta^*_{1i}} \right] \quad (6)$$

where

- **i.** n_V is the size of an internal-validation sample that has been drawn from a main-study sample of size n
- **ii.** $\hat{\gamma}_{1, n_v}, \hat{\gamma}_{2, n_v}$ are respective estimates of $\gamma_1, \gamma_2 \in \mathbb{R}^d$;
- **iii.** $\boldsymbol{Z}_i$ is a $1 \times d$ matrix containing the characteristics for subject i
- **iv.** $p_{jk}(\boldsymbol{Z}_i; \hat{\gamma}_{k, n_v}) = P(C_i = j \mid C^*_i = k, \boldsymbol{Z}_i, \hat{\gamma}_{k, n_v})$ for $j, k \in \{1, 2\}$ is the estimated conditional probability of the true cause $C = j$ given the observed cause $C^* = k$, for $k \in \{1, 2\}$,
- **v.** $\sum_{j=1}^{2} p_{jk}(\boldsymbol{Z}_i; \hat{\gamma}_{k, n_v}) = 1$

Henceforth, for all j and k, we shall refer to $p_{jk}(\boldsymbol{Z}_i; \gamma_k)$ 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. We assume that missing true event data are missing at random(MAR). That is, among the *non-censored*, the probability that the true cause is missing is independent of the true cause of failure conditional on the observed cause and the subject characteristics. That is, for $j, k \in \{0, 1, 2\}$,

$$\mathrm{P}(R_i = 0 \big| \boldsymbol{Z}_i, C_i = k, C_i^* = j) = \mathrm{P}(R_i = 0 \big| \boldsymbol{Z}_i, C_i^* = j) \tag{7}$$

The covariate matrix $\boldsymbol{Z}$ may also include *auxiliary* covariates that make the MAR assumption more plausible. From the MAR assumption defined in Equation 7, without losing generality, it follows that:

$$\begin{aligned} &\mathrm{P}(C_i = 1 \big| \boldsymbol{Z}_i, R_i = 0, C_i^* = j) = \mathrm{P}(C_i = 1 \big| \boldsymbol{Z}_i, R_i = 1, C_i^* = j) \\ &= \mathrm{P}(C_i = 1 \big| \boldsymbol{Z}_i, C_i^* = j) \end{aligned} \tag{8}$$

for $j \in \{0, 1, 2\}$.

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 et al., 2010). From a data analysis perspective, this means we can use predictive values estimated using data from those whose event data were validated to inform the predictive value estimates among those whose event data were not validated (that is, assuming that the validated and unvalidated subjects are drawn from the same population).

We model the predictive values, $p_{jk}(\boldsymbol{Z};\ \gamma_k)$ for $j, k \in \{1, 2\}$, parametrically using logistic regression. Therefore, we set up the predictive values as follows:

$$p_{12}[\boldsymbol{Z}; \gamma_2] = \frac{\exp(\boldsymbol{Z}^T \gamma_2)}{1 + \exp(\boldsymbol{Z}^T \gamma_2)} \tag{9}$$

$$p_{21}[\boldsymbol{Z}; \gamma_1] = \frac{\exp(\boldsymbol{Z}^T \gamma_1)}{1 + \exp(\boldsymbol{Z}^T \gamma_1)} \tag{10}$$

When we replace δ_{1i} and δ_{2i} with $\widetilde{\delta}_{1i}$ and $\widetilde{\delta}_{2i}$ respectively, the resulting pseudo-log-likelihood(estimated log-likelihood) is $m(\boldsymbol{\beta}; \hat{\gamma}_{n_v})$, where the overall parameter is $(\beta, \gamma) \in \mathbb{R}^{2(q+d)}$, with $\boldsymbol{\beta}$ being the parameter of interest, and $\boldsymbol{\gamma}$ being the nuisance parameter. The parameter $\boldsymbol{\gamma}$ is estimated by fitting logistic regression models 9 and 10 using the internal-validation data as stated above. Assuming the logistic regression models are correctly specified, $\hat{\gamma}_{n_v}$ will converge in probability to $\boldsymbol{\gamma}$. When we plug in $\hat{\gamma}_{n_v}$ into the log-

likelihood, our problem reduces to that of optimizing $m(\beta;\hat{\gamma}_{n_v})$, a pseudo-log-likelihood, and the resulting estimates are called pseudo-likelihood estimates.

3.3 Maximum pseudo-likelihood estimation

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

$$\boldsymbol{\Psi}_n^{(1)}(\beta_1;\hat{\gamma}_{n_v}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \tilde{\delta}_{1i}(R_i,\delta_{1i},\boldsymbol{Z}_i;\hat{\gamma}_{n_v})\left[\delta_{2i}^* - \frac{\exp(\boldsymbol{X}_i^T\beta_1)}{1+\exp(\boldsymbol{X}_i^T\beta_1)}\right] = \mathbf{0} \tag{11}$$

$$\boldsymbol{\Psi}_n^{(2)}(\beta_2;\hat{\gamma}_{n_v}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \tilde{\delta}_{2i}(R_i,\delta_{2i},\boldsymbol{Z}_i;\hat{\gamma}_{n_v})\left[\delta_{1i}^* - \frac{\exp(\boldsymbol{X}_i^T\beta_2)}{1+\exp(\boldsymbol{X}_i^T\beta_2)}\right] = \mathbf{0} \tag{12}$$

where $\hat{\gamma}_{n_v} = (\hat{\gamma}_{1,n_v},\hat{\gamma}_{2,n_v})$; n is the size of main-study sample, and n_v is the size of the internal-validation sample that has been selected from the main-study sample.

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

4 Asymptotic properties

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

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

Under the regularity conditions we proved the asymptotic properties of consistency and normality as stated in *Theorem 1* and *Theorem 2* below.

Theorem 1: The maximum pseudo-likelihood estimator $\hat{\beta}_n$ is a consistent estimator of β_0, that is,

$$\|\hat{\beta}_n - \beta_0\| \xrightarrow{p} 0$$

The detailed proof for consistency can be found in the supporting information.

Theorem 2: The maximum pseudo-likelihood estimator $\hat{\beta}_n$ is asymptotically normal, that is:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{\Omega})$$

where the asymptotic variance,

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

where $\boldsymbol{I}(\beta_0, \gamma_0; \mathbf{\Delta}_i) = -\mathrm{E}\left[\frac{\partial}{\partial\beta}\mathbf{\Psi}_n(\beta, \gamma_0; \mathbf{\Delta}_i, \boldsymbol{X}_i)\mid_{\beta=\beta_0}\right]$ is a $q \times q$ information matrix with respect to misclassification parameter, β, when the predictive-value parameter γ is known; and

$$\begin{aligned}\mathbf{W}(\beta_0, \gamma_0; \mathbf{\Delta}) &= \mathrm{E}\left[\mathbf{J}(\beta_0, \gamma_0; \mathbf{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, R)\boldsymbol{I}^{-1}(\gamma_0; \mathbf{\Delta}, \boldsymbol{Z})\dot{l}(\gamma_0|\mathbf{\Delta}, \boldsymbol{Z})\dot{m}(\beta_0, \gamma_0|\mathbf{\Delta}, \boldsymbol{X})^T\right] \\ &+ \mathrm{E}\left[\dot{m}(\beta_0, \gamma_0|\mathbf{\Delta}, \boldsymbol{X})\dot{l}(\gamma_0|\mathbf{\Delta}, \boldsymbol{Z})^T\boldsymbol{I}^{-1}(\gamma_0; \mathbf{\Delta}, \boldsymbol{Z})\mathbf{J}(\beta_0, \gamma_0; \mathbf{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, R)^T\right] \\ &+ \mathbf{J}(\beta_0, \gamma_0; \mathbf{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, R)\boldsymbol{I}^{-1}(\gamma_0; \mathbf{\Delta}, \boldsymbol{Z})\mathbf{J}(\beta_0, \gamma_0; \mathbf{\Delta}, \boldsymbol{X}, \boldsymbol{Z}, R)^T\end{aligned}$$

where

- **i.** $\dot{m}(\beta_0, \gamma_0; \mathbf{\Delta}_i) = \frac{\partial}{\partial\beta}m(\beta, \gamma_0; \mathbf{\Delta}_i, \boldsymbol{X}_i)\mid_{\beta=\beta_0}$ is a $q \times 1$ gradient (score function with respect to the pseudo-likelihood) with respect to the misclassification parameter, when the predictive-value parameter γ is known.
- **ii.** $\mathbf{J}(\beta_0, \gamma_0; \mathbf{\Delta}_i) = \left[\frac{\partial}{\partial\gamma}\mathbf{\Psi}_n(\beta_0, \gamma; \mathbf{\Delta}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i, R_i)\mid_{\gamma=\gamma_0}\right]$ is a $q \times d$ Jacobian matrix;
- **iii.** $\dot{l}(\gamma_0; \mathbf{\Delta}_i) = \frac{\partial}{\partial\gamma}l(\gamma; \mathbf{\Delta}_i, \boldsymbol{Z}_i)\mid_{\gamma=\gamma_0}$ is a $d \times 1$ gradient (score function) with respect to the predictive-value model parameter, γ.
- **iv.** $\boldsymbol{I}(\gamma_0; \mathbf{\Delta}_i) = -\mathrm{E}\left[\ddot{l}(\gamma_0; \mathbf{\Delta}_i, \boldsymbol{Z}_i)\right]$ is a $d \times d$ information matrix with respect to the predictive-value model parameter, γ.
- **v.** $s = \lim_{n\to\infty}\frac{n}{n_v}$, with n being the size of the main-study sample, and n_V being the size of the validation sample, that has been drawn out of the main-study sample.
- **vi.** Note: $\boldsymbol{X}$ and $\boldsymbol{Z}$ need not be the same: In other words, the covariates that go into the misclassification model *need not* be same as the covariates that go into the predictive-value model.

The asymptotic variance, $\mathbf{\Omega}$, can be estimated by replacing the parameter β_0 and γ_0 with their consistent estimators and the expectations with sample averages so that

$$\widehat{\mathbf{\Omega}}_n = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\psi}(\mathbf{\Delta}_i|\hat{\beta}_n, \hat{\gamma}_{n_v})\widetilde{\psi}(\mathbf{\Delta}_i|\hat{\beta}_n, \hat{\gamma}_{n_v})^T$$

where $\widetilde{\psi}(\boldsymbol{\Delta}_i \mid \beta_0, \gamma_0) = -\boldsymbol{I}^{-1}(\beta_0, \gamma_0; \boldsymbol{\Delta}_i)\left[\dot{m}(\beta_0, \gamma_0; \boldsymbol{\Delta}_i) + \sqrt{s}\mathbf{J}(\beta_0, \gamma_0; \boldsymbol{\Delta}_i)\boldsymbol{I}^{-1}(\gamma_0; \boldsymbol{\Delta}_i)\dot{l}(\gamma_0; \boldsymbol{\Delta}_i)\right]$ is the weighted score-contribution for subject-i. A detailed proof for asymptotic normality can be found in supporting material.

Remark 1: $s.\boldsymbol{I}^{-1}(\beta_0, \gamma_0)\mathbf{W}(\beta_0, \gamma_0, \boldsymbol{X}, \boldsymbol{Z})\boldsymbol{I}^{-1}(\beta_0, \gamma_0)$ is the additional variability associated with estimating the nuisance parameter, γ.

Remark 2: As $n_v \to n$, that is, as $s \to 1$, $\boldsymbol{\Omega} \approx \boldsymbol{I}^{-1}(\beta_0, \gamma_0; \boldsymbol{\Delta})$.

4.1 Implementing the pseudo-likelihood estimation in R

It is fairly simple to set up the estimating equation represented by Equation 11 (or 12) 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

$\widetilde{\delta}_{1i} = R_i \times \delta_{1i} + (1 - R_i) \times [p_{11}(\boldsymbol{Z}_i; \hat{\gamma}_{1, n_v})^{\delta_{1i}^*} p_{12}(\boldsymbol{Z}_i; \hat{\gamma}_{2, n_v})^{1 - \delta_{1i}^*}]$, for subject indices $i = 1, ..n$.

Under such a setup, a subject whose cause of failure was validated is weighted based on his/her validated outcome(0 versus 1), otherwise he/she 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, $p_{jk}(\boldsymbol{Z}; \gamma_k)$. We provide, in this manuscript's supplement, an example R function that implements the proposed standard error estimator as described in Section 4. We should also note that, as an alternative to the closed-form variance estimator, one could appeal to bootstrapping given the computational efficiency and the $\sqrt{n}$-consistency of the proposed pseudo-likelihood estimator.

5 Simulation Study

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, $T \sim \mathrm{WB}(\alpha, \lambda)$. Assume that censoring time, $V \sim \mathrm{Exp}(\eta)$. Under a competing risks scenario with two causes of failure, $C \in \{1, 2\}$, the survival outcome is represented as follows: $\{U = \min(T, V), C = k\}$, for $k \in \{0, 1, 2\}$; $C = 0$ for censored cases. Assuming proportional hazards, at $T = t$, the cause-specific hazard for cause-k is $h_k(t; \boldsymbol{Z}, \boldsymbol{\theta}_k) = \alpha_k \lambda_k t^{\alpha_k - 1} \exp(\boldsymbol{Z}^T \boldsymbol{\theta}_k)$, for $k = 1, 2$, where $\boldsymbol{Z} = (z_1, z_2)$ is the matrix of covariates, and $\boldsymbol{\theta}_k$ captures the multiplicative dependence between the cause-specific hazard for cause-k and the covariates $\boldsymbol{Z}$. In our simulation, we set $\alpha_1 = \alpha_2 = \alpha$.

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. We assume that those who are censored are never misclassified, that is, $C = 0$ if and only if $C^* = 0$. We will let $\mathrm{P}[C^* = 2 \mid C = 1, \boldsymbol{X}] = \pi_{21}^*$, and $\mathrm{P}[C^* = 1 \mid C = 2, \boldsymbol{X}] = \pi_{12}^*$ be the misclassification probabilities as defined in Section 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 $\boldsymbol{X}_{n\times 4} = [\mathbf{1}, \mathbf{t}, \boldsymbol{Z}_1, \boldsymbol{Z}_2]$. For each subject, we generate $M_i \sim Ber(I(C_i = 1)\,.\,\pi_{21i}^* + I(C_i = 2)\,.\,\pi_{12i}^*)$, the misclassification indicator for subject i, where $M_i = 1$ indicates that the outcome is misclassified. The observed cause of failure for subject i, C_i^*, is then defined as follows:

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

5.1 True outcomes missing at random (MAR)

In addition to exploring a situation where data are missing completely at random (MCAR), our simulations also explore 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, we explore 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 et al., 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

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

5.2 Conducting the simulation study

We set the simulation parameters as follows:

- **a.** *Misclassification parameters:* β_1 = (−0.4, −0.4, 0.5, −0.5), β_2 = (−0.4, −0.4, 0.5, −0.5);
- **b.** *Weibull proportional hazards parameters:* $\boldsymbol{\theta}_1$ *= (0.5, 1),* $\boldsymbol{\theta}_2$ = (−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. In the simulations, we 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 5.1.

iv. *predictive value model specification* (correct versus incorrect). When we set $\beta_1 = \beta_2$, for example $\beta_2 = (-0.3, 0.2, 0.5, 0.5)$, theory suggest that logistic regression may not be appropriate for modeling the predictive values. That is, when $\beta_1 \neq \beta_2$, the proposed logistic models 9 and 10 may not be suitable because the linearity assumption between the logit function and the parameters 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 the supporting material. 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, we performed 1000 replications and then obtained the following quantities: average estimate, $\hat{\beta}_{average} = \frac{1}{1000}\sum_{l=1}^{1000} \hat{\beta}_l$; absolute percent bias of average estimate, $100 \times \left| \frac{\hat{\beta}_{average} - \beta}{\beta} \right|$; 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. We repeated estimation using the EM algorithm, and compared the computational efficiency of the EM algorithm to that of the proposed pseudo-likelihood method.

5.3 Simulation Results

Results under MCAR—The datasets used in simulations with 20% double-sampling under MCAR are summarized in Figure 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 pseudolikelihood 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 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.

Computational efficiency—The comparison of results from the EM and pseudo-likelihood methods is presented in Table 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 pseudolikelihood estimation method. That being said, the proposed pseudo-likelihood estimation method was computationally faster than the EM algorithm. To compare computational efficiency, we 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 *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. At the different sample sizes, and under the computational restrictions of the computer used, the pseudolikelihood approach was found to converge, on average, 93.6 times faster than the EM algorithm.

6 Application 1: Estimating misclassification probabilities

6.1 Notation

In this application, we define C^* 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

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

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, $P[C^* = 2 | C = 1; \textit{covariates}]$.

6.2 Data

We 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 4.

6.3 Methods

The misclassification probabilities were modeled using the pseudo-likelihood method presented in this paper. First, we modeled the predictive value of death $P[C = 1|C^* = 2; \textit{covariates}]$ 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 $P[C = 2|C^* = 1; \textit{covariates}] = 0$.

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 (Lin et al., 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, we built a model for the misclassification probabilities, $P[C^* = 2|C = 1, \textit{covariates}]$. We also assessed model goodness-of-fit using the Supremum goodness-of-fit test at the 0.05 alpha level.

6.4 Results

The misclassification models resulting from performing a complete-case analysis and a pseudo-likelihood-based analysis are presented in Table 5. 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% 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% 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.

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

In this section we illustrate how the misclassification probabilities estimated from an external study can be used in a situation where no outcome validation has been performed. In the present analysis, we use the misclassification probabilities derived from a treatment program with an available internal-validation sample to inform the possible misclassification in a new treatment program that does not have outcome validation, and then use this information to adjust the observed estimator of the cumulative incidence of death in the program without a validation sample. In the case of our analysis, 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 6. 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 (Carroll et al., 2006; Spiegelman, 2010; Lyles et al., 2011). 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 used the same data as Edwards et al. (2019). The 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. In this application, we plugged the misclassification probability estimates from the pseudo-likelihood method as shown in Table 5, into the cumulative incidence estimator by Edwards et al. (2019) to estimate the cumulative incidence of death in the FACES cohort adjusting for possible differential death misclassification. The results of the adjustment are shown in Figure 3. In the FACES cohort, the naïve cumulative incidence function estimate of mortality at 12 months after ART initiation was about 1.9% (95% C.I.: 1.48%-2.36%), whereas the misclassification-adjusted cumulative incidence function estimate of mortality at 12 months was about 6.4% (95% C.I.: 4.89%-7.98%). That is, the misclassification-adjusted mortality was about 3.37 times the unadjusted mortality within the first year of follow-up.

8 Discussion

In this paper we present a pseudo-likelihood method of estimating binary misclassification probabilities in the presence of an internal validation sample. We 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, we 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, we 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-sample and finite-sample properties. 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, in cases where there are no auxiliary variables, complete-case analysis under the logistic regression model will yield consistent estimates (Little and Rubin). Even in this case, however, the proposed method is expected to provide more precise estimates. Lastly, we should note that, although we considered a special case with two events/causes of failure in this manuscript, the proposed method easily extends to situations with three or more causes of failure. In such situations, one needs to use multinomial models instead of a binary logistic models.

We concede that the proposed pseudo-likelihood approach is not a “panacea” or the only solution. The success of the pseudo-likelihood approach that we have presented also depends on the size of the internal-validation sample. Estimating γ_0 using a small internal-validation

sample can lead to imprecise estimates and to less than satisfactory asymptotic linearity approximations from distribution of $\hat{\gamma}_{n_v}$. Besides the pseudo-likelihood approach, 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 estimation described in this article can be easily implemented using the glm function in the R software. The variance estimator of the pseudo-likelihood estimator can be coded to R based on the closed-form formula presented in Section 4; alternatively, it can be computed via bootstrapping given the computational efficiency and the $\sqrt{n}$-consistency of the proposed pseudo-likelihood estimator. We have provided, in the supporting material associated with this manuscript, a sample R function that provides point estimates and standards errors according to the proposed methodology.

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. A remedy for the potential misspecification of this model is to consider flexible penalized parametric models such as those discussed by Zhang and Little (2009) (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 consider auxiliary covariates that make the MAR assumption plausible: Omitting important auxiliary covariates when building the predictive-value models can lead to biased estimation. For an ideal world wherein all important covariates are accounted for, our simulation studies suggested that the proposed pseudo-likelihood estimator exhibits some degree of robustness against the misspecification of the parametric predictive model. This peculiar finding may be explained by the theory of misspecified maximum likelihood estimators. The estimate $\hat{\gamma}_{n_v}$ under a misspecified predictive model is the minimizer of the Kullback-Leibler divergence between the assumed (misspecified)

model and the true model. Therefore, in such cases, our pseudo-likelihood estimator is the *closest,* with respect to the Kullback-Leibler divergence, to the true predictive model. This may explain the small bias observed in our simulation studies.

We hope we have convinced the reader 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 article, we 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 naïve 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 naïve 12-month cumulative incidence of about 1.9%. This change, in our opinion, delineates the importance of statistically principled ways of estimating misclassification probabilities.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the anonymous reviewers who reviewed our work: Their comments and insights greatly improved our manuscript.

This research was supported by the National Institute Of Allergy And Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Institute On Drug Abuse (NIDA), National Cancer Institute (NCI), and the National Institute of Mental Health (NIMH), in accordance with the regulatory requirements of the National Institutes of Health under Award Number U01 AI069911 East

Africa IeDEA Consortium. This research was also supported by the NIAID under Award Number R21 AI145662 “Estimating the cascade of HIV care under incomplete outcome ascertainment”. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research has also been supported by the National Institutes of Health grant R01 AI102710 “Statistical Designs and Methods for Double-Sampling for HIV/AIDS” and by the President’s Emergency Plan for AIDS Relief (PEPFAR) through USAID under the terms of Cooperative Agreement No. AID-623-A-12-0001 it is made possible through joint support of the United States Agency for International Development (USAID). The contents of this presentation are the sole responsibility of AMPATH and do not necessarily reflect the views of USAID or the United States Government.

9: Appendix

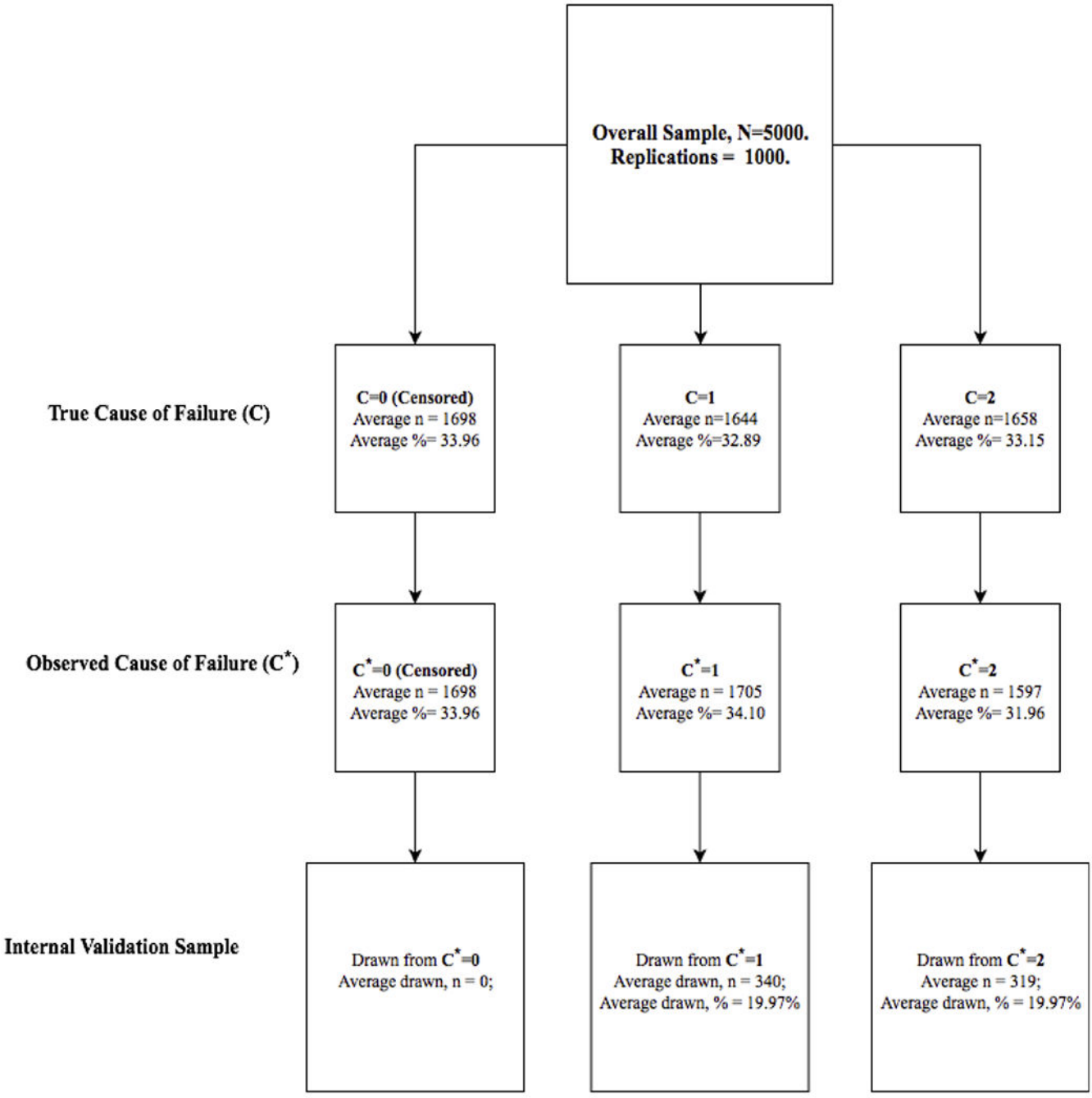

Figure 1.
Summary of simulation samples used when double sampling was set at 20%.

Table 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					Pseudolikelihood Estimator					
Model, (ds %)	**Cause**	**Parameter**	**True Value**	**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	1.578
		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	1.598
		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	1.623
		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	1.596
	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	1.921
		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	1.900
		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	1.755
		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	1.770
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	1.525
		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	1.682
		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	1.775
		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	1.851
	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	1.957
		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	1.931
		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	1.901
		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	1.894
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	1.310
		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	1.353
		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	1.286
		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	1.269
	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	1.364
		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	1.404
		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	1.358
		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
	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	1.329
		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	1.357
		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	1.319

Model, (ds %)	Cause	Parameter	True Value	Complete Case Estimator					Pseudolikelihood Estimator					RE
				Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP	
Incorrect (50%)	2	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	1.397
		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	1.404
		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	1.407
		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	1.358
		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	1.366

Table 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.

Model*,Planned DS%(Actual DS %)**	Cause	Parameter	True Value	Complete Case Estimator					Pseudo-likelihood Estimator				
				Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP
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	0.95
		β_{11} (t)	**−0.40**	−0.404	1.00	0.436	0.432	0.947	−0.399	0.25	0.338	0.328	0.93
		β_{21} (z_1)	**0.50**	0.526	5.20	0.516	0.514	0.947	0.528	5.60	0.394	0.386	0.95
		β_{31} (z_2)	**−0.50**	−0.514	2.80	0.176	0.182	0.953	−0.509	1.80	0.133	0.136	0.96
	2	β_{02}(Intercept)	**−0.40**	−0.489	22.25	0.375	0.360	0.932	−0.415	3.75	0.264	0.251	0.92
		β_{12} (t)	**−0.40**	−0.405	1.25	0.398	0.368	0.934	−0.407	1.75	0.265	0.253	0.94
		β_{22} (z_1)	**0.50**	0.517	3.40	0.497	0.500	0.943	0.521	4.20	0.363	0.365	0.94
		β_{32} (z_2)	**−0.50**	−0.518	3.60	0.185	0.176	0.938	−0.510	2.00	0.132	0.128	0.94
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	0.95
		β_{11} (t)	**−0.4**	−0.415	3.75	0.259	0.266	0.958	−0.408	2.00	0.218	0.219	0.95
		β_{21} (z_1)	**0.5**	0.490	2.00	0.322	0.320	0.947	0.495	1.00	0.263	0.261	0.94
		β_{31} (z_2)	**−0.5**	−0.512	2.40	0.114	0.112	0.943	−0.510	2.00	0.091	0.092	0.95
	2	β_{02}(Intercept)	**−0.4**	−0.460	15.00	0.225	0.224	0.940	−0.382	4.50	0.174	0.174	0.94
		β_{12} (t)	**−0.4**	−0.410	2.50	0.236	0.228	0.944	−0.411	2.75	0.178	0.176	0.94
		β_{22}(t)	**0.5**	0.487	2.60	0.311	0.312	0.952	0.478	4.40	0.256	0.250	0.93
		β_{32} (z_2)	**−0.5**	−0.512	2.40	0.113	0.109	0.938	−0.509	1.80	0.090	0.087	0.94

Model*,Planned DS%(Actual DS %)**	Cause	Parameter	True Value	Complete Case Estimator					Pseudo-likelihood Estimator				
				Estimate	% Bias	MCSD	ASE	CP	Estimate	% Bias	MCSD	ASE	CP
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	0.95
		β_{11} (t)	**−0.4**	−0.407	1.75	0.421	0.431	0.950	−0.402	0.50	0.317	0.320	0.94
		β_{21} (z_1)	**0.5**	0.509	1.80	0.533	0.514	0.943	0.497	0.60	0.386	0.375	0.94
		β_{31} (z_2)	**−0.5**	−0.516	3.20	0.174	0.181	0.951	−0.506	1.20	0.124	0.128	0.95
	2	β_{02}(Intercept)	**−0.3**	−0.366	22.00	0.364	0.356	0.939	−0.291	3.00	0.247	0.242	0.94
		β_{12} (t)	**0.2**	0.191	4.50	0.362	0.364	0.955	0.196	2.00	0.245	0.246	0.94
		β_{22} (z_1)	**0.5**	0.504	0.80	0.480	0.490	0.955	0.485	3.00	0.340	0.351	0.96
		β_{32} (z_2)	**0.5**	0.506	1.20	0.168	0.172	0.949	0.499	0.20	0.114	0.119	0.95
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	0.94
		β_{11} (t)	**−0.4**	−0.397	0.75	0.261	0.265	0.941	−0.396	1.00	0.210	0.216	0.95
		β_{21} (z_1)	**0.5**	0.501	0.20	0.324	0.319	0.949	0.495	1.00	0.267	0.256	0.92
		β_{31} (z_2)	**−0.5**	−0.508	1.60	0.111	0.112	0.951	−0.503	0.60	0.089	0.088	0.94
	2	β_{02}(Intercept)	**−0.3**	−0.372	24.00	0.227	0.222	0.937	−0.295	1.67	0.173	0.170	0.93
		β_{12} (t)	**0.2**	0.199	0.50	0.225	0.226	0.953	0.205	2.50	0.172	0.172	0.95
		β_{22} (z_1)	**0.5**	0.488	2.40	0.308	0.307	0.953	0.477	4.60	0.235	0.242	0.96
		β_{32} (z_2)	**0.5**	0.501	0.20	0.109	0.108	0.953	0.498	0.40	0.085	0.083	0.95

Table 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%.

			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

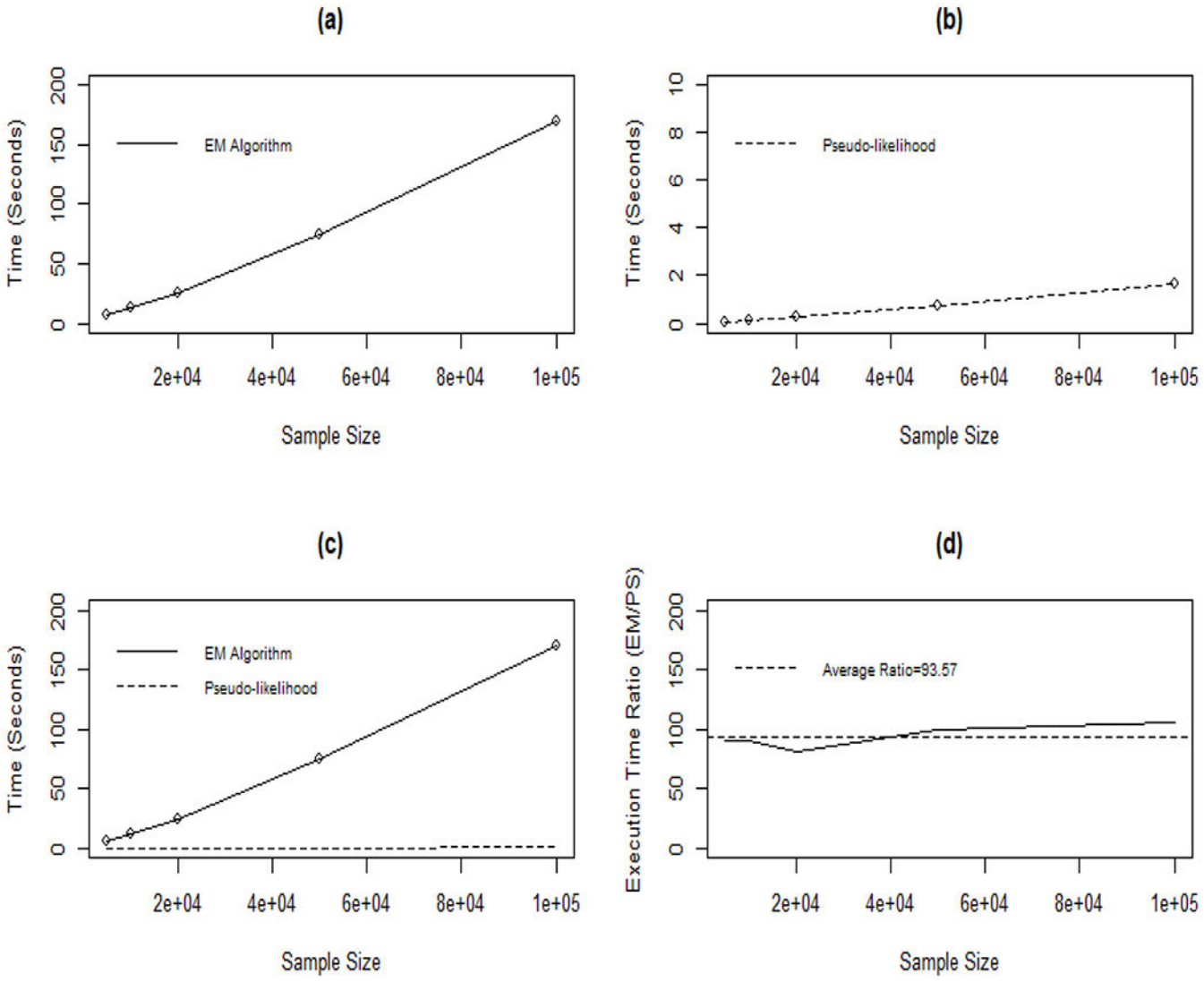

Figure 2.

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.

Table 4

Characteristics of patients involved in the misclassification 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.

		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 5

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.

	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

	Complete Case Analysis, N=3862				Pseudo-likelihood Method, N=28084			
	Estimate	**SE**	**Z**	**Pr(> |Z|)**	**Estimate**	**SE**	**Z**	**Pr(> |Z|)**
Study time (months)	0.025	0.0115	2.161	0.0307	0.058	0.0087	6.629	0.0000
I(3 ≤ Study time < 6)× (Study time - 3)	0.016	0.0601	0.269	0.7876	−0.031	0.0358	−0.868	0.3856
I(6 ≤ Study time < 12)× (Study time - 6)	−0.028	0.0379	−0.743	0.4574	−0.053	0.0232	−2.299	0.0215
I(Study time ≥ 12)×(Study time - 12)	−0.027	0.0155	−1.711	0.0871	−0.068	0.0107	−6.370	0.0000

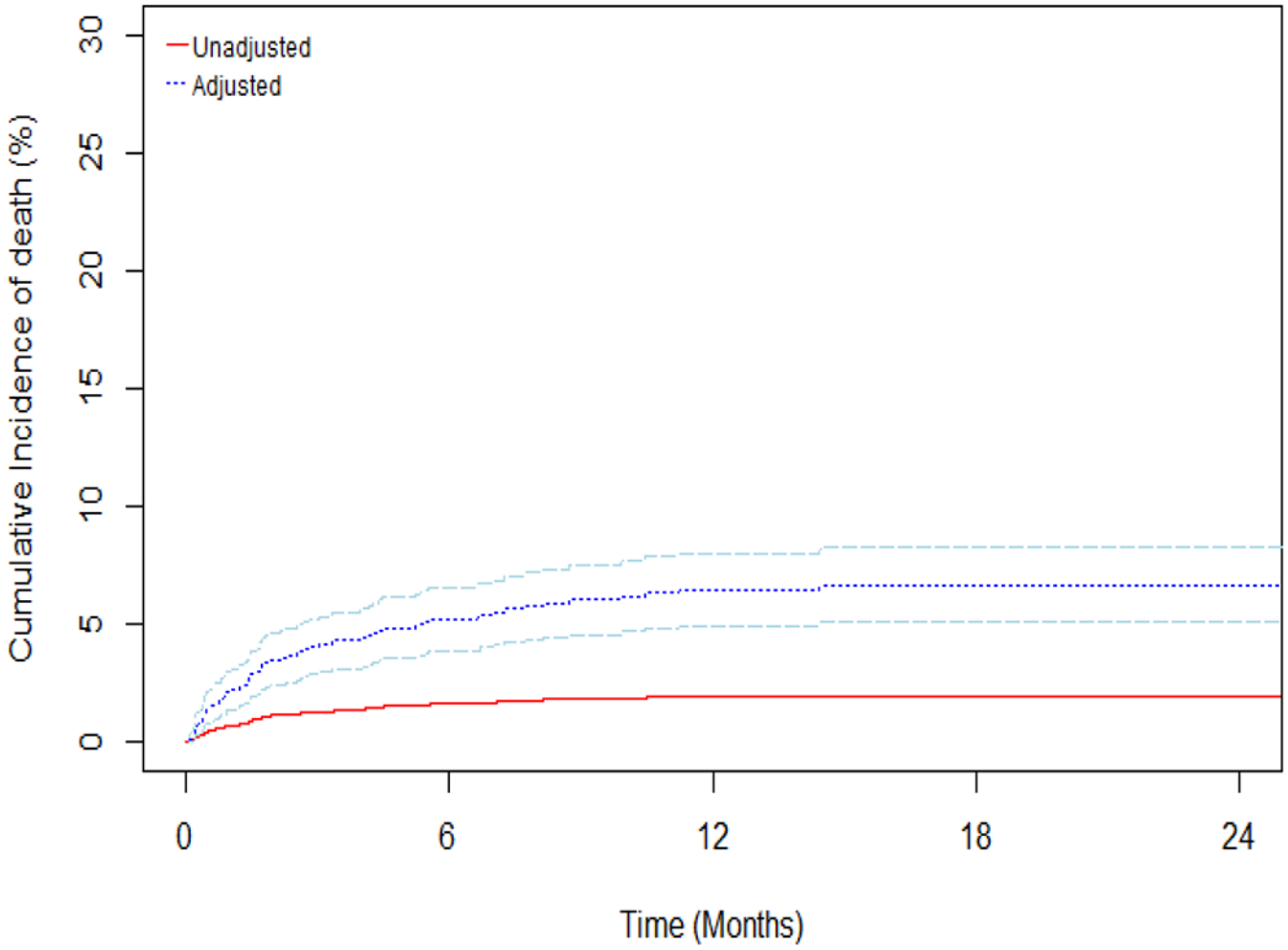

Figure 3.

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.

10 References

An MW et al. (2009) The need for double-sampling designs in survival studies: An application to monitor pepfar. Biometrics. 65 (1), 301–306. [PubMed: 18479488]

Bakoyannis G et al. (2010) Modelling competing risks data with missing cause of failure. Stat Med. 29 (30), 3172–3185. [PubMed: 21170911]

Bakoyannis G & Touloumi G (2012) Practical methods for competing risks data: A review. Statistical methods in medical research. 21 (3), 257–272. [PubMed: 21216803]

Bakoyannis G & Yiannoutsos CT (2015) Impact of and correction for outcome misclassification in cumulative incidence estimation. PloS one. 10 (9), e0137454. [PubMed: 26331616]

Bakoyannis G et al. (2020). Semiparametric regression and risk prediction with competing risks data under missing cause of failure. Lifetime Data Analysis. DOI: 10.1007/s10985-020-09494-1, in press.

Bakoyannis G, Zhang Y and Yiannoutsos CT (2019). Nonparametric inference for Markov processes with missing absorbing state. Statistica Sinica. DOI: 10.5705/ss.202017.0175, in press

Barron BA (1977) The effects of misclassification on the estimation of relative risk. Biometrics. 33 (2), 414–418. [PubMed: 884199]

Beyersmann J et al. (2009) Simulating competing risks data in survival analysis. Stat Med. 28 (6), 956–971. [PubMed: 19125387]

Brinkhof MWG et al. (2010) Adjusting mortality for loss to follow-up: Analysis of five art programmes in sub-saharan africa. PLOS ONE. 5 (11), e14149. [PubMed: 21152392]

Bross I (1954) Misclassification in 2 x 2 tables. Biometrics. 10 (4), 478–486.

Carroll RJ et al. (2006) Measurement error in nonlinear models: A modern perspective. CRC press.

Chen Y-H (2000) Miscellanea. a robust imputation method for surrogate outcome data. Biometrika. 87 (3), 711–716.

Dempster AP et al. (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society. Series B (methodological). 1–38.

Edwards JK et al. (2013) Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. Am J Epidemiol. 177 (9), 904–912. [PubMed: 24627573]

Egger M et al. (2011) Correcting mortality for loss to follow-up: A nomogram applied to antiretroviral treatment programmes in sub-saharan africa. PLoS Med. 8 (1), e1000390. [PubMed: 21267057]

Geng EH et al. (2008) Sampling-based approach to determining outcomes of patients lost to follow-up in antiretroviral therapy scale-up programs in africa. JAMA. 300 (5), 506–507. [PubMed: 18677022]

Gong G & Samaniego FJ (1981) Pseudo maximum likelihood estimation: Theory and applications. The Annals of Statistics. 861–869.

Greenland S (1988) Variance estimation for epidemiologic effect estimates under misclassification. Stat Med. 7 (7), 745–757. [PubMed: 3043623]

Hardt J et al. (2012) Auxiliary variables in multiple imputation in regression with missing x: A warning against including too many in small sample research. BMC Med Res Methodol.12 (1), 184. [PubMed: 23216665]

JK E et al. (2019) Nonparametric estimation of the cumulative incidence function under outcome misclassification using external validation data. Stat Med. 38 (29), 5429–5564. [PubMed: 31647135]

Lin DY et al. (2002) Model-checking techniques based on cumulative residuals. Biometrics. 58 (1), 1–12. [PubMed: 11890304]

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

Lu K & Tsiatis AA (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. Biometrics. 57 (4), 1191–1197. [PubMed: 11764260]

Lyles RH & Lin J (2010) Sensitivity analysis for misclassification in logistic regression via likelyhood methods and predictive value weighting. Stat Med. 29 (22), 2297–2309. [PubMed: 20552681]

Lyles RH et al. (2011) Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. Epidemiology. 22 (4), 589–597. [PubMed: 21487295]

Magder LS & Hughes JP (1997) Logistic regression when the outcome is measured with uncertainty. American Journal of Epidemiology. 146(2), 195–203. [PubMed: 9230782]

Meng X-L (1994) Multiple-imputation inferences with uncongenial sources of input. Statistical Science. 538–558.

Neuhaus JM (1999) Bias and efficiency loss due to misclassified responses in binary regression. Biometrika. 86 (4), 843–855.

Parke WR (1986) Pseudo maximum likelihood estimation: The asymptotic distribution. The Annals of Statistics. 355–357.

Pepe MS (1992) Inference using surrogate outcome data and a validation sample. Biometrika. 79 (2), 355–365.

Putter H et al. (2007) Tutorial in biostatistics: Competing risks and multi-state models. Stat Med. 26 (11), 2389–2430. [PubMed: 17031868]

Robins JM and Wang N (2000). Inference for imputation estimators. Biometrika. 87 (1), 113–124.

Rubin DB (1976) Inference and missing data. Biometrika. 63 (3), 581–592.

Spiegelman D et al. (2001) Efficient regression calibration for logistic regression in main study/ internal validation study designs with an imperfect reference instrument. Stat Med. 20 (1), 139–160. [PubMed: 11135353]

Spiegelman D (2010) Approaches to uncertainty in exposure assessment in environmental epidemiology. 31 (1), 149–163.

Tang L et al. (2015) Binary regression with differentially misclassified response and exposure variables. Stat Med. 34 (9), 1605–1620. [PubMed: 25652841]

Tenenbein A (1970) A double sampling scheme for estimating from binomial data with misclassifications. Journal of the American Statistical Association. 65 (331), 1350–1361.

Tilling K et al. (2016) Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. Journal of clinical epidemiology. 80107–115.

Touloumis A (2016) Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. The R Journal.

Wacholder S (1996) The case-control study as data missing by design: Estimating risk differences. Epidemiology. 144–150. [PubMed: 8834553]

Yiannoutsos CT et al. (2008) Sampling-based approaches to improve estimation of mortality among patient dropouts: Experience from a large pepfar-funded program in western kenya. PLoS One. 3 (12), e3843. [PubMed: 19048109]

Zhang G & Little R (2009) Extensions of the penalized spline of propensity prediction method of imputation. Biometrics. 65 (3), 911–918. [PubMed: 19053998]

Zhao Y et al. (2009) Likelihood methods for regression models with expensive variables missing by design. Biometrical Journal. 51 (1), 123–136. [PubMed: 19197954]