Bryan E. Shepherd* and Pamela A. Shaw

Errors in multiple variables in human immunodeficiency virus (HIV) cohort and electronic health record data: statistical challenges and opportunities

https://doi.org/10.1515/scid-2019-0015
Received October 15, 2019; accepted August 21, 2020; published online October 7, 2020

Abstract

Objectives: Observational data derived from patient electronic health records (EHR) data are increasingly used for human immunodeficiency virus/acquired immunodeficiency syndrome (HIV/AIDS) research. There are challenges to using these data, in particular with regards to data quality; some are recognized, some unrecognized, and some recognized but ignored. There are great opportunities for the statistical community to improve inference by incorporating validation subsampling into analyses of EHR data.

Methods: Methods to address measurement error, misclassification, and missing data are relevant, as are sampling designs such as two-phase sampling. However, many of the existing statistical methods for measurement error, for example, only address relatively simple settings, whereas the errors seen in these datasets span multiple variables (both predictors and outcomes), are correlated, and even affect who is included in the study.

Results/Conclusion: We will discuss some preliminary methods in this area with a particular focus on time-to-event outcomes and outline areas of future research.

Keywords: electronic health records; HIV; measurement error; misclassification; two-phase sampling.

Introduction

Routinely collected data, such as data derived from electronic health records, clinical medical charts, or administrative databases, are being increasingly used in human immunodeficiency virus (HIV) research. Their use allows investigators to obtain large amounts of data at relatively low costs, and therefore study questions that would not otherwise be feasible to investigate. A few examples presented at the 2019 Conference on Retroviruses and Opportunistic Infections (CROI) include studies of the continuum of care (MacKellar et al. 2019), studies of deaths due to opioids among persons living with HIV (Bosh et al. 2019), and studies of dolutegravir use among women at conception and its potential relationship with neural tube defects in their newborn infants (Mofenson 2019). These studies relied on the use of routinely collected data. Funding agencies have realized the importance of these data sources and have contributed significant resources to establishing networks such as the International Databases to evaluate acquired immunodeficiency syndrome (AIDS) consortium (www.iedea.org).

The use of routinely collected data for HIV research also comes with challenges (Weng et al. 2012). One of the greatest challenges is that these data are prone to errors that in some cases could substantially

***Corresponding author: Bryan E. Shepherd**, Biostatistics, Vanderbilt University, 2525 West End, Suite 11000, 37203 Nashville, Tennessee, USA, E-mail: bryan.shepherd@vanderbilt.edu. https://orcid.org/0000-0002-3758-5992
Pamela A. Shaw, Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

change study conclusions, and clinical care, if the data are taken at face value. A seminal example in the HIV literature comes from using data collected as part of routine care from HIV clinics to study mortality in the presence of high rates of loss to follow-up. Geng et al. (2008) traced a random sample of patients in Uganda who were lost to follow-up and found that a large portion of them had died; naive estimates using the readily available clinical data and standard Kaplan-Meier estimators of survival, which treat these lost individuals as independently censored, severely underestimated the probability of death after antiretroviral therapy (ART) initiation; three-year estimated mortality was approximately 2% using the clinic data vs. 12% when corrected based on the tracing data. Data errors, whether in the outcomes or the exposures, have the potential to bias study results when not addressed in the analytical methods used for analysis.

Several research groups have investigated the quality of routinely collected HIV data by performing data audits, also referred to as source data verification or data validation. The Caribbean, Central and South America network for HIV epidemiology (CCASAnet), for example, has sent teams of auditors to each of their study sites to compare data sent to the data coordinating center with data in the clinical medical records (Duda et al. 2012). Other groups have engaged in similar types of data quality initiatives (e.g., Kiragga et al. 2011); however, data validation can be expensive and time-consuming, so the practice is unfortunately not widespread (Ledergerber 2012). When using the electronic health record (EHR) as a data source, investigators often validate a subsample of records for a few key variables by going in detail through the EHR, including text fields; this information is often then used to confirm that diagnoses are accurate or to develop and validate computational phenotyping algorithms used to classify diagnoses to large numbers of patients throughout the EHR. For example, NA-ACCORD (IeDEA's North American region) validated end stage liver disease (ESLD) and end stage renal disease (ESRD) for a randomly selected 6% of their patients through comprehensive medical record reviews; they also examined algorithms to screen for events with readily available data to avoid infeasible chart review for the entire cohort (Kitahata et al. 2015). They found that their algorithms for ESRD had positive and negative predictive values of 82 and 100%, respectively, and 27 and 100% for ESLD.

There are important opportunities for statisticians to make a positive impact on studies using error-prone observational data. Two key areas of impact are 1) incorporating validation data into analyses, and 2) designing efficient validation studies. Once an investigator has performed an audit, they are often left wondering what to do with the information. Often the audit results only make it into discussions of study strengths and limitations. However, there is a rich statistical literature on how to incorporate data from a validation study into a larger study; the measurement error and missing data literatures are relevant (e.g., Carroll et al. 2006; Little and Rubin 2002). Designing an audit can be thought of as designing a two-phase sample; there is also a rich statistical literature on efficient two-phase sampling, which could be applied to validation sampling of routinely collected data to improve study estimates (e.g., Breslow and Chatterjee 1999; Gilbert, Yu, and Rotnitzky 2014; MacNamee 2005; Reilly 1996; Xu, Hui, and Grannis 2014; Zhao and Lipsitz 1992).

In practice, however, there are complications with the errors of routinely collected data that make it difficult to directly apply many existing methods. Some of this is because there are often errors across multiple variables – particularly when these variables are derived from the same error-prone underlying variables – and the existence and magnitude of these errors are correlated. We will provide a detailed illustration in Section 2. Similarly, complications specific to the use of routinely collected data also make designing efficient validation samples in this setting somewhat different from what has been studied elsewhere.

The purpose of this article is to highlight opportunities and challenges for dealing with error-prone routinely collected data in HIV research. We will describe some preliminary work in this area, with a particular focus on methods for time-to-event outcomes, and outline some impactful areas for statistical and implementation science research. In Section 2 we provide a real data illustration, in Section 3 we discuss estimation, in Section 4 we discuss validation designs, and in Section 5 we provide a few general conclusions. This manuscript is based on a presentation given at the 2019 Workshop on Statistical Challenges and Opportunities

in HIV/AIDS Research in the Era of Getting-to-Zero HIV Infections organized by the United States National Institute of Allergy and Infectious Diseases.

Illustration

CCASAnet is a research network that includes clinical data from HIV positive individuals living with HIV in Latin America, with sites in Argentina, Brazil, Chile, Haiti, Honduras, Mexico, and Peru. The goal of CCASAnet is to use this shared repository of HIV data to answer questions about the characteristics of the regional HIV epidemic. CCASAnet data are assembled at each site and then sent to a data coordinating center for merging and analysis. CCASAnet data protocols closely follow those of other HIV cohorts (http://iedea.github.io, https://www.hicdep.org). In short, CCASAnet data include a table of demographics and time-invariant data (e.g., date of HIV diagnosis), tables with laboratory information (e.g., repeated CD4 count and HIV-1 RNA measurements over time), a table with clinic visit information, a table with antiretroviral therapy data (e.g., regimen and dates of starting/stopping the regimen), a table of clinical events (e.g., onset of tuberculosis, Kaposi sarcoma, and dates), a table with vital status and follow-up information, and other tables corresponding to specialized data collection stemming from specific projects (e.g., substance abuse during a particular period of time, pregnancy outcomes). Data from the same patient across tables are linked with a common personal identification number.

As is common in studies using routinely collected observational data, analysis variables for CCASAnet studies are often derived from multiple variables across multiple tables. As an example that we will refer to throughout this manuscript, suppose we are interested in studying the time from ART initiation to an AIDS-defining event (ADE) and estimating its association with other variables, including CD4 count, at the time of ART initiation. The time-to-event variable is derived from the date of ART initiation in the ART table, the date of first ADE based on the clinical events table, and the last visit date or death date if no event occurred from the vital status table. CD4 count at ART initiation is derived from tables of laboratory values as the CD4 measurement taken the closest to the date of ART initiation, within a certain window. Patient eligibility for the study is also derived from the database. For example, one might only include patients who start ART prior to an ADE.

CCASAnet has performed multiple waves of data quality audits. An audit team from the data coordinating center, usually consisting of a clinician and an informaticist, typically spends about two days at each site comparing data sent to the coordinating center with data in the clinical medical records for a subset of records randomly selected by the data coordinating center. Audit procedures and findings have been published previously (Duda et al. 2012).

In 2013–2014, CCASAnet conducted on-site audits at seven HIV research sites. A total of 250 records were audited, including 14,995 values across 23 variables. The data audits uncovered some errors, with an overall discrepancy percentage of approximately 17%. The discrepancies included data mismatches (i.e., value in the database does not match the value in the charts) (6%), observations found in source document that were not in the database (7%), or values that could not be verified in the source document (4%). Some variables were quite reliable: e.g., sex at birth, which was found to be incorrect in only three (1%) records. Other variables, however, were much more prone to errors. For example, CD4 cell count was incorrect for 10.5% of entries, ART start data was incorrect for 20.5% of entries, and date of ADE was incorrect for 50% of entries (4% date mismatches, 43% ADE discovered in the clinical charts that was not in the database, and 3% ADE reported in database but not found in clinical charts).

Throughout, we assume that the audit values are correct, but this is not always the case. CCASAnet has also done studies that compare audit results between outside-auditors and self-audits, where investigators at the site audit their own data; we have found high, but not perfect, concordance (94%) between auditors (Lotspeich et al. 2019), and concordance does not guarantee correctness. And, of course, some events do not make it into the medical records. While data discrepancies may not be able to be completely resolved, the information these types of concordance studies provide can be used to inform sensitivity analyses that can

incorporate varying assumptions about the accuracy of the observed data to evaluate the robustness of study results across a range of plausible scenarios.

Data errors are not just an issue for international cohorts. For example, we have seen similar non-trivial error rates when validating electronic health record data from persons living with HIV attending the Vanderbilt Comprehensive Care Clinic (Giganti et al. 2020).

Because analysis datasets use derived variables, errors in the original variables often propagate across multiple variables, meaning that errors will be dependent across derived variables. Using the example above, an error in the date of ART initiation will lead to an error in the censored-failure time (Y^*) and likely an error in CD4 count at ART initiation (X^*). Errors in the date of ADE will lead to additional errors in the censored-failure time (Y^*) and the ADE indicator (D^*). Finally, errors in date of ART initiation and/or date of ADE may lead to improper inclusion/exclusion. We denote the true and error prone indicators of inclusion criteria being met as V and V^*, respectively; true and error prone variables are similarly denoted for other variables and Z denotes error-free covariates. These interrelated errors likely produce non-additive errors with a complicated correlation structure, and methods to address these errors will likely need to be flexible and not overly reliant on distributional assumptions about the nature of the error.

Analysis methods

The measurement error literature provides a foundation of methods to incorporate validation data into analyses, but much work remains to be able to address the full complexity of our setting with dependent errors across multiple variables. For example, for time-to-event studies with covariate measurement error, where (X^*, Y, D) are available on all subjects and X is available only on a validation sample, the literature is rich with a variety of methods for addressing this measurement error (e.g., Carroll et al. 2006; Cole, Chu, and Greenland 2006; Prentice 1982; Xie, Wang, and Prentice 2001). There is also a literature that considers errors in the event indicator (X, Y, D^*) (e.g., Balasubramanian and Lagakos 2003; Gravel et al. 2018; Gu, Ma, and Balasubramanian 2015; Meier, Richardson, and Hughes 2003), and there are a few papers on errors only in failure times (X, Y^*, D) (Dodd et al. 2011; Hong et al. 2012; Korn, Dodd, and Freidlin 2010; Oh et al. 2018; Skinner and Humphreys 1999). Gravel et al. (2018) considered error in either the timing or the classification in a competing risk setting (X, Y^*, D^*). But except for recent work described below, we are not familiar with methods to simultaneously address covariate and failure time errors (X^*, Y^*, D). Our situation is even more complicated with errors in event indicators and whether someone qualifies for inclusion in the study (X^*, Y^*, D^*, V^*). In addition, many of the existing methods make assumptions on the structure of the error (e.g., additive error), which are often not realistic in our setting. Therefore, it is doubtful that existing methods can be directly applied with the validation data to correct for data errors in the CCASAnet analysis of time from ART initiation until ADE.

Here we provide a brief summary of some of the statistical methods used to address measurement error, preliminary work expanding these methods to address correlated errors across multiple variables, and our thoughts on the methods' potential to address more complicated settings seen in practice with errors in routinely collected data. Table 1 provides a quick summary for reference and comparison across the methods.

Moment-based estimators

Moment-based estimators are simple approaches for addressing covariate measurement error. With continuous Y and X and classical measurement error (i.e., $X^* = X + U$ where $E(U) = 0$, $Var(U) = \sigma_u^2$ is constant, and U independent of other variables), if the model $E(Y|X^*) = \alpha_0 + \alpha_1 X^*$ is fit, then it is well known that $\alpha_1 = \beta_1 \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$, where σ_x^2 is the variance of X and β_1 is the true association between Y and X that one would obtain from the model $E(Y|X) = \beta_0 + \beta_1 X$. A simple moment-based correction is then to estimate σ_x^2 and σ_u^2 from the validation data, frequently selected as a simple random sample (SRS), and then to divide the naive estimate of α_1 by the estimated attenuation factor $\sigma_x^2/(\sigma_x^2 + \sigma_u^2)$. Moment estimators can also be adjusted to handle stratified SRS. When Y is also measured

Table 1: Informal summary of analysis methods.

Analysis method	Key assumptions for consistency[a]	Validation sampling designs[b]	Efficiency[d]	Adaptability/ Flexibility[e]	Ease of implementation with standard software
Moment-based estimators	Outcome model correctly specified, non-differential additive error, error model correctly specified, generally consistent only for linear model	Simple or stratified random sampling[c]	Moderate	Low	Simple
Regression calibration	Outcome model correctly specified, non-differential additive error, calibration model correctly specified, generally consistent only for linear model	Simple or stratified random sampling[c]	Moderate	Low	Simple
SIMEX	Outcome model correctly specified, non-differential additive error or binary misclassification, extrapolation function properly specified	Simple or stratified random sampling[c]	Varies by setting	Low	Simple
Multiple imputation	Outcome model correctly specified, imputation model correctly specified	Probability-based sampling	High	Moderate	Varies by model complexity
Likelihood-based estimators	Outcome model correctly specified, error model correctly specified	Probability-based sampling	High	Moderate	Varies by model complexity
Bayesian estimators	Outcome model correctly specified, error model correctly specified	Probability-based sampling	High	Moderate	Varies by model complexity
Design-based estimators		Probability-based sampling, non-zero sampling probability for all records	Low to moderate	High	Simple

[a]Key assumptions for standard approaches to be consistent, recognizing that modifications can be made to relax some assumptions. By consistency, we mean consistent for the parameter we would estimate if we had correct data on all subjects. We recognize that design-based estimators make some assumptions (e.g., a regular estimator in the validation sample), but these assumptions are generally mild and made by the other estimators so we did not include them in this table.
[b]The types of validation sampling for which the method is generally applicable.
[c]Can be adapted to probability-based sampling using inverse-probability weighting when there are non-zero sampling probabilities for all records.
[d]Efficiency in terms of variance when assumptions are met and sample size is large.
[e]Ability to handle a large number of error-prone variables and a wide range of error structures.

with error and this error is correlated with the error in X, it can be shown that the naive estimate of γ_1 based on the regression of $E(Y^*|X^*)= \gamma_0 + \gamma_1 X^*$ has expectation $\beta_1\sigma_x^2/(\sigma_x^2 + \sigma_u^2) + Cov(Y - Y^\star, U)/(\sigma_x^2 + \sigma_u^2)$. Therefore, an unbiased estimate of β_1 can be obtained by fitting the naive model to the full dataset, estimating $\sigma_x^2, \sigma_u^2, \frac{1}{2}$ and $Cov(Y - Y^*, U)$ from the validation data, and then plugging these estimates into the equation of the expectation of γ_1 and solving for β_1. This approach is easily extended to include other covariates (both error-prone and error-free) (Shepherd and Yu 2011). To our knowledge, moment-based estimators have not been extended beyond this relatively simple setting, but their utility/tractability for more complicated settings (e.g., non-additive errors in (X^*, Y^*, D^*) with time-to-event data) is uncertain.

Regression calibration

Regression calibration (RC) is a popular method to address covariate measurement error due to its simplicity and wide applicability (Carroll et al. 2006). The idea is to singly impute the unobserved X with an estimate of

$E(X|X^*, Z)$, where Z may be other precisely observed covariates, and perform the desired regression with the imputed exposure. For linear models that are correctly specified, RC results in asymptotically unbiased estimation. For non-linear models, RC has some asymptotic bias; however, in many settings, the bias is modest (Prentice 1982; Shaw and Prentice 2012). RC can be applied with a validation sample, typically a SRS or a stratified SRS, to estimate $E(X|X^*, Z)$. We extended RC to address correlated errors in covariates and a continuous outcome, applying calibration estimators fit on a reliability, validation or calibration subset (Shaw, He, and Shepherd 2018). We then applied this approach to correlated additive error in both the censored failure time Y^* and exposure X^*. In this method, the usual estimate $\widehat{X} = \widehat{E}(X|X^*, Z)$ of the unobserved X is obtained, along with an estimate of the true censored failure time, $\widehat{Y} = \widehat{E}(Y|X^*, Z)$, which is used to create calibrated risk sets in the Cox score equation (Oh et al. 2019). We further refined this method through successive calibrations across the risk sets, a method previously seen to reduce bias of the regression calibration estimator for covariate error (Xie, Wang, and Prentice 2001). We found that our RC estimators improved upon the naive estimator for modest log-hazard ratios, but that for large log-hazard ratios the method had appreciable bias. Bias can also be seen for RC when the distribution of the errors is not well-approximated by a normal distribution (Shaw and Prentice 2012). Given these biases observed with correlated errors in X^* and Y^*, and more generally with RC, extensions of RC to also address errors in D^* and V^* may not be particularly fruitful.

SIMEX

Simulation and extrapolation (SIMEX) is an approximate method where one incorporates additional error into an error-prone variable through simulation, estimates the relationship between the error variance and coefficient estimates, and then extrapolates this relationship to obtain estimates when the error variance is zero (Cook and Stefanski 1994). The variance of the error in the observed data is either assumed known or estimated using a validation sample based on SRS. SIMEX has generally been applied to address covariate measurement error (e.g., Alexeeff, Carroll, and Coull 2016; He, Yi, and Xiong 2007; Parveen, Moodie, and Brenner 2017), although in recent work we used it to address measurement error in the observed censored failure time (Oh et al. 2018). To apply SIMEX in our setting with errors across multiple variables, one would simulate multidimensional, correlated error and assess the relationship between the error variance-covariance matrix and coefficient estimates. For example, with additive errors in continuous Y and X, one would simulate data varying the variance of the errors in Y and X and their covariance. One would estimate the relationship between these various errors and the coefficient of interest and then extrapolate back to the setting where there is no error. Even with this relatively simple setting with additive, correlated errors in Y and X, interesting analytical issues arise, as the extrapolation is across multiple dimensions. Holcomb (1999) describes a bivariate SIMEX method in which the relationship between two variables (X^*, Y^*) that were themselves derived from a common set of error-prone variables was investigated. However, addressing errors in (X^*, Y^*, D^*, V^*), as in our CCASAnet time-to-event study, using SIMEX seems particularly challenging because of both the dimension and mix of misclassification and measurement error.

Multiple imputation (MI)

Measurement error can be thought of as a missing data problem (Carroll et al. 2006), where we have complete data $(X^*, Y^*, D^*, V^*, X, Y, D, V, Z)$ on a subsample of validated records and incomplete data (X^*, Y^*, D^*, V^*, Z) for the remainder. Because validation subsets are selected by design, the missing data mechanism is missing at random (MAR), and standard missing data methods are applicable (Little and Rubin 2002). If the validation sample is a SRS then the data are actually missing completely at random (MCAR); but a strength of multiple imputation (MI) is that it can handle other, more efficient MAR sampling designs, which will be discussed in Section 4. MI has been used to account for misclassified binary and continuous outcomes (Edwards et al. 2013); and in time-to-event models, covariate misclassification (Cole, Chu, and Greenland 2006). In earlier work, we

applied MI to address correlated errors in X^* and Y^* (Shepherd, Shaw, and Dodd 2012). In recent work, we applied MI to address correlated errors in X^*, Y^*, D^*, and V^* (Giganti et al. 2020). In that analysis, we built imputation models based on simplified discrete-time failure models that allowed us to incorporate the error-prone values of these variables and many additional time-varying covariates to impute the correct values. Using the CCASAnet time from ART to ADE example, indicators of ART start and an ADE would be imputed (yes/no) in each month based on time-fixed and time-varying covariates and the error-prone values of these variables in the unvalidated data. A strength of MI is that it is able to address fairly complicated settings. When multiple variables are derived from underlying original variables, one can impute those original variables (e.g., month of ART initiation or ADE) and re-derive analysis variables (e.g., time from ART initiation to ADE), thus creating realistic dependencies between imputed variables (Van Buuren 2012). In addition, the unvalidated value of the variable is often a very good predictor of the correct value of that variable. There are some caveats to using MI, however. The key assumption underlying unbiased estimation for MI is that the imputation model is properly specified. In practice, it can be hard to avoid model misspecification. There are often bias-variance trade-offs to consider. In addition, the standard Rubin's rules formulas for computing the variance of MI estimators are often biased because of incompatibility (also known as uncongeniality) between the imputation and analysis models (Xie and Meng 2017). This problem can be corrected using alternative variance estimators (Robins and Wang 2000), but these can be complicated to implement (Giganti and Shepherd 2020).

Likelihood-based estimation

Another approach for addressing measurement error is to put models on the various components of the error, as well as the outcome, and employ a maximum likelihood approach to estimate the parameters of interest. One can consider the error-prone data as auxiliary variables that can improve estimation based on the likelihood of the observed error-free true data, where the error-free data is obtained at the second phase of a two-phase sampling design (e.g. validation subset). The general form of the likelihood for two-phase samples is well-known (Scott and Wild 1997). Likelihood-based estimation can handle any sampling that, conditional on the error-prone data, is independent of the correct data (i.e., the unvalidated data are MAR). Most prior likelihood-based work has not considered errors in X^* and Y^*, with the exception being a paper by Tang et al. (2015), that considered binary X^* and Y^*, both of which are misclassified with errors both differential and dependent; their approach is fully parametric. Others have proposed semiparametric solutions that can be applied to situations with errors only in X^* (Tao, Zeng, and Lin 2017). In recent work, we have extended these semiparametric methods to handle correlated errors in X^* and Y^*, for general X^* and continuous (Tao et al. 2020) or binary (Lotspeich et al. 2020) Y^*. These approaches leave unspecified the model for the distribution of the errors conditional on other variables. Hence, they are more robust than fully parametric models, but still quite efficient. We hope to extend them to time-to-event settings where there may also be misclassification of D^* to be able to handle situations such as the CCASAnet time-to-ADE analysis. Such models are feasible, but in practice it may prove difficult to propose realistic and tractable models; estimates are biased if models are not properly specified. These semiparametric techniques also have difficulties including many covariates, and may require data reduction techniques in practice. Boe, Tinker, and Shaw (2020) considered a combination of regression calibration with a likelihood approach to handle both misclassified discrete time to event outcome and exposure error. Huang et al. (2018) considered the problem of a misclassified outcome in EHR data and proposed an integrated likelihood approach that integrates out nuisance parameters, in their case the unknown sensitivity and specificity of an outcome classification algorithm. Such an approach is related to other likelihood-based approaches mentioned above except that it does not jointly model the error-prone and error-free data, but incorporates uncertainty in the sensitivity and specificity estimates through the use of a prior that is placed in the integrated likelihood. Similar approaches may be useful to explore for our setting, though computational complexity, already a challenge for the misclassified outcome, will likely also be a challenge

when there are errors in multiple variables. And choosing priors for complicated settings with errors across multiple variables is somewhat daunting.

Bayesian estimation

The Bayesian paradigm offers a conceptually simple approach to handling measurement error. One can factor the joint density of the data (X^*, Y^*, D^*, X, Y, D, Z) in a way that allows one to specify separate models for exposure, outcome and measurement error. With these models and specified prior distributions for the unknown parameters, one derives the posterior distribution for the parameters of interest from the observed data. Whether or not the model has exposure measurement error only, outcome error only, or both, does not change the basic approach. Gustafson (2003) provides an overview for Bayesian methods that handle exposure measurement error. Bartlett and Keogh (2018) outline several advantages of the Bayesian approach over standard measurement error correction approaches, including regression calibration, multiple imputation and likelihood approaches. These advantages include a flexible modeling framework, the ability to sample directly from the posterior distribution and not rely on large sample assumptions, an integrated framework to handle error and missing data, and practical computation with available software. As a likelihood-based approach, Bayesian methods can handle any standard two-phase sampling design (i.e., MAR); however, they do require extensive parametric model specifications, which may be difficult to realistically specify in practice. For example, correct specification of the joint likelihood of (X^*, Y^*, D^*, X, Y, D, Z) in the CCASAnet time-to-ADE analysis would be daunting.

There is a large body of work that focuses on how to handle covariate measurement error and misclassification (Gustafson 2003), including more recent works that look at addressing covariate error in propensity scores (Hong, Rudolph, and Stuart 2017) and a mixed effects quantile regression approach to handle covariates measured with error and an outcome subject to right censoring, with application to HIV/AIDS data (Tian, Tang, and Tian 2018). Similar to the frequentist framework, there is comparatively less work focused on outcome error. Speybroeck et al. (2013) outlined a Bayesian approach for estimating prevalence when the disease outcome is subject to misclassification. Others have considered Bayesian methods for estimating the relationship between a misclassified response and precise covariates (e.g., see Daniel Paulino, Soares, and Neuhaus 2003; Gerlach and Stamey 2007; Li et al. 2019). Hubbard et al. (2019) developed a Bayesian latent class model for binary outcome classification (phenotyping) for the EHR setting, where no validation data were available and covariates were subject to high levels of missingness according to a missing not at random pattern. Such a framework has the potential to be expanded to also include error-prone or misclassified covariate data. Important challenges for settings with more complex error structures may remain for this approach; in particular, model identifiability is a known challenge for latent class models (Gustafson 2005) and model identifiability for latent class models with misclassified outcomes is an active area of research for both Bayesian and non-Bayesian frameworks (Duan et al. 2019; Xia and Gustafson 2018).

Design-based estimation

Design-based estimation includes inverse probability weighting (IPW) and generalized raking, which is sometimes called 'calibration' in the survey sampling literature (Sarndal 2007). If certain records are more likely to be validated than others (e.g., cases are over-sampled), then it is important to account for the sampling design in the analysis. Some of the methods discussed above naturally do this (e.g., likelihood-based estimation), but others do not without some tweaks (e.g., moment-based estimation). The most popular way to account for the design is to use IPW. Let R_i be the indicator that a record i=1, … , n is in the validation (phase 2) sample, let π_i be its sampling probability, and $U_i(\beta)$ be an estimating equation for the parameter of interest β. Then the IPW estimator solves the equation $\sum_{i=1}^{n} R_i U_i(\beta)/\pi_i = 0$. However, IPW is known to be inefficient since it ignores the information in the

unvalidated data; a more efficient design-based estimator is to employ generalized raking (Deville, Särndal, and Sautory 1993). The basic idea behind generalized raking is to calibrate the inverse probability weights using an auxiliary variable available in the full cohort that contains information useful for estimating the target parameter of interest. Specifically, from the larger cohort (i.e., the phase 1 sample), we derive an auxiliary variable A_i with information useful for estimation. To gain efficiency, design weights, π_i, are adjusted so they remain close to their original values while satisfying the constraint that the weighted sum of A_i on the phase 2 sample equals its known sum on the phase 1 cohort. The raking estimator re-weights the phase 2 estimating equation with these weights (g_i/π_i), solving $\sum_{i=1}^{n} R_i g_i U_i(\beta)/\pi_i = 0$ for β. This can be thought of as calibrating the phase 2 sample to the phase 1 sample. For estimating a regression coefficient, useful auxiliary variables will be ones that are highly correlated with the influence functions for the target regression fit to the error-free data, since the sum of the influence functions on the population will estimate the regression slope parameter (Breslow et al. 2009; Lumley, Shaw, and Dai 2011). The influence function is not observed because the validated values of the variables are unknown for individuals not selected for the phase 2 sample. But one can approximate these functions with A, the influence functions from the target regression fit with the error-prone variables, or estimators such as the RC or MI estimators described above. The closer the estimating equation is to the one based on the true data, the more efficient the estimation should be; although in practice, we have seen that it is hard to do much better than just using influence functions from the regression fit with error-prone variables (Oh et al. 2019). This is the focus of future research. Design-based estimators have several attractive features including (1) they generally make fewer assumptions regarding the error structure than the other approaches described above, and (2) they are able to handle a wide variety of analyses, variables, and error structures. For example, Oh et al. (2019) addressed simultaneous errors in (X^*, Y^*, D^*) with time-to-event data using raking, and demonstrated that it was unbiased under minimal assumptions (i.e., MAR with π_i>0 for all i) and was more robust and achieved a smaller mean squared error than RC in several settings. Generalized raking is capable of handling the CCASAnet analysis of time from ART initiation to ADE (Oh et al. 2019).

Designs

The choice of the validation subsample can have a large impact on the bias and efficiency of estimates using error-prone routinely collected data. Statisticians can play an important role by ensuring that the validation sample is well-chosen and by furthering research on efficient designs. First, it is clear that the validation sample should be a probabilistic sample. This is necessary to ensure that missing at random assumptions, required to obtain unbiased estimates by the analytical methods outlined in the previous section, hold. Second, different probabilistic sampling schemes in combination with different estimation procedures can lead to more or less efficient estimation. Given the finite resources available to investigators, it can be very beneficial to target records for validation that will maximize information gained. We focus on this problem in this section.

Two-phase sampling

Two-phase sampling has been well-studied (e.g., Breslow and Chatterjee 1999; Gilbert, Yu, and Rotnitzky 2014; MacNamee 2005; Reilly 1996; Xu, Hui, and Grannis 2014; Zhao and Lipsitz 1992) and generally refers to predictors and outcomes being measured on all subjects (phase 1) and a more expensive measurement on only a subsample (phase 2). The two-phase sampling literature is very relevant for designing validation studies of routinely collected data. In our setting, phase 1 consists of the error-prone routinely collected data available on all subjects and phase 2 is the audit or validation sample.

With rare dichotomous outcomes, it is well known that over-sampling the cases improves the efficiency of odds ratios in two-phase designs (Prentice and Pyke 1979). Case-control and case-cohort designs and their variations are popular and efficient (Prentice 1986). These designs are a form of sampling the extremes. With time-to-event outcomes, extreme tail sampling takes the form of oversampling records with a short

time-to-event and records with a long time-to-censoring (Lawless 2018). In the CCASAnet example, this suggests one would want to validate records from patients who experienced an ADE early after ART initiation and from patients who did not have an ADE (i.e., were censored) after extensive follow-up. With continuous outcomes, sampling extremes of the outcomes or extremes of the residuals have been proposed (Lin, Zeng, and Tang 2013). In recent work, Tao, Zeng, and Lin (2019) showed that under the null that β=0, optimal sampling for likelihood-based estimation (which could also include MI and Bayesian estimators) corresponds to finding the sampling rule R that maximizes $E[Rvar(score|R{=}1,\ Z)\ var(X|X^*,Z)]$ where *score* denotes the score function (i.e., partial derivative of the log-likelihood with respect to β). With continuous outcomes, this corresponds to sampling extremes of the weighted residual, where the weight is given by $SD(X|X^*,\ Z)$, the standard deviation of X conditional on X^* and Z. In our setting, $SD(X|X^*, Z)$ is higher among those Z that are more error prone, which leads to a sensible sampling strategy. For example, if there were no errors at a particular site (Z=*siteC*), then $SD(X|X^*,\ Z{=}siteC){=}0$ and one would not gain information relevant to estimating β by performing audits at that site.

For design-based estimators (e.g., IPW and generalized raking), optimal two-phase designs follow somewhat different rules because one is not relying on a model. For example, design-based estimation does not work with extreme tail sampling described above because the probability of selection for some records is zero, so inverse probability weights are undefined. Strategies for efficient design-based sampling can be gleaned from the survey sampling literature (Sarndal, Swensson, and Wretman 2003). With given strata, Neyman-allocation, i.e., sampling proportional to the product of the stratum size and standard deviation, is optimal (Neyman 1934). For design-based estimation, to improve efficiency it would be desirable to form strata using the influence function for the parameter of interest given the correct data; again, the true influence function is not known, but a good choice for strata would be to use the expectation of the influence function given phase 1 data. Efficiency can improve by choosing strata such that based on Neyman-allocation, an equal number of records will be selected from each stratum (Waterhouse 1983); in general, higher numbers of strata also improve efficiency (Lumley 2011). We have seen that sampling based on these principles can lead to more efficient designs (Amorim et al. 2020).

Optimal validation sampling can be more challenging with routinely collected data. The event of interest may contain errors, so therefore case-control sampling on the phase 1 data, for instance, may not be optimal given that it is not known who is a true case. It can still be advantageous to choose samples based on strata of the unvalidated, error-prone data. In an example considered by Oh et al. (2019), which considered time to first ADE, more than half the error prone events were incorrect classifications; however, in this case the sensitivity was very high and so case-control sampling based on D^* resulted in nearly all the true cases being selected for the validation subset – a design that turned out to be highly efficient. Wang et al. (2017) considered a genotype stratified validation sample for linked EHR-genotype data with a misclassified EHR-derived binary outcome and proposed a maximum likelihood estimator for testing the genetic association that incorporates both the phase 1 and phase 2 data. Their proposed estimator was seen to have more efficiency than case-control sampling on the error prone case status.

In classical two-phase designs, observed variables in the phase 1 sample are typically assumed to be predictive of the expensive covariate, but they are rarely assumed to be surrogates for the true covariate. With errors in variables, the correlation between the true covariate and the observed covariate is often quite high and this information can be exploited to improve efficiency. Because the regression parameter for a target variable is asymptotically equivalent to the mean of the efficient influence functions (Lumley, Shaw, and Dai 2011), this influence function, $IF(X, Y)$, is the optimal stratifying variable with design-based regression analyses but is unknown because it is a function of the true variables. When X^* and Y^* are good proxies for X and Y, $IF(X^*, Y^*)$ can be an excellent stratification variable (Amorim et al. 2020).

Multi-wave sampling

Multi-wave sampling may be particularly beneficial for research with error-prone routinely collected data. Multi-wave validation designs can be thought of as having at least three sampling phases: obtain error-prone

data on all subjects (phase 1); audit an initial subset of records (phase 2, wave 1); then target additional audits based on results of the initial audit to improve efficiency (phase 2, wave 2). In practice, an initial audit (phase 2, wave 1) is often a quality-control check to uncover potential data problems. This first audit could then prompt a second whose purpose is no longer only to check data quality but to improve precision of estimates correcting for errors. Based on first audit results, a second audit may oversample records with more influence on the error-corrected estimators. For example, in a multi-site study, investigators may want to focus their audit resources on sites with particularly error-prone data. Importantly, optimal strategies for two-phase sampling depend on and are sensitive to unknown parameter values (MacNamee 2005; Xu, Hui, and Grannis 2014). For example, optimal designs for model-based estimation when $\beta = 0$ require knowledge of $SD(X|X^*, Z)$ which is unknown without a pilot sample (Tao, Zeng, and Lin 2019). Multi-wave audit designs can better estimate these parameters, and therefore maximize resources.

Adaptive sampling designs using pilot data to modify sampling have been advocated in other settings (Breslow and Chatterjee 1999; Fedorov, Wu, and Zhang 2012; Lohr 1990; McIsaac and Cook 2015; Wittes and Brittain 1990). For the setting of an error prone X^*, Reilly and Pepe (1995) proposed an efficient validation design for the mean score estimator that relies on pilot data. McIsaac and Cook (2015) further developed this idea by developing an adaptive phase 2 sampling design where the validation sample is divided into an initial pilot sample, from which nuisance parameters necessary for optimal mean score sampling are estimated, and then sampling of the remaining validation sample is adapted to achieve the estimated optimal proportions. Han et al. (2019) extended this work to the setting of discrete Cox proportional hazards models and studied its usefulness for validation designs considering continuous time to event outcomes. 'Three-stage' sampling has been investigated in the context of measurement error in both outcomes and predictors for linear models (Holcroft, Rotnitzky, and Robins 1997); however, stages 2 and 3 referred to measurements of the true outcome (stage 2) and true predictors (stage 3). In our setting, validation of outcomes and predictors generally occurs during the same sampling wave.

Conclusions and future research priorities

Electronic health record data and other routinely collected data are increasingly used for HIV research. These data are prone to errors; biostatisticians are in a unique position to impact science using these types of data. First, we should be strong advocates for the validation of data. Second, we can help investigators design efficient validation samples. Third, we can perform analyses that correctly incorporate the validation design and efficiently combine validated and unvalidated data.

There are also many opportunities for statistical methods development in this area. First, with regards to estimation it would be useful to extend existing measurement error methods to handle more complex error structures seen in practice. We have provided some guidance on which types of methods we believe are more or less amenable to the complexities of electronic health records data, but many extensions are possible. With that said, for these types of methods to have an impact on HIV/AIDS and more general biomedical research, there must be a focus on practicalities. Simple methods with some bias or inefficiency may be better than complex methods. User-friendly software implementations are also critical.

Second, research into more efficient designs for selecting validation samples is needed. Although we outlined some preliminary work, there are many open areas of research. Most existing designs ignore the fact that there may be correlated errors across multiple variables. Optimal sampling designs for likelihood-based estimation procedures have not been formally derived except in special cases, and deserve further investigation. Preliminary work has focused on efficient estimation of a single regression coefficient. In some cases, a prediction model may be the ultimate goal; efficient designs for prediction would presumably require simultaneous efficient estimation of regression coefficients for all predictor variables. In practice, in a research network like CCASAnet, there are often many high priority studies/outcomes, but insufficient funds to perform a validation study for each of them; pragmatic validation strategies to efficiently address many research

questions warrant additional study. In addition, different outcomes may require different validation designs. For example, validation of ADE through chart review is quite different from validation of mortality by contact tracing patients who are lost to follow-up. Also, we believe multi-wave validation strategies hold great promise. However, many questions remain that are specific to multi-wave sampling. For example, what should be the size of the initial audit?

Finally, a critical area of future research is to prospectively apply these designs and methods to address important HIV/AIDS questions. Ultimate uptake of these methods by biomedical researchers will be based on visibly successful implementation of these approaches in practical settings. We encourage statisticians engaged in research using routinely collected biomedical data to implement these designs and methods.

Acknowledgments: We would like to acknowledge our collaborators (both statistical and clinical), including Ran Tao, Gustavo Amorim, Mark Giganti, Sarah Lotspeich, Kyunghee Han, Eric Oh, Lillian Boe, Thomas Lumley, Timothy Sterling, Catherine McGowan, and Stephany Duda. We would also like to generally acknowledge the CCASAnet database managers and data abstractors. This work was funded in part by NIH grants R01AI131771, U01AI069923 (CCASAnet), and PCORI R-1609-36207.
Research funding: This work was funded in part by NIH grants R01AI131771, U01AI069923 (CCASAnet), and PCORI R-1609-36207.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Informed consent was obtained from all individuals included in this study.
Ethical approval: Research involving human subjects complied with all relevant national regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013), and has been approved by the authors' Institutional Review Board (xxxx) or equivalent committee. (xxx-Nr.: xx/x) **OR** The local Institutional Review Board deemed the study exempt from review.

References

Alexeeff, S. E., R. J. Carroll, and B. Coull. 2016. "Spatial Measurement Error and Correction by Spatial SIMEX in Linear Regression Models when Using Predicted Air Pollution Exposures." *Biostatistics* 17: 377–89.

Amorim, G., R. Tao, S. Lotspeich, P. Shaw, T. Lumley, and B. Shepherd. 2020. "Two-Phase Sampling Designs for Data Validation in Settings with Measurement Error." (submitted).

Balasubramanian, R., and S. Lagakos. 2003. "Estimation of a Failure Time Distribution Based on Imperfect Diagnostic Tests." *Biometrika* 90: 171–82.

Bartlett, J. W., and R. H. Keogh. 2018. "Bayesian Correction for Covariate Measurement Error: A Frequentist Evaluation and Comparison with Regression Calibration." *Statistical Methods in Medical Research* 27: 1695–708.

Boe, L. A., L. F. Tinker, and P. A. Shaw. 2020. "An Approximate Quasi-Likelihood Approach for Error-Prone Failure Time Outcomes and Exposures." arXiv preprint arXiv:2004.01112.

Bosh, K. A., N. Crepaz, X. Dong, S. Lyss, M. Mendoza, and A. J. Mitsch. 2019. "Opioid Overdose Deaths Among Persons with HIV Infection, United States, 2011–2015." *Conference on Retroviruses and Opportunistic Infections*. Seattle, WA.

Breslow, N., and N. Chatterjee. 1999. "Design and Analysis of Two-Phase Studies with Binary Outcome Applied to Wilms Tumour Prognosis." *Applied Statistics* 48: 457–68.

Breslow, N., T. Lumley, C. Ballantyne, L. Chambless, and M. Kulich. 2009. "Improved Horvitz–Thompson Estimation of Model Parameters from Two-Phase Stratified Samples: Applications in Epidemiology." *Statistics in Biosciences* 1: 32–49.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective*. Boca Raton: Chapman and Hall/CRC.

Cole, S. R., H. Chu, and S. Greenland. 2006. "Multiple-Imputation for Measurement-Error Correction." *International Journal of Epidemiology* 35: 1074–81.

Cook, J. R., and L. A. Stefanski. 1994. "Simulation-Extrapolation Estimation in Parametric Measurement Error Models." *Journal of the American Statistical Association* 89: 1314–28.

Daniel Paulino, C., P. Soares, and J. Neuhaus. 2003. "Binomial Regression with Misclassification." *Biometrics* 59: 670–5.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. "Generalized Raking Procedures in Survey Sampling." *Journal of the American Statistical Association* 88: 1013–20.

Dodd, L., E. Korn, B. Freidlin, R. Gray, and S. Bhattacharya. 2011. "An Audit Strategy for Progression-Free Survival." *Biometrics* 67: 1092–9.

Duan, R., M. Cao, Y. Ning, M. Zhu, B. Zhang, A. McDermott, H. Chu, X. Zhou, J. H. Moore, J. G. Ibrahim, D. O. Scharfstein, and Y. Chen. 2019. "Global Identifiability of Latent Class Models with Applications to Diagnostic Test Accuracy Studies: A Gröbner Basis Approach." *Biometrics* 76: 98–108.

Duda, S., B. Shepherd, C. Gadd, D. R. Masys, and C. C. McGowan. 2012. "Measuring the Quality of Observational Study Data in an International HIV Research Network." *PLoS One* 7: e33908.

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

Fedorov, V., Y. Wu, and R. Zhang. 2012. "Optimal Dose-Finding Designs with Correlated Continuous and Discrete Responses." *Statistics in Medicine* 31: 217–34.

Geng, E. H., N. Emenyonu, M. B. Bwana, D. V. Glidden, and J. N. Martin. 2008. "Sampling-based Approach to Determining Outcomes of Patients Lost to Follow-Up in Antiretroviral Therapy Scale-Up Programs in Africa." *Journal of the American Medical Association* 300: 506–7.

Gerlach, R., and J. Stamey. 2007. "Bayesian Model Selection for Logistic Regression with Misclassified Outcomes." *Statistical Modelling* 7: 255–73.

Giganti, M., and B. Shepherd. 2020. "Multiple Imputation Variance Estimation in Studies with Missing or Misclassified Inclusion Criteria." *American Journal of Epidemiology*, https://doi.org/10.1093/aje/kwaa153 (Epub ahead of print).

Giganti, M., P. Shaw, G. Chen, S. Bebawy, M. Turner, T. Sterling, and B. Shepherd. 2020. "Accounting for Dependent Errors in Predictors and Time-To-Event Outcomes Using Electronic Health Records, Validation Samples, and Multiple Imputation." *Annals of Applied Statistics* 14: 1045–61.

Gilbert, P. B., X. Yu, and A. Rotnitzky. 2014. "Optimal Auxiliary-Covariate-Based Two-Phase Sampling Design for Semiparametric Efficient Estimation of a Mean or Mean Difference, with Application to Clinical Trials." *Statistics in Medicine* 33: 901–17.

Gravel, C. A., A. Dewanji, P. J. Farrell, and D. Krewski. 2018. "A Validation Sampling Approach for Consistent Estimation of Adverse Drug Reaction Risk with Misclassified Right-Censored Survival Data." *Statistics in Medicine*, https://doi.org/10.1002/sim.7854 (Epub ahead of print).

Gu, X., Y. Ma, and R. Balasubramanian. 2015. "Semiparametric Time to Event Models in the Presence of Error-Prone, Self-Reported Outcomes—With Application to the Women's Health Initiative." *The Annals of Applied Statistics* 9: 714–30.

Gustafson, P. 2003. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. Boca Raton: Chapman and Hall/CRC.

Gustafson, P. 2005. "On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables." *Statistical Science* 20: 111–40.

Han, K., T. Lumley, B. Shepherd, and P. Shaw. 2019. "Design and Analysis of Two-Phase Samples in Discrete-Time Survival Analysis with Error-Prone Exposures." In *Joint Statistical Meetings of the American Statistical Association*. Denver, CO.

He, W., G. Y. Yi, and J. Xiong. 2007. "Accelerated Failure Time Models with Covariates Subject to Measurement Error." *Statistics in Medicine* 20: 4817–32.

Holcomb, J. 1999. "Regression with Covariates and Outcome Calculated from a Common Set of Variables Measured with Error: Estimation Using the SIMEX Method." *Statistics in Medicine* 26: 2847–62.

Holcroft, C., A. Rotnitzky, and J. Robins. 1997. "Efficient Estimation of Regression Parameters from Multistage Studies with Validation of Outcome and Covariates." *Journal of Statistical Planning and Inference* 65: 349–74.

Hong, S., N. Schmitt, A. Stone, and J. Denne. 2012. "Attenuation of Treatment Effect Due to Measurement Variability in Assessment of Progression-Free Survival." *Pharmaceutical Statistics* 11: 394–402.

Hong, H., K. E. Rudolph, and E. A. Stuart. 2017. "Bayesian Approach for Addressing Differential Covariate Measurement Error in Propensity Score Methods." *Psychometrika* 82: 1078–96.

Huang, J., R. Duan, R. A. Hubbard, Y. Wu, J. A. Moore, H. Xu, and Y. Chen. 2018. "PIE: A Prior Knowledge Guided Integrated Likelihood Estimation Method for Bias Reduction in Association with Studies Using Electronic Health Records Data." *Journal of the American Medical Informatics Association* 25: 345–52.

Hubbard, R. A., J. Huang, J. Harton, A. Oganisian, G. Choi, L. Utidjian, I. Eneli, L. C. Bailey, and Y. Chen. 2019. "A Bayesian Latent Class Approach for EHR-Based Phenotyping." *Statistics in Medicine* 38: 74–87.

Kiragga, A., B. Castelnuovo, P. Schaefer, T. Muwonge, and P. Easterbrook. 2011. "Quality of Data Collection in a Large HIV Observational Clinic Database in Sub-Saharan Africa: Implications for Clinical Research and Audit of Care." *Journal of the International AIDS Society* 14: 3.

Kitahata, M., D. Drozd, H. Crane, S. E. Van Rompaey, K. N. Althoff, S. J. Gange, M. B. Klein, G. M. Lucas, A. G. Abraham, V. Lo Re, J. McReynolds, W. B. Lober, A. Mendes, S. P. Modur, Y. Jing, E. J. Morton, M. A. Griffith, A. M. Freeman, and R. D. Moore. 2015. "Ascertainment and Verification of End-Stage Renal Disease and End-Stage Liver Disease in North American AIDS Cohort Collaboration on Design and Research." *AIDS Research and Treatment* 2015: 923194.

Korn, E. L., L. E. Dodd, and B. Freidlin. 2010. "Measurement Error in the Timing of Events: Effect on Survival Analyses in Randomized Clinical Trials." *Clinical Trials* 7: 626–33.

Lawless, J. 2018. "Two-phase Outcome-Dependent Studies for Failure Times and Testing for Effects of Expensive Covariates." *Lifetime Data Analysis* 24: 28–44.

Ledergerber, B. 2012. "Data Quality in Cohort Collaborations: Should We Let Sleeping Dogs Lie?" In *16th International Workshop on HIV Observational Databases*. Athens, Greece.

Li, L., A. Jara, M. J. García-Zattera, and T. E. Hanson. 2019. "Marginal Bayesian Semiparametric Modeling of Mismeasured Multivariate Interval-Censored Data." *Journal of the American Statistical Association* 114: 129–45.

Lin, D., D. Zeng, and Z. Tang. 2013. "Quantitative Trait Analysis in Sequencing Studies under Trait-Dependent Sampling." *Proceedings of the National Academy of Sciences, USA* 110: 12247–52.

Little, R., and D. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: Wiley.

Lohr, S. L. 1990. "Accurate Multivariate Estimation Using Triple Sampling." *Annals of Statistics* 18: 21615–33.

Lotspeich, S., M. Giganti, M. Maia, R. Vieira, D. Machado, R. Succi, S. Ribeiro, M. Pereira, B. Shepherd, C. McGowan, and S. Duda. 2019. "Self-audits as Alternative to Travel-Audits for Improving Data Quality in the Caribbean, Central and South America Network for HIV Epidemiology." (submitted).

Lotspeich, S., B. Shepherd, E., G. Amorim, P. Shaw, and R. Tao. 2020. Submitted for publication. "Efficient odds ratio estimation using error-prone data from a multi-national HIV research cohort." (submitted).

Lumley, T., P. Shaw, and J. Dai. 2011. "Connections between Survey Calibration Estimators and Semiparametric Models for Incomplete Data." *International Statistical Review* 79: 200–20.

Lumley, T. 2011. *Complex Surveys: A Guide to Analysis Using R*, Vol. 565. John Wiley & Sons.

MacKellar, D., R. Nelson, R. Thompson, I. Casavant, S. Pals, I. Pathmanathan, J. Cardoso, D. Ujamaa, E. Yufenyuy, K. Sleeman, V. Chivurre, N. Chicuecue, K. Oladapo, A. Couto, and A. Vergara. 2019. "Fifty-Percent Reduction in HIV Incidence in Chokwe District, Mozambique, 2014–2017." In *Conference on Retroviruses and Opportunistic Infections*. Seattle, WA.

MacNamee, R. 2005. "Optimal Design and Efficiency of Two-Phase Case-Control Studies with Error-Prone and Error-Pree Exposure Measures." *Biostatistics* 6: 590–603.

McIsaac, M. A., and R. J. Cook. 2015. "Adaptive Sampling in Two-Phase Designs: A Biomarker Study for Progression in Arthritis." *Statistics in Medicine* 34: 2899–912.

Meier, A. S., B. A. Richardson, and J. P. Hughes. 2003. "Discrete Proportional Hazards Models for Mismeasured Outcomes." *Biometrics* 59: 947–54.

Mofenson, L. M. 2019. "Update on Antiretroviral Drugs and Birth Defects." In *Conference on Retroviruses and Opportunistic Infections*. Seattle, WA.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97: 558–606.

Oh, E. J., B. E. Shepherd, T. Lumley, and P. A. Shaw. 2018. "Considerations for Analysis of Time-To-Event Outcomes Measured with Error: Bias and Correction with SIMEX." *Statistics in Medicine* 37: 1276–89.

Oh, E. J., B. E. Shepherd, T. Lumley, and P. A. Shaw. 2019. "Raking and Regression Calibration: Methods to Address Bias from Correlated Covariate and Time-To-Event Error." arXiv preprint arXiv:1905.08330.

Parveen, N., E. Moodie, and B. Brenner. 2017. "Correcting Covariate-Dependent Measurement Error with Non-zero Mean." *Statistics in Medicine* 36: 2786–800.

Prentice, R. L., and R. Pyke. 1979. "Logistic Disease Incidence Models and Case-Control Studies." *Biometrika* 66: 403–11.

Prentice, R. L. 1982. "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model." *Biometrika* 69: 331–42.

Prentice, R. L. 1986. "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials." *Biometrika* 73: 1–11.

Reilly, M., and M. S. Pepe. 1995. "A Mean Score Method for Missing and Auxiliary Covariate Data in Regression Models." *Biometrika* 82: 299–314.

Reilly, M. 1996. "Optimal Sampling Strategies for Two-Stage Studies." *American Journal of Epidemiology* 143: 92–100.

Robins, J., and N. Wang. 2000. "Inference for Imputation Estimators." *Biometrika* 87: 113–24.

Sarndal, C., B. Swensson, and J. Wretman. 2003. *Model Assisted Survey Sampling*. New York: Springer Verlag.

Sarndal, C. 2007. "The Calibration Approach in Survey Theory and Practice." *Survey Methodology* 33: 99–119.

Scott, A., and C. Wild. 1997. "Fitting Regression Models to Case-Control Data by Maximum Likelihood." *Biometrika* 84: 57–61.

Shaw, P. A., and R. L. Prentice. 2012. "Hazard Ratio Estimation for Biomarker-Calibrated Dietary Exposures." *Biometrics* 68: 397–407.

Shaw, P., J. He, and B. Shepherd. 2018. "Regression Calibration to Correct Correlated Errors in Outcome and Exposure." arXiv preprint arXiv:1811.10147.

Shepherd, B. E., and C. Yu. 2011. "Accounting for Data Errors Discovered from an Audit in Multiple Linear Regression." *Biometrics* 67: 1083–91.

Shepherd, B. E., P. A. Shaw, and L. E. Dodd. 2012. "Using Audit Information to Adjust Parameter Estimates for Data Errors in Clinical Trials." *Clinical Trials* 9: 721–9.

Skinner, C. J., and K. Humphreys. 1999. "Weibull Regression for Lifetimes Measured with Error." *Lifetime Data Analysis* 5: 23–37.

Speybroeck, N., B. Devleesschauwer, L. Joseph, and D. Berkvens. 2013. "Misclassification Errors in Prevalence Estimation: Bayesian Handling with Care." *International Journal of Public Health* 58: 791–5.

Tang, L., R. Lyles, C. King, D. Celentano, and Y. Lo. 2015. "Binary Regression with Differentially Misclassified Response and Exposure Variables." *Statistics in Medicine* 34: 1605–20.

Tao, R., D. Zeng, and D. Y. Lin. 2017. "Efficient Semiparametric Inference under Two-Phase Sampling, with Applications to Genetic Association Studies." *Journal of the American Statistical Association* 112: 1468–76.

Tao, R., D. Zeng, and D. Y. Lin. 2019. "On Optimal Two-Phase Designs." *Journal of the American Statistical Association* (in press). https://doi.org/10.1080/01621459.2019.1671200.

Tao, R., S. C. Lotspeich, P. A. Shaw, and B. E. Shepherd. 2020. "Efficient Semiparametric Inference for Two-Phase Studies with Outcome and Covariate Measurement Errors." (submitted).

Tian, Y., M. Tang, and M. Tian. 2018. "Joint Modeling for Mixed-Effects Quantile Regression of Longitudinal Data with Detection Limits and Covariates Measured with Error, with Application to AIDS Studies." *Computational Statistics* 33: 1563–87.

Van Buuren, S. 2012. *Flexible Imputation of Missing Data*. Boca Raton: Chapman and Hall/CRC.

Wang, L., S. M. Damrauer, H. Zhang, A. X. Zhang, R. Xiao, J. H. Moore, and J. Chen. 2017. "Phenotype Validation in Electronic Health Records Based Genetic Association Studies." *Genetic Epidemiology* 41: 790–800.

Waterhouse, W. 1983. "Do symmetric Problems Have Symmetric Solutions?." *The American Mathematical Monthly* 90: 378–87.

Weng, C., P. Appelbaum, G. Hripcsak, I. Kronish, L. Busacca, K. W. Davidson, and J. T. Bigger. 2012. "Using EHRs to Integrate Research with Patient Care: Promises and Challenges." *Journal of the American Medical Informatics Association* 19: 684–7.

Wittes, J., and E. Brittain. 1990. "The Role of Internal Pilot Studies in Increasing the Efficacy of Clinical Trials." *Statistics in Medicine* 9: 65–72.

Xia, M., and P. Gustafson. 2018. "Bayesian Inference for Unidirectional Misclassification of a Binary Response Trait." *Statistics in Medicine* 37: 933–47.

Xie, X., and X. Meng. 2017. "Dissecting Multiple Imputation from a Multiphase Inference Perspective: What Happens when God's, Imputer's and Analysts Models Are Uncongenial?." *Statistica Sinica* 27: 1485–545.

Xie, S. X., C. Y. Wang, and R. L. Prentice. 2001. "A Risk Set Calibration Method for Failure Time Regression by Using a Covariate Reliability Sample." *Journal of the Royal Statistical Society: Series B* 63: 855–70.

Xu, H., S. L. Hui, and S. Grannis. 2014. "Optimal Two-Phase Sampling Design for Comparing Accuracies of Two Binary Classification Rules." *Statistics in Medicine* 10: 500–13.

Zhao, L. P., and S. Lipsitz. 1992. "Designs and Analysis of 2-Stage Studies." *Statistics in Medicine* 11: 769–82.