The Annals of Applied Statistics
2020, Vol. 14, No. 2, 1045–1061
https://doi.org/10.1214/20-AOAS1343
© Institute of Mathematical Statistics, 2020

ACCOUNTING FOR DEPENDENT ERRORS IN PREDICTORS AND TIME-TO-EVENT OUTCOMES USING ELECTRONIC HEALTH RECORDS, VALIDATION SAMPLES AND MULTIPLE IMPUTATION

BY MARK J. GIGANTI[1,*], PAMELA A. SHAW[2], GUANHUA CHEN[3], SALLY S. BEBAWY[4,‡], MEGAN M. TURNER[4,§], TIMOTHY R. STERLING[4,¶] AND BRYAN E. SHEPHERD[1,†]

[1]*Department of Biostatistics, Vanderbilt University,* [*]*mark.giganti@vanderbilt.edu;* [†]*bryan.shepherd@vanderbilt.edu*

[2]*Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, shawp@pennmedicine.upenn.edu*

[3]*Department of Biostatistics and Medical Informatics, University of Wisconsin, gchen25@wisc.edu*

[4]*Vanderbilt University School of Medicine,* [‡]*sally.furukawa@vumc.org;* [§]*megan.turner@vumc.org;* [¶]*timothy.sterling@vanderbilt.edu*

Data from electronic health records (EHR) are prone to errors which are often correlated across multiple variables. The error structure is further complicated when analysis variables are derived as functions of two or more error-prone variables. Such errors can substantially impact estimates, yet we are unaware of methods that simultaneously account for errors in covariates and time-to-event outcomes. Using EHR data from 4217 patients, the hazard ratio for an AIDS-defining event associated with a 100 cell/mm^3 increase in CD4 count at ART initiation was 0.74 (95%CI: 0.68–0.80) using unvalidated data and 0.60 (95%CI: 0.53–0.68) using fully validated data. Our goal is to obtain unbiased and efficient estimates after validating a random subset of records. We propose fitting discrete failure time models to the validated subsample and then multiply imputing values for unvalidated records. We demonstrate how this approach simultaneously addresses dependent errors in predictors, time-to-event outcomes, and inclusion criteria. Using the fully validated dataset as a gold standard, we compare the mean squared error of our estimates with those from the unvalidated dataset and the corresponding subsample-only dataset for various subsample sizes. By incorporating reasonably sized validated subsamples and appropriate imputation models, our approach had improved estimation over both the naive analysis and the analysis using only the validation subsample.

1. Introduction.

An alarming number of studies are raising concerns regarding the quality of routinely collected electronic health record (EHR) data and, consequently, misleading findings (e.g., Chan, Fowles and Weiner (2010), Duda et al. (2012), Floyd et al. (2012)). Some errors, such as values falling outside specific ranges, can be identified using computerized data quality checks; other errors are harder to detect. For example, the date of treatment initiation may be incorrectly recorded, but the documented date is within the follow-up period. Furthermore, the existence and magnitude of errors may be correlated across multiple variables. For example, if the treatment initiation date is incorrect, then lab values at the time of treatment initiation and the calculated time from initiation to some event are also likely incorrect. To identify such errors, all relevant error-prone variables would have to be verified; however, such a resource-intensive process may not be feasible in settings with limited funding.

An alternative to verification of all records is to perform data audits or validation in a subset of records. This is generally done by selecting a random set of records and verifying

Received August 2018; revised March 2020.

Key words and phrases. Correlated errors, electronic health records, measurement error, multiple imputation, survival analysis.

data accuracy for key variables. If nontrivial error rates are revealed, one might remove the data in question or reenter all data. These options, however, seem unsatisfactory. A more appealing option would be to incorporate the audit or validation data into the analysis.

The data available following an audit—an error-prone measurement for all records and a "gold standard" measurement for a subset of records—resembles the data one might need to correct for measurement error. While the statistical literature regarding measurement error is substantial, most methods involving time-to-event data focus only on covariate measurement error. These methods include regression calibration (Prentice (1982), Shaw and Prentice (2012)), corrected score methods (Nakamura (1990), Huang and Wang (2000)), conditional score methods (Tsiatis and Davidian (2001)), joint models (Wulfsohn and Tsiatis (1997)) and SIMEX (Cook and Stefanski (1994), Li and Lin (2003)). There have also been select studies related to time-to-event outcome measurement error, with methods corresponding to errors in event indicators (Magaret (2008), Richardson and Hughes (2000), Hunsberger, Albert and Dodd (2010)) or the failure time (Skinner and Humphreys (1999), Korn, Dodd and Freidlin (2010)); unlike linear regression, unbiased errors in failure times result in biased estimates (Oh et al. (2018)). While correlated errors in covariates and uncensored outcomes have been previously considered (Shepherd and Yu (2011), Shepherd, Shaw and Dodd (2012)), no existing methods address situations with errors in both the covariates and the time-to-event outcome. Given that errors in EHR data typically occur across multiple variables and these errors are generally correlated, the current measurement error literature is not equipped to handle such multidimensional errors seen in practice.

Measurement error with a validation subsample can also be thought of as a missing data problem (Little and Rubin (2014)). Because audited records are typically selected by a simple random sample or a random sample stratified on observed data (e.g., unvalidated disease status), the missing data mechanism is missing at random and standard methods for addressing missing data could be applicable. For example, a multiple imputation (MI) approach could be employed by fitting models using the complete validated data and imputing missing values for unvalidated records. This approach has been implemented in previous studies for various measurement error scenarios, including mismeasured binary (Edwards et al. (2013)) and continuous (Shepherd, Shaw and Dodd (2012)) outcomes as well as measurement error in the exposure of a time-to-event outcome (Cole, Chu and Greenland (2006)). Although relevant, none of this work considers the situation where there are errors, likely correlated, in both predictors and the time-to-event outcome. Furthermore, it does not address errors in indicators of patient eligibility that determine whether a patient should be included in the analysis. These situations, common in practical applications of EHR data for analyses, add considerable complexity.

In this manuscript we describe and implement a multiple imputation-based strategy to account for correlated errors in both predictors and time-to-event outcomes as well as analysis eligibility. We illustrate our approach using unvalidated and validated EHR data from a large HIV clinic, the Vanderbilt Comprehensive Care Clinic (VCCC). In Section 2 we present our motivating example. In Section 3 we formalize our problem and present our strategy for obtaining improved estimates after partial data validation. We first assign a notation for the variables and errors in our motivating example, and then we coarsen these variables into discrete-time for analysis. In Sections 4 and 5 we demonstrate and evaluate our approach using data from the VCCC and simulations. In Section 6 we discuss our results and suggest areas for future research.

2. Motivating example.

In this study we analyzed data on 4217 HIV-positive patients who established care at the VCCC between 1998 and 2011. Briefly, the VCCC is an outpatient clinic that provides primary and subspecialty care for persons living with HIV. As part of routine treatment and care, data relevant to the patient's clinical experience were collected over time. This included both time-invariant variables such as demographic characteristics as well as time-varying variables corresponding to laboratory measurements (e.g., CD4 counts), pharmacy dispensations, opportunistic infections and vital status. Data before enrollment were also recorded, usually during the initial visit, based on patient recall and outside medical records. The median length of follow-up after enrollment was 3.2 years (interquartile range [IQR]: 1.1–6.8). The majority of patients were male (76%), and the median age at enrollment was 38 years (IQR: 31–45).

Data at the VCCC were collected and electronically recorded by health care providers, typically nurses and physicians. Research protocols mandated that chart reviews were performed for all VCCC records to validate key variables. A team of data abstractors performed the data validation. After this comprehensive chart review process, two datasets were available. The first dataset, which we refer to as the unvalidated dataset, contained the values entered for all 4217 records prior to the chart review. The second dataset, which we refer to as the validated dataset, contained the recorded values for the same 4217 records after thorough chart review. Throughout this study we consider the validated dataset to be correct.

For this study, we considered the association between CD4 count at time of antiretroviral therapy (ART) initiation and the time from ART initiation until first AIDS-defining event (ADE). Specifically, we calculated the incidence of ADE using Kaplan–Meier methods and the hazard ratio (HR) for a 100 cell/mm^3 increase in CD4 count at ART initiation using a multivariable Cox proportional hazards regression model. All patients included in the analysis cohort were adults ($\geq$ 18 years) at time of ART initiation. Patients were excluded if they started ART prior to enrollment, had an indeterminate ART start date or had a documented ADE prior to ART initiation. These inclusion and exclusion criteria are common for HIV studies.

We performed the same statistical analysis for both the unvalidated and validated datasets. The incidence of ADE was higher in the unvalidated dataset across the entire study period. The estimated incidence of ADE at five years was 19.7% (95% confidence interval [CI]: 17.1%–22.2%) for the unvalidated dataset and 8.3% (95% CI: 6.6%–10.0%) for the validated dataset. A 100 cell/mm^3 increase in CD4 count at ART initiation was associated with a much weaker decrease in the hazard of ADE in the unvalidated dataset (HR: 0.74; 95%CI: 0.68–0.80) compared to the validated dataset (HR: 0.60; 95%CI: 0.53–0.68).

There were many discrepancies between the unvalidated and validated datasets. In the unvalidated dataset 1743 patients satisfied the criteria for inclusion in the analysis cohort. In the validated dataset 1580 patients met all inclusion criteria. A total of 1392 met the inclusion criteria for both analysis cohorts, suggesting 539 (13%; 351 wrongly included and 188 wrongly excluded) patients were incorrectly classified in the unvalidated dataset. Among patients included in both analysis cohorts, there was discordancy in variables indicating baseline CD4 count (5%), ADE status (9%) and time from ART initiation to an ADE or end of study (32%). Table 1 includes further details comparing the unvalidated and validated datasets.

As with most studies using EHR data, the variables used in our analyses were primarily derived variables (e.g., baseline CD4 was determined by identifying the laboratory measurement in one table with a date closest to the first ART initiation date in a separate table). Discrepancies in derived variables were mostly due to errors in the indicators and dates of ART and ADE. Among all 4217 patients there were 1745 (41%) patients with an incorrect ART start date and 1253 (30%) patients with an incorrect ADE (or end of follow-up) date. All discrepancies in the baseline CD4 count were due to discrepancies in the ART start date.

TABLE 1
Comparison of variables in the unvalidated and validated datasets among the 4217 *patients*

	Notation (see Section 3)	Discrepancy magnitude n or median(IQR)
All patients		4217
Different ART start date	$T_0 \neq T_0^*$	1745 (41.4%)
Discrepancy in ART start dates (days)	$T_0 - T_0^* \mid T_0 \neq T_0^*$	14 (−222, 37)
Different ADE date	$T_E \neq T_E^*$	1253 (29.7%)
Discrepancy in ADE date (days)	$T_E - T_E^* \mid T_E \neq T_E^*$	14 (−5, 165)
Met inclusion criteria in both datasets	$W = 1, W^* = 1$	1392
Different ADE status	$D \neq D^*$	130 (9.3%)
ADE in unvalidated, no ADE in validated	$D^* = 1, D = 0$	116
ADE in validated, no ADE in unvalidated	$D^* = 0, D = 1$	14
Different time from ART initiation to ADE	$Y \neq Y^*$	441 (31.7%)
Discrepancy in time from ART to ADE (days)	$Y - Y^* \mid Y \neq Y^*$	1 (−357, 20)
Different baseline CD4 count	$X_1 \neq X_1^*$	76 (5.4%)
Discrepancy in baseline CD4	$X_1 - X_1^* \mid X_1 \neq X_1^*$	22 (−23, 88)

Abbreviations: ADE, AIDS-defining event. ART, antiretroviral therapy. IQR, interquartile range.

3. Our approach.

In the previous section we showed that estimates using just the unvalidated dataset were markedly biased. These findings highlight, at least in our setting, the importance of validating EHR data. Our goal in this study is to obtain low bias and low variance estimates after validating only a subsample of the EHR. In this section we formalize the problem analytically and describe our analysis approach.

3.1. *Notation.* Let T_B denote the date of enrollment, T_0 the date of ART initiation, T_E the date of first ADE and T_C the last follow-up ("end of study") date. Note that some patients will not initiate ART or experience an ADE during the study; in such instances, as is convention in time-to-event analyses, T_0 and T_E are assumed to be unobserved dates after T_C. Using these dates, we derive the variables corresponding to the outcome: the time from ART initiation to the first of ADE or end of study, $Y = \min(T_E, T_C) - T_0$ and an indicator of an ADE, $D = I(T_E \leq T_C)$.

Let $X(t) = (X_1(t), X_2(t), \dots, X_p(t))$ denote a vector of p covariates for a patient on a given date, t. For example, let $X_1(t)$ denote CD4 count on date t. While this notation allows each covariate to change values over time, we note that some covariates may be time-invariant. Since we are interested in values at time of ART initiation (T_0), we define a vector of "baseline" variables as $X = X(T_0) = (X_1(T_0), X_2(T_0), \dots, X_p(T_0))$, where $X_1(T_0)$ corresponds to baseline CD4 count and $X_2(T_0), \dots, X_p(T_0)$ correspond to the remaining baseline covariate values. Finally, let W denote whether a patient was included in the analysis cohort. Patients were included if they started ART after enrollment ($T_B \leq T_0 < T_C$) and if they did not have an ADE before starting ART ($T_0 < T_E$). The quadruplet (W, X, Y, D) represents the data for our time-to-event analyses from the validated records (i.e., the gold standard).

In our specific application, among those meeting inclusion criteria ($W = 1$), we are interested in estimating the incidence of an event at time t, $P(T_E - T_0 \leq t)$ and the hazard ratio in the proportional hazards model, $\lambda(t \mid X) = \lambda_0(t) \exp(\beta X)$. In our example the last follow-up may be determined by a number of different censoring mechanisms (administrative censoring, lost to follow-up, death). Except where specified differently, the date of the last follow-up T_C is considered to be the minimum time amongst these potential events. While deaths may act as a competing event, this approach reflects the common practice of investigators treating deaths as noninformative, censored observations when the frequency of death is low. Note that even with noninformative censoring, Kaplan–Meier estimates of incidence are biased if death is treated as a censoring event because it is not possible for someone to have an ADE after death; however, for small numbers of deaths, this bias can be small and that is why it is often ignored in practice. In contrast, Cox regression may still be appropriate if one is interested in cause-specific hazards.

Since patient records are potentially error-prone, we use asterisks to denote data from the unvalidated records. Let the unvalidated time of ART initiation be T_0^*, the unvalidated time of ADE be T_E^*, the end of study date in the unvalidated dataset be T_C^* and the date-specific vector of covariates in the unvalidated dataset be $X^*(t)$.

The derived variables corresponding to the outcome in the unvalidated dataset are $D^* = I(T_E^* \leq T_C^*)$ and $Y^* = \min(T_E^*, T_C^*) - T_0^*$. The unvalidated baseline predictor variables are defined as $X^* = X^*(T_0^*)$. Let W^* be an indicator for inclusion, defined as $I(T_B \leq T_0^* < T_C^*)I(T_0^* < T_E^*)$. We denote the unvalidated data used for analyses as the quadruplet (W^*, X^*, Y^*, D^*).

Finally, let $V = 1$ denote that data validation was performed for all variables. For records with $V = 1$, we have (W^*, X^*, Y^*, D^*) and (W, X, Y, D), whereas for those records with $V = 0$ we have only (W^*, X^*, Y^*, D^*). In our VCCC example, $V = 1$ for all records, so an analyst would ignore the error-prone unvalidated data and draw inference using only the validated data. We will consider the situation where $V = 1$ for only a subsample of patients.

As highlighted in the Introduction, there are methods for time-to-event outcome studies regarding covariate (X^*, Y, D), event indicator (X, Y, D^*) or time-to-event (X, Y^*, D) measurement error. However, methods for simultaneously dealing with errors in predictors, event indicators and times-to-event (X^*, Y^*, D^*) have not been considered, and because of potential dependence between these errors, it is not possible to simply sequentially apply existing methods. Furthermore, for our motivating example we also need to consider errors with the inclusion criteria (W^*, X^*, Y^*, D^*).

3.2. *Multiple imputation*: *Model fitting and time discretization.* Our strategy is to approach this as a missing data problem where the quadruplet (W^*, X^*, Y^*, D^*) is available for all records and the quadruplet (W, X, Y, D) is missing for those with $V = 0$. This requires the construction of a model for the joint distribution of (W, X, Y, D) conditional on (W^*, X^*, Y^*, D^*). The model will be fit using a subsample of records with $V = 1$; values for the remaining records $(V = 0)$ will be imputed using these fitted models. Therefore, the primary challenge is obtaining adequate models.

Consider the factorization of the distribution of (W, X, Y, D) conditional on (W^*, X^*, Y^*, D^*),

$$\begin{aligned} f(W, X, Y, D \mid W^*, X^*, Y^*, D^*) = {} & f(W \mid W^*, X^*, Y^*, D^*) \\ & \times f(X \mid W, W^*, X^*, Y^*, D^*) \\ & \times f(Y \mid W, X, W^*, X^*, Y^*, D^*) \\ & \times f(D \mid W, X, Y, W^*, X^*, Y^*, D^*), \end{aligned} \tag{3.1}$$

where $f(\cdot)$ denotes a generic probability density/mass function. With this factorization each component of (3.1) could be fit using existing data from the subsample of records with $V = 1$. However, these models would be constructed using derived variables that are functions of other error-prone variables, $(T_0, T_E, T_C, X(t))$, likely making their predictive ability poor. For example, errors in (W, X, Y) are frequently due to errors in the date of ART initiation (T_0), and a marked drop in a time-varying covariate, viral load, is highly predictive of T_0 that someone has begun ART; however, it is unclear how to incorporate such information into models based on the derived variables (W, X, Y).

Another strategy, which we adopt here and refer to as time-discretized modeling, divides time into intervals (e.g., months) and assesses values for variables during each interval. This approach employs a well-known strategy for modeling time-to-event data using pooled logistic regression (D'Agostino et al. (1990), Efron (1988)). Similar approaches have been implemented with marginal structural models (Hernán, Brumback and Robins (2001)) and ecological statistics (McClintock et al. (2014), Turchin (1998)), where discretization is used to allow for time-varying covariates and to reduce computationally-intensive tasks.

Here, variables are divided into monthly intervals, indexed by m, since the date of enrollment ($m = 0$). Specifically, let $\mathcal{A}_m$ be an indicator for a patient initiating at least one different ART drug during month m; if a patient is not on ART or continues the same ART regimen as the previous month, they are assigned $\mathcal{A}_m = 0$. Let $\mathcal{D}_m$ be an indicator of an ADE occurring during month m. Let $\mathcal{X}_m$ correspond to the most recent covariate values observed during month m, and, finally, let $\mathcal{C}_m$ be an indicator that the last follow-up visit for a patient occurred during month m.

Let $\overline{\mathcal{A}} = \{\mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_{M_{\text{post}}}\}$ designate the complete set of monthly new ART drug indicators in the validated dataset, where M_{post} denotes the longest possible length of follow-up (in months) among all patients. The variables $\overline{\mathcal{D}}, \overline{\mathcal{C}}$ and $\overline{\mathcal{X}}$ are similarly defined. For the unvalidated dataset we have $\overline{\mathcal{A}^*}$, $\overline{\mathcal{D}^*}$, $\overline{\mathcal{C}^*}$ and $\overline{\mathcal{X}^*}$. This notation implicitly assumes that the date of enrollment is correct in the unvalidated dataset; this assumption is met in the VCCC dataset but could be relaxed by using some other date to anchor time.

With this framework we can construct a model for the joint distribution of the variables in the validated dataset $(\overline{\mathcal{X}}, \overline{\mathcal{C}}, \overline{\mathcal{A}}, \overline{\mathcal{D}})$ conditional on the distribution of the variables in the unvalidated dataset $(\overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$ by decomposing it into separate components,

$$\begin{aligned} f(\overline{\mathcal{X}}, \overline{\mathcal{C}}, \overline{\mathcal{A}}, \overline{\mathcal{D}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) &= f(\overline{\mathcal{X}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ &\quad \times f(\overline{\mathcal{C}} \mid \overline{\mathcal{X}}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ &\quad \times f(\overline{\mathcal{A}} \mid \overline{\mathcal{C}}, \overline{\mathcal{X}}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ &\quad \times f(\overline{\mathcal{D}} \mid \overline{\mathcal{A}}, \overline{\mathcal{C}}, \overline{\mathcal{X}}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}). \end{aligned} \tag{3.2}$$

With this decomposition we directly model discretized versions of the original variables that are in error, rather than downstream, derived variables. By incorporating error-prone and corrected variables in models, we account for potential dependencies in errors across variables. Time-varying covariates are also easier to incorporate. For example, the probability of starting a new ART regimen in a given month, $\mathcal{A}_m$, can be modeled conditional on the unvalidated indicator of starting a new ART regimen for that month, $\mathcal{A}_m^*$, and time-varying covariates $\overline{\mathcal{X}}$ such as viral load prior to, during and after month m. Specific implementation details are given in the next section.

Component models can be fit using the records with validated data (i.e., those with $V = 1$) using appropriate methods (e.g., binary variables can be modeled using logistic regression). When predicting values for the remaining records (i.e., those with $V = 0$), we account for the uncertainty in the prediction model using a multiple imputation procedure. First, we draw an independent sample of the parameter estimates of the fitted models (e.g., we sample from a multivariate normal distribution with the mean as the parameter estimates and variance as the variance-covariance matrix of the parameter estimates). Using $(\overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$ and these randomly drawn parameter estimates, we impute values of $(\overline{\mathcal{X}}, \overline{\mathcal{C}}, \overline{\mathcal{A}}, \overline{\mathcal{D}})$ for all records with $V = 0$.

Having successfully imputed discretized versions of the original variables, values are converted back to the unit of measurement of the original variables using a fixed conversion.

For example, an imputed ART initiation two months after enrollment would be reported as 60 days after enrollment. We choose to "undiscretize" our imputed values to incorporate full information from our validated data in the time-to-event analyses. However, this undiscretization is not necessary; the impact of this choice will be evaluated in sensitivity analyses. With these imputed "undiscretized" values, we derive imputed versions for the variables to be used in our Cox regression and Kaplan–Meier analyses, denoted as $(\widehat{W}, \widehat{X}, \widehat{Y}, \widehat{D})$. Thus, we generate a complete dataset consisting of the true, observed values of the audited records and the predicted values of the unaudited records, denoted as

$$(W^c, X^c, Y^c, D^c) = \begin{cases} (\widehat{W}, \widehat{X}, \widehat{Y}, \widehat{D}) & \text{if } V = 0, \\ (W, X, Y, D) & \text{if } V = 1. \end{cases}$$

We then repeat the process of randomly sampling parameter estimates, predicting values and combining datasets, until we have B complete datasets. Here, B is the number of imputations performed. For each of the B complete datasets, we obtain estimates using Kaplan–Meier and Cox regression methods. The parameter estimates from these procedures are then averaged across iterations. To properly account for uncertainty in the setting of incompatible imputation and analysis models, we use the multiple imputation variance estimator proposed by Robins and Wang (2000) to calculate confidence intervals.

For this time-discretized modeling and imputation (TDMI) approach to yield unbiased estimates in large samples, standard assumptions for the validity of multiple imputation must be met (Van Buuren (2018)). The missing at random assumption can be translated as $V \perp\!\!\!\perp (\overline{\mathcal{X}}, \overline{\mathcal{C}}, \overline{\mathcal{A}}, \overline{\mathcal{D}}) \mid (\overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$, or that conditional on the observed data, selection for validation (V) is independent of the correct values. Another key assumption is that the imputation model, $f(\overline{\mathcal{X}}, \overline{\mathcal{C}}, \overline{\mathcal{A}}, \overline{\mathcal{D}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$, is properly specified. This requires the identification of covariates in the unvalidated dataset predictive of values in the validated dataset as well as a model that properly specifies the relationships. When the imputation model is not properly specified, parameter estimates may be biased and confidence interval coverage probabilities may differ from nominal levels (Carpenter, Kenward and Vansteelandt (2006), McIsaac and Cook (2017)). The TDMI approach handles differential measurement error through the imputation model (i.e., it requires no assumption of nondifferential measurement error), but covariates associated with differential error need to be correctly included in the model. Finally, because we are estimating parameters defined on an approximately continuous time scale (days) after imputing data from models on a discrete time scale (months), we assume that the discrete time scale is a good approximation to the continuous time scale which has been seen by others to be the case as long as the unit of discretized time is not too coarse (e.g., D'Agostino et al. (1990), Efron (1988)).

3.3. *Implementation details.* In this section we highlight key simplifications and noteworthy specifications that were made in our application of the TDMI to EHR data from the VCCC. Full model details are in the posted analysis code, http://biostat.mc.vanderbilt.edu/ArchivedAnalyses.

The end of study date for each patient did not vary between the unvalidated and validated records, and, thus, we did not need to model $\overline{\mathcal{C}}$ as it was perfectly predicted by $\overline{\mathcal{C}^*}$. This simplification allowed us to model

$$\begin{aligned} f(\overline{\mathcal{X}}, \overline{\mathcal{A}}, \overline{\mathcal{D}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) = {} & f(\overline{\mathcal{X}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ & \times f(\overline{\mathcal{A}} \mid \overline{\mathcal{X}}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ & \times f(\overline{\mathcal{D}} \mid \overline{\mathcal{A}}, \overline{\mathcal{X}}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}). \end{aligned} \tag{3.3}$$

Laboratory measurements (e.g., CD4 count and viral load) also did not vary between the unvalidated and validated records. Thus, we also did not need to model $\overline{\mathcal{X}}$ as it was perfectly predicted by $\overline{\mathcal{X}^*}$. However, these time-varying variables were not necessarily collected monthly. In most instances we carried forward values from months where previous measurements were available. In instances where no laboratory measurement was available at the time of enrollment, we needed to input values. Let $\mathcal{X}^*_{1,0}$ and $\mathcal{X}^*_{2,0}$ denote CD4 count and viral load at time of enrollment, respectively, and denote the remaining covariates that comprise $\overline{\mathcal{X}^*}$ as $\overline{\mathcal{X}^{C*}}$. Our imputation model for these two covariates was

$$\begin{aligned} f(\mathcal{X}^*_{1,0}, \mathcal{X}^*_{2,0} \mid \overline{\mathcal{X}^{C*}}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) &= f(\mathcal{X}^*_{1,0} \mid \overline{\mathcal{X}^{C*}}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ &\quad \times f(\mathcal{X}^*_{2,0} \mid \mathcal{X}^*_{1,0}, \overline{\mathcal{X}^{C*}}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}). \end{aligned} \tag{3.4}$$

Because we were only interested in the time of first ART initiation, not all subsequent ART changes, we were able to model the time of first ART initiation directly. Specifically, let $\mathcal{A}^1_m = \max_{k \le m}(\mathcal{A}_k)$ be the indicator that ART had been initiated prior to or during month m. Instead of $f(\overline{\mathcal{A}} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$, we used $f(\overline{\mathcal{A}^1} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*})$ as our model of ART status. This can be further simplified to

$$\begin{aligned} f(\overline{\mathcal{A}^1} \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) &= f(\mathcal{A}^1_0 \mid \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) \\ &\quad \times \prod_{m=1}^{M_{\text{post}}} f(\mathcal{A}^1_m \mid \mathcal{A}^1_{m-1}, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}), \end{aligned} \tag{3.5}$$

where $\Pr(\mathcal{A}^1_m = 1 \mid \mathcal{A}^1_{m-1} = 1, \overline{\mathcal{X}^*}, \overline{\mathcal{C}^*}, \overline{\mathcal{A}^*}, \overline{\mathcal{D}^*}) = 1$. Note that in this model we conditioned on $\overline{\mathcal{A}^*}$, the unvalidated vector of new ART drug indicators, rather than $\overline{\mathcal{A}^{1*}} = \{\mathcal{A}^{1*}_0, \mathcal{A}^{1*}_1, \dots, \mathcal{A}^{1*}_{M_{\text{post}}}\}$, the unvalidated vector of the indicator of having initiated ART, because $\overline{\mathcal{A}^*}$ is richer than $\overline{\mathcal{A}^{1*}}$ and may improve modeling (e.g., if the first date of ART initiation in the unvalidated data is incorrect, the second date of ART initiation in the unvalidated data might be a good candidate for the true first date of ART initiation). This model was fit using pooled logistic regression.

Although we were similarly interested in the first date of ADE, we chose to model all ADEs (i.e., the complete vector $\overline{\mathcal{D}}$) rather than just focusing on the first. Unlike ART status, the variables associated with a given ADE were not likely to differ based on the ordering of the ADE. Because ADE at a specific time is a binary variable, logistic regression was again used for model fitting.

Many VCCC patient records included data for months prior to enrollment; for example, dates of ART used prior to enrollment may have been included in the patient record. It was important to include this information in the analysis (e.g., a patient starting ART prior to enrollment does not meet analysis eligibility criteria). Thus, the time-discretized variables included time prior to enrollment, for example, $\overline{\mathcal{A}} = \{\mathcal{A}_{M_{\text{pre}}}, \dots, \mathcal{A}_{-1}, \mathcal{A}_0, \mathcal{A}_1, \dots, \mathcal{A}_{M_{\text{post}}}\}$ where M_{pre} designated the longest length of preenrollment follow-up among all patients. For this study $M_{\text{pre}} = -100$ and $M_{\text{post}} = 167$. We constructed separate imputation models for before ($m < 0$) and after ($m \ge 0$) enrollment. Therefore, we fit a total of six imputation models: two linear regression models (CD4 count and viral load at enrollment) and four logistic regression models (ART initiation prior to enrollment, ART initiation on or after enrollment, ADE prior to enrollment and ADE on or after enrollment).

A total of 30 covariates, $\overline{\mathcal{X}^*}$, were used for the imputation models based on their clinical relevance and a priori belief that they might be predictive of validated values. Time-invariant covariates (calendar year of enrollment, age at first visit and sex) were attributed to each person-month observation. Time-varying covariates included months since enrollment, current CD4, previous and next CD4, current viral load, previous and next viral loads. We note that only a select number of covariates were considered relevant predictors for all six models. Models used to impute missing viral load and CD4 count values at time of enrollment only included time-invariant covariates and lab measurements at enrollment. For the ART and ADE models restricted to the months preceding enrollment, no lab measurements were included. For the ART status model, additional time-varying covariates included an indicator for any ART drug initiation during the entirety of follow-up. Since the type of ADE in the unvalidated dataset was highly predictive of the presence or absence of an ADE in the validated data, we included dummy variables corresponding to 14 specific ADEs in the unvalidated data for the ADE status model. All continuous variables were modeled using restricted cubic splines. Twenty iterations were used for the MI procedure.

Following imputation, values were switched from a discrete time scale to a continuous time scale. When undiscretizing imputed dates of ART or ADE, an imputed event date was assigned to the first day in the monthly interval. This was due to practical considerations as most ART initiations occurred on the same day as enrollment and we wanted imputed ART initiations during that monthly interval to map back to day 0 rather than say day 15 or 30. The cumulative incidence of ADE over time was estimated using the Kaplan–Meier method. To model the association between baseline CD4 and time to ADE, we fit a multivariable Cox regression model that adjusted for sex, age and year of enrollment. We applied a square-root transformation to the baseline CD4 count prior to model fitting and reported the estimated hazard ratio corresponding to an increase in baseline CD4 count from 100 cells/mm^3 to 200 cells/mm^3 (i.e., a 100 cell/mm^3 increase). The Efron method was used to handle ties. Due to the relatively infrequent occurrence of death prior to ADE, deaths prior to ADE were treated as censored in both Kaplan–Meier and Cox regression analyses.

Several sensitivity analyses were also performed. The TDMI approach included a step where we switch back from the discrete time scale to the continuous scale prior to analysis. In one sensitivity analysis we fit a multivariable pooled logistic regression model for the association between baseline CD4 and time to ADE using the imputed discrete-time variables. As a separate sensitivity analysis we implemented the TDMI approach using reduced imputation models with just the unvalidated variables corresponding to ART and ADE for a given month and six predictor variables (sex, age at first visit, months since enrollment, year of enrollment, current CD4 and current viral load). Finally, we also estimated the cumulative incidence of ADE treating death as a competing risk.

4. Results. For this section we applied our TDMI procedure to data from the VCCC and compared its performance with estimates obtained via alternative strategies. Specifically, we selected a simple random sample of patient records to serve as our validation subsample. For the records that were not randomly selected, we ignored the validation data (i.e., we pretended that no validation data were available). We then applied our TDMI approach using the unvalidated data on all records together with the validation subsample. TDMI estimates were compared to the gold standard postaudit estimates using the fully validated data for all patient records. Additionally, estimates from our multiple imputation approach were compared to two nonimputation-based estimates: the naive preaudit estimates using only the unvalidated data for all patient records and estimates calculated using only the subset of validated records (i.e., complete-case estimates).

In Figure 1 we show the estimated cumulative incidence of ADE over time using the unvalidated dataset, the fully validated dataset, the complete-case analysis using a simple random sample of 1000 validated records and the TDMI strategy using the same 1000 validated records. For this particular sample the TDMI estimates appeared to have similarly

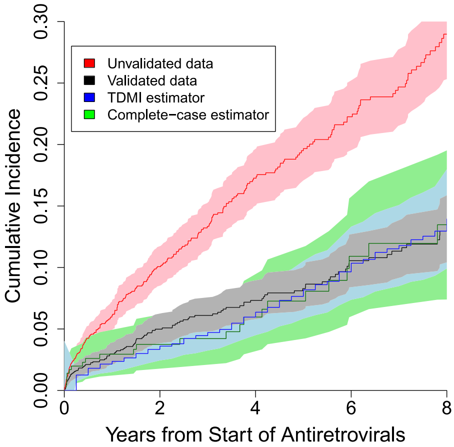

FIG. 1. *Estimated cumulative incidence of AIDS-defining event over time using unvalidated, validated, time-discretized modeling and imputation (TDMI) and complete-case approaches. Estimates for TDMI and complete-case approaches are based on one randomly selected iteration.*

small bias but narrower confidence intervals compared to the complete-case estimates. Both the complete-case and the TDMI estimates were closer to the gold standard estimates than the naive estimates at most time points. Specifically, the TDMI estimate of the incidence of ADE at five years was 8.2% (95% CI: 5.8%–10.5%) and the complete-case estimate was 7.3% (95% CI: 3.8%–10.7%), compared to the naive estimate of 19.7% (95% CI: 17.1%–22.2%) and the gold standard estimate of 8.3% (95% CI: 6.6%–10.0%).

Similarly, the estimated HR for the association between a 100 cell/mm^3 increase in baseline CD4 count and ADE was 0.59 (95% CI: 0.53–0.66) for the TDMI approach and 0.51 (95% CI: 0.39–0.68) for the complete-case approach compared to the naive estimate of 0.74 (95%CI: 0.68–0.80) and the gold standard estimate of 0.60 (95%CI: 0.53–0.68). The TDMI estimate based on a pooled logistic regression model that kept time discretized was 0.60 (95% CI: 0.50–0.73).

While promising, these results were based on a single validation sample. Because we had the fully validated data on all patient records, we were able to repeat the process many times, compare estimates to the fully validated data and empirically study the performance of our TDMI approach. To quantitatively compare approaches, we calculated the difference and the squared difference between each candidate estimate of the five-year incidence and the log HR to their corresponding estimates based on the complete validated data (i.e., the gold standard estimates) for 1000 replications. The mean difference (bias), variance and mean squared difference (mean squared error; MSE) for each candidate estimator (TDMI, complete-case and naive) were then calculated.

Using an audit size of 1000, the MSE of the TDMI estimator for ADE incidence at five years was similar but slightly lower than that of the complete-case estimator (2.1×10^{-4} vs. 2.4×10^{-4}). This result was driven by the TDMI estimator's lower variance (6.6×10^{-5} vs. 2.4×10^{-4}), despite a larger absolute bias (-0.0120 vs. 0.0004). The MSE for the TDMI estimator for the log HR was substantially lower than that of the complete-case estimator (0.003 vs. 0.015). The bias and variance of the TDMI estimator for the log HR were 0.030 and 0.002, respectively, compared with 0.001 and 0.015 for the complete-case estimator. The MSE for the TDMI approach using a pooled logistic regression model that kept time on a discrete scale yielded a similar MSE (0.004).

Our original selection of an audit size of 1000 records was chosen a priori but was arbitrary. To assess how the TDMI approach performed for varying audit sizes, we repeated the entire exercise for various audit sample sizes, ranging from $n = 300$ to $n = 4000$ records. The MSEs

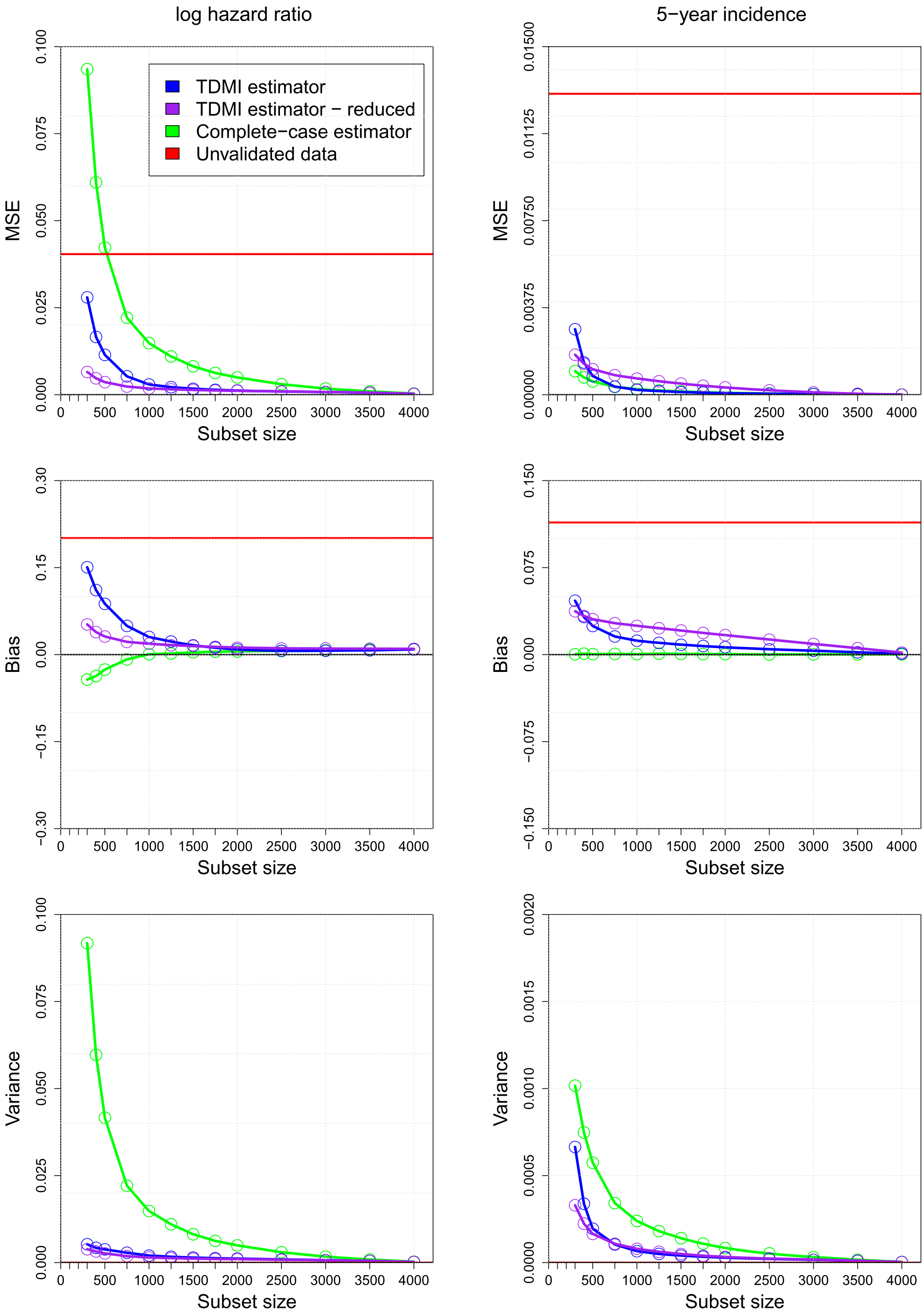

FIG. 2. *Mean squared errors* (*top row*), *bias* (*middle row*) *and variance* (*bottom row*) *for estimates of the log hazard ratio* (*first column*) *and five-year incidence of AIDS-defining event* (*second column*) *from each candidate estimator* (*including a TDMI estimator where fewer predictor variables are included in the imputation model*) *and various audit sizes. Estimates are calculated as the average of* 1000 *replications.*

for the estimated log HR and incidence at five years across various audit sizes are shown in Figure 2 as well as a bias-variance decomposition of MSEs. At all validation sample sizes, the TDMI estimators had lower MSE than the naive estimators. For the log HR, the TDMI estimator beat the complete-case estimator at all audit sample sizes. When estimating the incidence of ADE at five years, the MSE for the TDMI approach tended to be higher than the complete-case analysis for audit sample sizes less than 750 but slightly lower thereafter. In general, the TDMI estimator was less variable but more biased, particularly at the smaller audit sizes, than the complete-case estimator.

Figure 2 also includes results using the TDMI approach with a reduced imputation model that included fewer predictor variables. TDMI estimates of the log HR using the reduced imputation models had substantially lower MSE than the full model when the audit sample size was small (750 or less) but fairly similar MSE thereafter; the MSEs remained lower relative to the complete case estimator at all subset sample sizes. In contrast, the TDMI approach based on these reduced imputation models performed worse when estimating the incidence of ADE at five years at all audit sample sizes above 400. The bias of these new estimates was such that the TDMI estimator had a higher MSE than the complete case estimator at all audit sample sizes. These results are intuitive and highlight the challenges of model fitting at varying audit sizes. The original model was chosen for an audit size of 1000 records; with small audit sizes, fitting similarly complicated models can lead to over-fitting and resulting bias, as seen with the log HR. In contrast, in the reduced model we did not include specific types of ADEs which were very predictive of having any ADE; therefore, the estimated incidence of ADE from the reduced model was more biased, likely due to poor model specification.

Among the 1580 patients in the validated analysis dataset, there were 99 (6.3%) deaths that occurred prior to an ADE. Within the first five years, there were 65 (4.1%) deaths prior to an ADE. We also considered a time-to-event analysis model where death was treated as a competing risk. Using the full validated dataset, cumulative incidence estimates from the competing risks model were similar to Kaplan–Meier estimates (Web Figure 2). We again calculated the incidence of ADE at five years for varying audit sizes using the TDMI approach with the full and reduced imputation models as well as the complete case-approach. In Web Figure 3 we show the bias, variance and MSEs across various audit sizes. Findings were similar to those regarding the Kaplan–Meier estimates of the cumulative incidence of ADE at five years. The MSE for the TDMI approach tended to be higher than the complete-case analysis for audit sample sizes less than 500 but fairly similar thereafter.

5. Simulation. We conducted a simulation study to better understand how the TDMI approach performs when the imputation model is misspecified. Simulated data were based on a simplified version of the VCCC example. In this section we briefly highlight important features of the data generation and analyses needed to interpret the results. Following the recommendation of Burton et al. (2006), complete details are provided in a simulation protocol (Web Appendix C) and all R code is available at http://biostat.mc.vanderbilt.edu/ArchivedAnalyses.

The simulated cohort included 4000 subjects. Two datasets were generated to correspond to the unvalidated and validated datasets. Each subject was assigned two continuous variables, X_1 and X_2, which, for simplicity, were time-invariant and error-free. Indicator variables corresponding to whether the subject was active ($\mathcal{C}_m$), initiated at least one different ART drug ($\mathcal{A}_m$) and had an ADE ($\mathcal{D}_m$) were generated at each month $m = 0, \dots, 99$. Error-prone values of these variables, $\mathcal{A}_m^*$, $\mathcal{D}_m^*$, and $\mathcal{C}_m^*$, were also generated such that the true and error-prone variables were highly dependent. In particular, $(\mathcal{A}_m, \mathcal{D}_m, \mathcal{C}_m)$ were generated from models that induced correlation between the true values, $\mathcal{A}_m^*$, $\mathcal{D}_m^*$, $\mathcal{C}_m^*$, X_1, X_2, and time, m.

The parameters $(\sigma, \beta_2, \gamma_2)$ represent, respectively, the covariance between X_1 and X_2; the log-odds of $\mathcal{A}_m$ for X_2 after conditioning on m, X_1, $\mathcal{A}_m^*$ and $\mathcal{D}_m^*$; and the log-odds of $\mathcal{D}_m$ for X_2 after conditioning on m, X_1, $\mathcal{D}_m^*$ and $\mathcal{A}_m$. A total of 12 scenarios were constructed by varying $(\sigma, \beta_2, \gamma_2)$.

Analysis datasets for the validated, (W, X_1, X_2, Y, D) and unvalidated $(W^*, X_1, X_2, Y^*, D^*)$ data were subsequently derived by undiscretizing the data. A subset of 1000 subjects were randomly selected to represent an audited cohort $(V = 1)$. For the remaining 3000 records with $V = 0$, the validated variables $(\mathcal{A}_m, \mathcal{D}_m, \mathcal{C}_m)$ and their derived variables (W, Y, D) were masked.

The parameters of interest were $P(T_E - T_0 \leq 60)$ estimated using Kaplan–Meier techniques and β from the Cox proportional hazards model, $\lambda(m \mid X_1) = \lambda_0(m) \exp(\beta X_1)$. The TDMI procedure was implemented to the partially validated data with $B = 20$ imputation replications to estimate the parameters of interest. Two candidate sets of imputation models were considered for the TDMI procedure: (i) perfectly specified models for $\overline{\mathcal{A}}$ and $\overline{\mathcal{D}}$ that included X_1 and X_2 and (ii) misspecified models that did not include X_2. Note that imputation models for $\overline{\mathcal{C}}$ were always perfectly specified.

For each scenario, estimates and corresponding 95% confidence intervals for both the perfectly specified and misspecified TDMI implementations were generated for 1000 independent replications. We assessed the relative bias, MSE and the coverage for 95% confidence intervals. The true value for the two parameters of interest were approximated with empirical estimates from a sample size of 500,000 using only validated records. The Monte Carlo simulation error of the coverage was estimated using a bootstrap method (Koehler, Brown and Haneuse (2009)).

Table 2 shows the relative bias, MSE and coverage of the log HR using the TDMI approach under the different simulation settings. When the imputation model was correctly specified, the estimated log HR was approximately unbiased and 95% confidence intervals had coverage at or just below the nominal level (93%–95%). When the imputation model was incorrectly specified, relative bias increased and coverage decreased as the relative strength of association for the omitted covariate increased. For example, with $(\beta_2, \gamma_2) = (0, 0)$, models not including X_2 were correctly specified, so bias was low and coverage was near the nominal level. In contrast, with $(\beta_2, \gamma_2) = (1, 2)$, failure to include X_2 in imputation models led to substantial bias and very poor coverage.

TABLE 2

Summary of simulation results for time-discretized modeling and imputation (TDMI) log hazard ratio estimates from Cox regression with different levels of misspecification in the imputation model[*]

Fixed values		Imputation Model					
		Correctly specified			Misspecified		
(β_2, γ_2)	σ	Bias (%)	MSE	Coverage	Bias (%)	MSE	Coverage
(1, 2)	−0.25	−0.1%	0.0025	0.94	29.7%	0.1227	0.09
(1, 2)	0	−1.0%	0.0020	0.94	36.6%	0.1189	0.04
(1, 2)	0.25	−0.3%	0.0018	0.95	46.6%	0.1203	0.03
(0.5, 1)	−0.25	0.7%	0.0042	0.94	17.9%	0.0955	0.13
(0.5, 1)	0	0.1%	0.0038	0.94	18.3%	0.0812	0.20
(0.5, 1)	0.25	0.6%	0.0039	0.95	18.4%	0.0664	0.22
(0.25, 0.5)	−0.25	0.8%	0.0054	0.93	6.4%	0.0213	0.64
(0.25, 0.5)	0	0.4%	0.0047	0.95	6.0%	0.0179	0.70
(0.25, 0.5)	0.25	−0.4%	0.0048	0.94	5.0%	0.0136	0.79
(0, 0)	−0.25	−0.5%	0.0052	0.93	−0.4%	0.0050	0.93
(0, 0)	0	0.3%	0.0044	0.95	0.4%	0.0046	0.94
(0, 0)	0.25	−0.2%	0.0048	0.94	−0.1%	0.0048	0.94

[*] Population parameters were estimated with empirical estimates from a sample size of 500,000 using only validated records.

TABLE 3

*Summary of simulation results for time-discretized modeling and imputation (TDMI) parameter estimates from Kaplan–Meier estimation for $P(T_E - T_0 \leq 60)$ with different levels of misspecification in the imputation model**

Fixed values		Imputation Model					
		Correctly specified			Misspecified		
(β_2, γ_2)	σ	Bias (%)	MSE	Coverage	Bias (%)	MSE	Coverage
$(1, 2)$	−0.25	−0.0%	0.0001	0.95	−14.3%	0.0120	0.00
$(1, 2)$	0	−0.0%	0.0001	0.95	−13.7%	0.0116	0.00
$(1, 2)$	0.25	0.0%	0.0001	0.94	−10.9%	0.0079	0.00
$(0.5, 1)$	−0.25	0.1%	0.0001	0.92	−2.3%	0.0005	0.60
$(0.5, 1)$	0	−0.0%	0.0001	0.94	−2.3%	0.0005	0.61
$(0.5, 1)$	0.25	0.1%	0.0001	0.94	−1.5%	0.0003	0.76
$(0.25, 0.5)$	−0.25	−0.0%	0.0001	0.94	−0.5%	0.0001	0.91
$(0.25, 0.5)$	0	0.0%	0.0001	0.93	−0.4%	0.0001	0.93
$(0.25, 0.5)$	0.25	−0.1%	0.0001	0.93	−0.5%	0.0001	0.92
$(0, 0)$	−0.25	−0.0%	0.0001	0.93	−0.0%	0.0001	0.93
$(0, 0)$	0	0.2%	0.0001	0.93	0.2%	0.0001	0.94
$(0, 0)$	0.25	−0.0%	0.0001	0.94	−0.0%	0.0001	0.95

*Population parameters were estimated with empirical estimates from a sample size of 500,000 using only validated records.

Simulation results for Kaplan–Meier TDMI estimates of the incidence of an event by $m = 60$ are shown in Table 3. Conclusions were largely the same as for the log HR. When the imputation model was correctly specified, estimates of the 60-month incidence were approximately unbiased with coverage probabilities at or just below the nominal level (92%–95%). When the imputation model was incorrectly specified, absolute bias increased and coverage decreased as the relative strength of association for the omitted covariate increased. For both parameters, the Monte Carlo simulation error was 1.6% or lower for coverage estimates, 0.4% or lower for relative bias estimates and 0.0024 or lower for MSE estimates for all scenarios.

6. Discussion.

Using EHR data from an HIV cohort, we have illustrated the bias that can arise by ignoring data errors, and we have proposed a missing data analysis solution that incorporates validation data to address multidimensional errors in time-to-event analyses. To our knowledge, this is the first study to simultaneously address errors in both predictors and outcomes in a time-to-event analysis. We were also able to address errors in study eligibility. The TDMI approach did not outperform the complete-case approach under all scenarios, but we are encouraged that it led to improved estimation under most conditions, particularly when estimating the log HR.

The TDMI procedure is subject to various assumptions generally similar to those required for multiple imputation in standard missing data settings (Van Buuren (2018)). The key missing at random assumption (in fact, the stronger missing completely at random assumption) was satisfied in our example as the audited sample was a simple random sample. This assumption can also be satisfied in more complicated settings, where subjects are sampled based on the observed unvalidated data with known probabilities, but will likely be violated if the validation sample is one of convenience.

Another basic assumption underlying the TDMI approach is that the imputation model is properly specified. This is difficult in practice. Despite our best efforts—the incorporation of over 30 covariates, both time-fixed and time-varying exposures—estimates for our approach were still biased, especially at smaller validation sample sizes. Results from both our reduced model TDMI and the simulation study highlight potential challenges with model misspecification. Model overfitting can be a problem at smaller validation sample sizes, as seen by our reduced model TDMI out-performing the original model TDMI at small audit sample sizes when estimating the log HR. But the reduced model was not sufficiently rich to obtain good estimates for the incidence of ADE at modest audit sizes. We are currently studying generalized raking methods to combine potentially biased but efficient estimators like the TDMI estimator with unbiased but less efficient complete-case estimators (Lumley, Shaw and Dai (2011)).

We considered alternative modeling approaches (e.g., classification and regression trees, random forests, support vector machines and linear discriminant analysis) but, ultimately, fit logistic regression models. Given the improbability of knowing, a priori, which model will perform best for a certain setting, it might be worthwhile to add a preliminary step that selects the most appropriate model through cross-validation or some other model-selection procedure in the audit subsample.

We estimated standard errors using the Robins and Wang (2000) imputation variance estimator instead of the more popular (and easier to implement) approach proposed by Rubin (Little and Rubin (2014)) because of incompatibility between imputation and analysis models. There were two sources of incompatibility in our setting. First, the unit of observation was different between the imputation model (subject-month observations) and the analysis model (subject-level observations). Second, our study had exclusion criteria that removed observations from the analysis model that contributed information to the imputation model. Standard errors calculated using Rubin's rule in our setting led to inflated standard errors and conservative confidence intervals (e.g., coverage of 98–99% in simulations, data not shown).

There are potential issues from coarsening the data into time intervals when fitting the imputation models. One potential issue with discretization is a loss of information. However, losses of information due to discretization to the level of months will be minimal in clinical settings where visits typically occur no more than monthly. As an additional sensitivity analysis, we coarsened the data from the validated records into monthly intervals and reestimated the incidence of ADE at five years for a validation subsample of size 1000. Using the estimate from the nondiscretized fully validated data for all patients as the gold standard, the MSE of the discretized version of the complete-case estimator (2.4×10^{-4}) was nearly identical to the nondiscretized complete-case estimator (2.4×10^{-4}), suggesting little loss of information. Other potential issues may arise when we convert our imputed discrete-time data (in months) back to the original measurement scale (days). This conversion is not completely necessary; estimates were similar when we fit discrete-time pooled logistic regression models compared to estimates when we switched back to time measured in days and fit Cox regression models. Note that, even after the data are converted back to the original measurement, the imputed failure times may still be grouped. When we use Cox regression, we are thus fitting grouped proportional hazards models (Tutz and Schmid (2016)). To account for ties, we used Efron's partial likelihood to approximate the discrete hazard model partial likelihood.

Although our analyses focused on Kaplan–Meier and Cox regression estimates, a strength of our multiple imputation approach is that we could have also performed other estimation procedures. For example, we estimated the cumulative incidence treating death as a competing risk and obtained similar results. As another example, instead of excluding those with ART prior to enrollment, we could have applied a TDMI-like approach that incorporated these patients in an analysis using an estimation procedure that accounted for left-truncation. The flexibility of the TDMI approach to handle other estimation procedures is important because analyses of EHR data typically require addressing multiple problems simultaneously (e.g., confounding, missing data and informative censoring). Methods for dealing with these other sources of bias could potentially be applied to the multiply imputed dataset without substantial modification. Of course, the performance of our approach may vary across analysis methods, as we saw in this study.

In our example dataset there were only errors in two variables, date of ART and date of ADE, so we did not illustrate how our TDMI approach performs when there are errors in the predictor variable beyond those induced by errors in the date of ART initiation. Although we believe general performance would be similar to that seen in the analyses presented here, computation would be more complicated and require at least one additional model.

Future research will consider improving the efficiency of these methods by applying principles of two-phase sampling, such as oversampling exposures or events that are rare or considered a priori to be more error-prone.

Acknowledgments. We thank our coinvestigators at the VCCC for the use of their data. This work was conducted, in part, using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University, Nashville, TN. This research was supported, in part, by the National Institutes of Health (P30 AI110527, R01 AI093234, R01 AI131771, U01 AI069923 and U01 AI069918) and the Patient Centered Outcome Research Institute (R-1609-36207).

SUPPLEMENTARY MATERIAL

Web-based supplementary materials (DOI: 10.1214/20-AOAS1343SUPP; .pdf). We provide more information regarding the VCCC cohort, additional details of model specification, and additional figures as online supplementary materials (Giganti et al. (2020)).

REFERENCES

BURTON, A., ALTMAN, D. G., ROYSTON, P. and HOLDER, R. L. (2006). The design of simulation studies in medical statistics. *Stat. Med.* **25** 4279–4292. MR2307592 https://doi.org/10.1002/sim.2673

CARPENTER, J. R., KENWARD, M. G. and VANSTEELANDT, S. (2006). A comparison of multiple imputation and doubly robust estimation for analyses with missing data. *J. Roy. Statist. Soc. Ser. A* **169** 571–584. MR2236921 https://doi.org/10.1111/j.1467-985X.2006.00407.x

CHAN, K. S., FOWLES, J. B. and WEINER, J. P. (2010). Electronic health records and reliability and validity of quality measures: A review of the literature. *Med. Care Res. Rev.* **67** 742–752.

COLE, S. R., CHU, H. and GREENLAND, S. (2006). Multiple-imputation for measurement-error correction. *Int. J. Epidemiol.* **35** 1074–1081.

COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

D'AGOSTINO, R. B., LEE, M.-L., BELANGER, A. J., CUPPLES, L. A., ANDERSON, K. and KANNEL, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat. Med.* **9** 1501–1515.

DUDA, S. N., SHEPHERD, B. E., GADD, C. S., MASYS, D. R. and MCGOWAN, C. C. (2012). Measuring the quality of observational study data in an international HIV research network. *PLoS ONE* **7** e33908.

EDWARDS, J. K., COLE, S. R., TROESTER, M. A. and RICHARDSON, D. B. (2013). Accounting for misclassified outcomes in binary regression models using multiple imputation with internal validation data. *Am. J. Epidemiol.* **177** 904–912.

EFRON, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.* **83** 414–425. MR0971367

FLOYD, J. S., HECKBERT, S. R., WEISS, N. S., CARRELL, D. S. and PSATY, B. M. (2012). Use of administrative data to estimate the incidence of statin-related rhabdomyolysis. *J. Am. Med. Assoc.* **307** 1580–1582.

GIGANTI, M. J., SHAW, P. A., CHEN, G., BEBAWY, S. S., TURNER, M. M., STERLING, T. R. and SHEPHERD, B. E. (2020). Supplement to "Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples and multiple imputation." https://doi.org/10.1214/20-AOAS1343SUPP

HERNÁN, M. A., BRUMBACK, B. and ROBINS, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J. Amer. Statist. Assoc.* **96** 440–448. MR1939347 https://doi.org/10.1198/016214501753168154

HUANG, Y. and WANG, C. Y. (2000). Cox regression with accurate covariates unascertainable: A nonparametric-correction approach. *J. Amer. Statist. Assoc.* **95** 1209–1219. MR1804244 https://doi.org/10.2307/2669761

HUNSBERGER, S., ALBERT, P. S. and DODD, L. (2010). Analysis of progression-free survival data using a discrete time survival model that incorporates measurements with and without diagnostic error. *Clin. Trials* **7** 634–642.

KOEHLER, E., BROWN, E. and HANEUSE, S. J.-P. A. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *Amer. Statist.* **63** 155–162. MR2750076 https://doi.org/10.1198/tast.2009.0030

KORN, E. L., DODD, L. E. and FREIDLIN, B. (2010). Measurement error in the timing of events: Effect on survival analyses in randomized clinical trials. *Clin. Trials* **7** 626–633.

LI, Y. and LIN, X. (2003). Functional inference in frailty measurement error models for clustered survival data using the SIMEX approach. *J. Amer. Statist. Assoc.* **98** 191–203. MR1965685 https://doi.org/10.1198/016214503388619210

LITTLE, R. J. and RUBIN, D. B. (2014). *Statistical Analysis with Missing Data* **333**. John Wiley & Sons.

LUMLEY, T., SHAW, P. A. and DAI, J. Y. (2011). Connections between survey calibration estimators and semi-parametric models for incomplete data. *Int. Stat. Rev.* **79** 200–220. https://doi.org/10.1111/j.1751-5823.2011.00138.x

MAGARET, A. S. (2008). Incorporating validation subsets into discrete proportional hazards models for mismeasured outcomes. *Stat. Med.* **27** 5456–5470. MR2542363 https://doi.org/10.1002/sim.3365

MCCLINTOCK, B. T., JOHNSON, D. S., HOOTEN, M. B., HOEF, J. M. V. and MORALES, J. M. (2014). When to be discrete: The importance of time formulation in understanding animal movement. *Mov. Ecol.* **2** 21. https://doi.org/10.1186/s40462-014-0021-6

MCISAAC, M. and COOK, R. J. (2017). Statistical methods for incomplete data: Some results on model misspecification. *Stat. Methods Med. Res.* **26** 248–267. MR3592724 https://doi.org/10.1177/0962280214544251

NAKAMURA, T. (1990). Corrected score function for errors-in-variables models: Methodology and application to generalized linear models. *Biometrika* **77** 127–137. MR1049414 https://doi.org/10.1093/biomet/77.1.127

OH, E. J., SHEPHERD, B. E., LUMLEY, T. and SHAW, P. A. (2018). Considerations for analysis of time-to-event outcomes measured with error: Bias and correction with SIMEX. *Stat. Med.* **37** 1276–1289. MR3777974 https://doi.org/10.1002/sim.7554

PRENTICE, R. L. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika* **69** 331–342. MR0671971 https://doi.org/10.1093/biomet/69.2.331

RICHARDSON, B. A. and HUGHES, J. P. (2000). Product limit estimation for infectious disease data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1** 341–354.

ROBINS, J. M. and WANG, N. (2000). Inference for imputation estimators. *Biometrika* **87** 113–124. MR1766832 https://doi.org/10.1093/biomet/87.1.113

SHAW, P. A. and PRENTICE, R. L. (2012). Hazard ratio estimation for biomarker-calibrated dietary exposures. *Biometrics* **68** 397–407. MR2959606 https://doi.org/10.1111/j.1541-0420.2011.01690.x

SHEPHERD, B. E., SHAW, P. A. and DODD, L. E. (2012). Using audit information to adjust parameter estimates for data errors in clinical trials. *Clin. Trials* **9** 721–729.

SHEPHERD, B. E. and YU, C. (2011). Accounting for data errors discovered from an audit in multiple linear regression. *Biometrics* **67** 1083–1091. MR2829243 https://doi.org/10.1111/j.1541-0420.2010.01543.x

SKINNER, C. J. and HUMPHREYS, K. (1999). Weibull regression for lifetimes measured with error. *Lifetime Data Anal.* **5** 23–37. MR1750340 https://doi.org/10.1023/A:1009674915476

TSIATIS, A. A. and DAVIDIAN, M. (2001). A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika* **88** 447–458. MR1844844 https://doi.org/10.1093/biomet/88.2.447

TURCHIN, P. (1998). *Quantitative Analysis of Movement*: *Measuring and Modeling Population Redistribution in Animals and Plants* **1**. Sinauer Associates Sunderland.

TUTZ, G. and SCHMID, M. (2016). *Modeling Discrete Time-to-Event Data*. *Springer Series in Statistics*. Springer, Cham. MR3497009 https://doi.org/10.1007/978-3-319-28158-2

VAN BUUREN, S. (2018). *Flexible Imputation of Missing Data*. CRC Press/CRC.

WULFSOHN, M. S. and TSIATIS, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53** 330–339. MR1450186 https://doi.org/10.2307/2533118