Three-phase generalized raking and multiple imputation estimators to address error-prone data

Gustavo Amorim1 | Ran Tao1,2 | Sarah Lotspeich1,3 | Pamela A. Shaw4 | Thomas Lumley5 | Rena C. Patel6 | Bryan E. Shepherd1

1Department of Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, USA
2Vanderbilt Genetics Institute, Vanderbilt University Medical Center, Nashville, Tennessee, USA
3Department of Statistical Sciences, Wake Forest University, Winston-Salem, North Carolina, USA
4Biostatistics Division, Kaiser Permanente Washington Health Research Institute, Seattle, Washington, USA
5Department of Statistics, University of Auckland, Auckland, New Zealand
6Department of Medicine, University of Washington, Seattle, Washington, USA

Validation studies are often used to obtain more reliable information in settings with error-prone data. Validated data on a subsample of subjects can be used together with error-prone data on all subjects to improve estimation. In practice, more than one round of data validation may be required, and direct application of standard approaches for combining validation data into analyses may lead to inefficient estimators since the information available from intermediate validation steps is only partially considered or even completely ignored. In this paper, we present two novel extensions of multiple imputation and generalized raking estimators that make full use of all available data. We show through simulations that incorporating information from intermediate steps can lead to substantial gains in efficiency. This work is motivated by and illustrated in a study of contraceptive effectiveness among 83,671 women living with HIV, whose data were originally extracted from electronic medical records, of whom 4732 had their charts reviewed, and a subsequent 1210 also had a telephone interview to validate key study variables.

KEYWORDS
data audits, design-based estimator, electronic medical records, measurement error, multiple imputation, three-phase design

1 INTRODUCTION

Millions of women living with HIV are on antiretroviral therapy (ART) and some form of contraception. There is some evidence that a commonly prescribed antiretroviral drug, efavirenz, reduces the efficacy of contraceptive implants at preventing unwanted pregnancies.1,2 A large study using routinely collected electronic health record (EHR) data from over 80,000 women with HIV in Kenya was performed to investigate this potential association. Because of concerns about the quality of the EHR data, chart reviews were performed on a probabilistic sample of approximately 5000 women; chart reviews involved validating ART regimens, contraceptive methods, pregnancy status, and associated dates based on a thorough review of sampled women's medical records. A second stage of validation via telephone interview was...
conducted on a subsample of over 1200 women whose charts were reviewed. Those women who received the telephone interview had fully validated data whereas those whose charts were reviewed but did not have a telephone interview had partially validated data.

In Patel et al, they presented findings from this Kenyan study to a clinical audience. Among women on contraceptive implants, we found that women on efavirenz were approximately three times more likely to become pregnant than those on another commonly prescribed antiretroviral, nevirapine. The analyses of that manuscript involved extending a statistical technique, generalized raking (GR), to allow us to incorporate all three phases of data collection into the analysis. These extensions were novel and resulted in substantial gains in power over existing approaches, while maintaining a high level of robustness to potential model misspecification. The purpose of the current article is to statistically describe and justify the choice of our methods, as well as to develop alternative multiple imputation based analysis approaches, which may be attractive due to their potential efficiency gains to address this analytic challenge. Routinely collected data are being used more frequently for biomedical research, these methods are likely to be applicable in other settings.

Using statistical terminology, the Kenyan study can be thought of as a nested three-phase design to address measurement error across multiple variables. Phase-1 is the EHR data collection, phase-2 is the chart reviews, and phase-3 is the telephone interviews.

Traditional two-phase designs collect information that may be costly or hard to obtain on a subsample of patients in the original dataset. Studies that address measurement error with a validation subsample can be thought of as two-phase designs where phase-1 contains error-prone surrogates of the variable of interest. However, until recently, most of the statistical literature on measurement error focused on settings with only covariate measurement error; our setting - and similar settings using EHR data - includes multiple variables, including outcomes and predictors, subject to measurement error, and the presence and magnitude of errors may be dependent across variables.

Complicated measurement error settings have been addressed by recognizing that the problem can be cast as a missing data problem. Complete data are only available for those with fully validated data, and the validated values of variables are missing for those that are not in the validation set. Under the assumption that validation data are missing at random (which is often the case since researchers decide which records to validate), missing validated data can be multiply imputed based on models fit in the phase-2 sample. Alternatively, validated data in the phase-2 sample can be upweighted to represent the larger phase-1 sample using inverse probability weights (IPW) or other more efficient weighting approaches such as GR. GR estimators improve upon the efficiency of IPW estimators without extra modeling assumptions, and asymptotically, the best augmented inverse probability weighting (AIPW) estimators and are quite easy to implement with standard software. A newer body of literature has developed and applied both MI and GR techniques to analyze two-phase EHR data with errors across multiple variables.

However, we are not aware of any study that has addressed measurement error across multiple variables in a three-phase design. Although there are several examples in the literature of three-phase designs, methods for the analysis of data resulting from these designs are lacking, restricted to specific simpler or different situations, and do not easily extend to our setting with errors across multiple variables. Many of the current approaches in the literature ignore the phase-2 (partially validated) data in analyses. As will be seen in this manuscript, there is the potential for substantial gains in efficiency by incorporating partially validated data into analyses.

In this manuscript, we propose and illustrate three estimators that make use of all three phases of data collection to address measurement error across multiple variables. Specifically, we develop a multiple imputation (MI) estimator, a GR estimator, and a third estimator based on both MI and GR. The MI estimator is based on a sequential, multistage imputation approach that treats validated data as missing for those not selected for validation, and then multiply imputes them sequentially across the multiple phases. The GR estimator accounts for both stages of sampling by calibrating weights using auxiliary information in both the phase-1 and phase-2 samples. Lastly, the estimator based on both GR and MI seeks to improve the efficiency of the GR method by using the imputed dataset. The performance of our estimators, in terms of efficiency and bias, is examined via simulations modeled after the complex three-phase validation study that motivates this work. We show substantial gains in efficiency when all validation steps are properly incorporated into the analyses. Finally, using the database of Patel et al, all proposed methods are used to examine whether efavirenz-containing ART reduces the effectiveness of contraceptive implants to prevent pregnancies. We show that, similar to our simulations, using information from all validation steps leads to smaller variances in real data.
2 | MOTIVATING EXAMPLE

In this section, we provide a few more details about our motivating data example. A retrospective analysis was conducted on a longitudinal cohort of women living with HIV, ages 15 to 45 years, from western Kenya from January 1, 2011 to December 31, 2015. The main objective was to assess whether efavirenz-containing ART decreased effectiveness of contraceptive implants as measured by pregnancy. Contraceptive methods were documented at each clinic visit and the leading contraceptives were categorized as (1) implants, (2) depomedroxyprogesterone acetate (DMPA), (3) less-effective family planning (LEFP) methods, or (4) no contraceptive method (no family planning). Similarly, ART regimens were recorded at each visit and leading ART regimens were categorized as one of the following: (1) efavirenz-containing ART, (2) nevirapine-containing ART, (3) protease inhibitor (PI)-containing ART, or (4) no ART. We note that there were a few other contraceptive methods and ART regimens in the EHR, but as they accounted for less than 4% of person-years in total and were of less interest to study investigators, they were excluded from the analyses presented here. Because of these exclusions, the numbers presented in this manuscript slightly differ from the numbers presented in Patel et al. The date of incident pregnancy (ie, conception) was estimated using reported last menstrual period, estimated gestational age, or date of delivery. Person-time during pregnancy (ie, time between incident pregnancy until delivery or termination of pregnancy) was excluded from analyses.

The primary goal of this study was to investigate whether efavirenz-containing ART, compared to nevirapine-containing ART, decreased the effectiveness of contraceptive implants among women living with HIV. We calculated the incidence rate ratio (IRR) for the effectiveness of efavirenz-containing ART with contraceptive implant, compared to nevirapine-containing ART with contraceptive implant, via a Poisson regression model with pregnancy status as the outcome; contraceptive, ART regimen, as well as their interactions, as exposures of interest; time as an offset (patient follow-up time was divided into time intervals with each interval corresponding to a certain combination of ART and contraception); and age at baseline and study site as covariates. No lag was incorporated in the analysis. A model for the time-varying exposure was not included.

To address potential errors in the EHR data, a three-phase sampling scheme was devised. The \( n_1 = 83,671 \) women with EHR data (phase-1) were stratified into 32 groups based on combinations of pregnancy status (yes/no), ART regimen (efavirenz, nevirapine, PI, and no ART), and contraceptive use (implants, DMPA, LEFP, and no contraceptive method). A total of \( n_2 = 47,322 \) women were then sampled for initial validation via paper chart review (phase-2). Different sampling probabilities were defined for each stratum, with women in more clinically relevant strata at least once during their follow-up period (eg, pregnancy while on implant) more likely to be selected for chart review than those in strata of lower priority (eg, not on ART and not pregnant). More information on sampling probabilities can be found in Patel et al. The chart review validation step consisted of reviewing ART regimens, types of contraceptives, pregnancy status, and all relevant dates. A total of \( n_3 = 1,210 \) women were then further selected for telephone interviews (phase-3). Patients who had their charts reviewed (ie, selected for phase-2) were re-stratified, based again on their EHR data, as follows: (1) contraceptive implant and pregnant, (2) contraceptive implant and not pregnant, (3) DMPA and pregnant, or (4) other. This hierarchical order from (1) to (4), with (1) being the most clinically relevant, was used to uniquely assign patients that were in more than one group during their follow-up. Patients were then sorted in terms of priority (from 1 to 4) and approached accordingly for telephone interviews; the phase-3 sample was essentially a convenience sample based on this prioritization. In the telephone interviews, ART regimens, types of contraceptives, pregnancy status, and all relevant dates were again validated. If a woman could not remember her ART regimen or type of contraceptive during the telephone interview, the information in the chart review was used.

In summary, we have EHR data on \( n_1 = 83,671 \) women (phase-1); chart review data on a subset of \( n_2 = 47,322 \) women (phase-2); and telephone interview data (in addition to chart review data) on a subset of \( n_3 = 1,210 \) women (phase-3). We consider those \( n_3 = 1,210 \) records with both chart review and telephone interview data to be fully validated as our gold-standard, whereas the \( n_2 - n_3 = 3,522 \) records with only chart review data are partially validated. A total of 12,432 pregnancies were recorded in the EHR, from 11,259 different women, while 1116 and 388 pregnancies were recorded in the charts and telephone interviews, respectively.

Figure 1 displays the monthly concordance of ART regimens, contraceptive types, and pregnancy status across the three sampling phases. To obtain this figure we discretized time into monthly intervals, so that the EHR, chart review, and telephone interview datasets were all in the same scale and the recorded values could be compared in a monthly basis. Misclassification rates were, in general, fairly small among ART regimens, especially when comparing data from the EHR to chart reviews. For example, among the patients selected into phase-2 that had listed use of a nevirapine-containing ART regimen in the EHR, in about 95% of the months of follow up, use of a nevirapine-containing ART was also recorded.
in their charts. This concordance rate remained high when comparing the chart review to the telephone interview but decreased slightly when comparing the EHR directly to telephone interview data. In contrast, error rates in contraceptive use were substantially higher. For example, considering women selected into phase-2 with no family planning in the EHR, of the months they listed as no family planning, 26% were actually found to be on implant according to the chart review. Slightly higher error rates were seen when comparing the telephone interview and EHR datasets. The number and time (month) of pregnancies differed substantially across the three datasets, as displayed in Figure 1C. Considering only women selected for chart review, 997 pregnancies were recorded in the EHR, of which only 354 were substantiated.
TABLE 1  Estimated incidence rate ratio (IRR) for the effect of efavirenz-containing ART compared to nevirapine-containing ART on pregnancy among women on implant contraceptives, accounting for the sampling schemes.

| Method                                      | IRR    | 95% CI          | SD(log(I RR)) |
|---------------------------------------------|--------|-----------------|---------------|
| Naive analyses\(^a\)                       |        |                 |               |
| EHR (phase-1 only)                         | 2.13   | (1.63, 2.78)    | 0.14          |
| Chart review (phase-2 only)                | 2.12   | (1.58, 2.86)    | 0.15          |
| Telephone interview (phase-3 only)         | 3.04   | (1.82, 5.08)    | 0.26          |
| Two- and three-phase analysis\(^b\)        |        |                 |               |
| IPW                                         | 3.82   | (1.39, 10.50)   | 0.52          |
| Two-phase GR                                | 4.09   | (1.50, 11.12)   | 0.51          |
| Two-phase GR + MI                          | 3.61   | (1.40, 9.31)    | 0.48          |
| Three-phase GR                             | 4.08   | (1.64, 10.12)   | 0.46          |
| Three-phase GR + MI                        | 3.06   | (1.27, 7.37)    | 0.45          |
| Two-phase MI                               | 3.54   | (1.81, 7.21)    | 0.34          |
| Three-phase MI                             | 1.81   | (1.03, 3.10)    | 0.31          |

Abbreviations: ART, antiretroviral treatment; CI, confidence interval; EHR, electronic medical records; GR, generalized raking; IPW, inverse probability weighting; IRR, incidence rate ratio; MI, multiple imputation; SD, standard deviation.

\(^a\) Individual analysis for each dataset (EHR, chart review, and telephone interview). Pregnancy was modeled via a Poisson regression with contraceptive type, ART, and their interaction as main exposures, while adjusting for age and study site (AMPATH or FACES) and using follow-up time as offset.

\(^b\) The same regression model as in the naive analysis was used, with weighting adjustments when required (IPW and GR).

in the same month by the medical chart, leading to a misclassification rate of nearly 65%. Considering fully validated women, there were 277 and 237 pregnancies found in the EHR and chart review, respectively, of which only 104 and 133, respectively, were in the same months as reported in the telephone interview. Some errors were due to incorrect recording of pregnancy status, whereas others were due to incorrect recording of pregnancy start dates.

We start by analyzing each dataset separately, without accounting for the three-phase sampling scheme. Results are shown at the top of Table 1. All methods suggested that women on implants who were taking efavirenz-containing ART had a significantly higher risk of pregnancy than those taking nevirapine-containing ART. However, the point estimates differed, which is not surprising given the high degree of misclassification across the datasets. The IRR obtained from the telephone interview dataset was nearly 50% higher than that using the error-prone EHR dataset only. However, these separate analyses are not only inefficient as they do not take into account all information simultaneously, but they may also be biased because they ignore the data errors and/or the sampling design. In the next section, we discuss approaches that incorporate all three datasets into a single analysis to provide valid and efficient estimates for the parameters of interest.

3  METHODS

3.1  Multiple imputation

As mentioned in Section 1, validation studies are essentially missing data problems. The fully validated data are only available for a small subsample of the study population, for example, patients selected for telephone interview. Records with only EHR data or only selected for chart review are not fully validated and are still error-prone. Multiple imputation consists of imputing the truth, that is, correct contraceptive types, ART regimens, and pregnancy status from the telephone interview stage, multiple times for all unvalidated (phase-1, EHR data) or partially validated (phase-2, chart review data only) subjects. Conventional analyses are then performed with the complete/fully imputed data. Final estimates for the parameters of interest and their respective variances can then be computed by combining results from the multiple imputations.\(^{23}\) MI estimates will be consistent if (1) data are missing at random (ie, selection of records for validation is independent of the true values of the variables conditional on observed/prevalidation variables) and (2) the imputation and analysis models are properly specified.
However, direct application of MI in complex settings, for example, in longitudinal studies with time-varying covariates or different follow-up times, are not straightforward. This is because validating time-varying variables can lead to datasets that are difficult to merge: for example, a patient’s EHR could have $K_1$ time intervals, with each interval corresponding to a certain combination of ART, contraceptive, and pregnancy; her chart review could potentially have a different number of intervals (say, $K_2$), with different starting and ending dates, and so could her telephone interview (say, $K_3$). In addition, correctly modeling the functional relationship between the time-varying covariates, such as ART regimen and contraceptive use, may be hard, and model misspecification may lead to biased estimates. To overcome these issues, we extend the time-discretized MI of Giganti et al. Specifically, we deconstruct the follow-up time into 1-month intervals. We then define exposures and outcomes as present/absent in each interval. We decompose the joint distribution of the discretized versions of the exposure and outcome variables using chained equations. We fit these chain equations in the completely validated data and then use them to impute for unvalidated records. Error-prone variables and other error-free variables available for all subjects in the study are used as covariates in the imputation procedure. With our three-phase MI procedure, we use the telephone interview to impute the missing data into the chart review dataset and then we use this completed dataset among those with chart reviews to further impute data into the EHR dataset.

Specifically, let $X_{1,m}$ and $X_{2,m}$ denote the true, fully validated ART regimen and contraceptive method, respectively, reported during the telephone interview in month $m$. Let $X^*_{1,m}$ and $\tilde{X}_{1,m}$ denote the error-prone versions of $X_{1,m}$, for $j = 1, 2$, which are observed in the EHR (phase-1) and chart review (phase-2), respectively. Let $X_1$ be an error-free covariate or set of covariates, such as age at baseline and study site, which are available for all participants from the EHR. The variable $Y_m$ denotes the fully validated pregnancy status for month $m$ reported in the telephone interview, with $Y^*_m$ and $\tilde{Y}_m$ being the error-prone outcomes observed from the EHR and chart review, respectively. Let $R_1$ be the indicator that a participant was selected for chart review and $R_2$ be the indicator that a participant who was selected for chart review was further selected for telephone interview; let $R = R_1 R_2$. Define $O_m$ as the vector $(Y_m, X_{1,m}, X_{2,m})$, $O^*_m$ as $(Y^*_m, X^*_{1,m}, X^*_{2,m})$, and $\tilde{O}_m$ as $(\tilde{Y}_m, \tilde{X}_{1,m}, \tilde{X}_{2,m})$. Finally, let $Y, X_1, X_2, O, O^*$, and $\tilde{O}$ denote the vectors / matrices formed by stacking $Y_m, X_{1,m}, X_{2,m}, O_m, O^*_m$, and $\tilde{O}_m$, respectively, across all $m$ within a subject. Therefore, $(O^*, X_3, R_1, R_2)$ is observed on all subjects, $O$ is only observed in those with $R_1 = 1$, and $O$ is only observed in those with $R_2 = 1$.

The three-phase MI requires two imputations: imputing fully validated data for those who had chart reviews but no telephone interview (ie, $R_1 (1 - R_2) = 1$) and then imputing fully validated data for those who only had EHR data ($R_1 = 0$).

To complete the first step, we initially need to fit models for $Y_m, X_{1,m}$, and $X_{2,m}$ among those with fully validated data ($R_2 = 1$); these models can be written as $f_y(Y_m | O^*, \tilde{O}, X_3)$, $f_{x_1}(X_{1,m} | O^*, \tilde{O}, X_3, Y)$, and $f_{x_2}(X_{2,m} | O^*, \tilde{O}, X_3, Y, X_1)$, respectively. Note that these models are functions of all error-prone variables observed in both the phase-1 and phase-2 data. As $Y_m$ is a binary variable and $X_{1,m}$ and $X_{2,m}$ are multinomial variables, in our application we used a logistic regression model for $f_y(\cdot)$ and multinomial logistic regression models for $f_{x_1}(\cdot)$ and $f_{x_2}(\cdot)$. When fitting these models, instead of conditioning on data during the entire follow-up period, we only conditioned on data during the previous, current, and subsequent months; specifically,

$$f_y \left( Y_m | O^*, \tilde{O}, X_3 \right) = f_y \left( Y_m | O^*_{m-1}, O^*_m, O^*_{m+1}, \tilde{O}_{m-1}, \tilde{O}_m, \tilde{O}_{m+1}, X_3 \right),$$

$$f_{x_1} \left( X_{1,m} | O^*, \tilde{O}, X_3, Y \right) = f_{x_1} \left( X_{1,m} | O^*_{m-1}, O^*_m, O^*_{m+1}, \tilde{O}_{m-1}, \tilde{O}_m, \tilde{O}_{m+1}, X_3, Y_m, Y_{m-1}, Y_{m+1} \right),$$

and

$$f_{x_2} \left( X_{2,m} | O^*, \tilde{O}, X_3, Y, X_1 \right) = f_{x_2} \left( X_{2,m} | O^*_{m-1}, O^*_m, O^*_{m+1}, \tilde{O}_{m-1}, \tilde{O}_m, \tilde{O}_{m+1}, X_3, Y_m, Y_{m-1}, Y_{m+1}, X_{1,m-1}, X_{1,m}, X_{1,m+1} \right).$$

Complete model specifications for these and all other models are in Appendix A.

Validated data for those in phase-2 but not phase-3 (ie, $R_1 (1 - R_2) = 1$) are then imputed from the fitted models of $f_y(\cdot)$, $f_{x_1}(\cdot)$, and $f_{x_2}(\cdot)$ in a standard manner. First, parameter coefficients are randomly selected from the multivariate normal sampling distribution of the fitted coefficients; and second, $Y_m, X_{1,m}$, and $X_{2,m}$ are imputed from the models using the sampled parameter coefficients. Denote the imputed values as $Y^\text{imp}_m$, $X^\text{imp}_{1,m}$, and $X^\text{imp}_{2,m}$. Note that to impute from $f_{x_2}(\cdot)$, $Y^\text{imp}_j$ is used in place of $Y_j$, and similarly, to impute from $f_{x_1}(\cdot)$, $Y^\text{imp}_j$ and $X^\text{imp}_{1,j}$ are used in place of $Y_j$ and $X_{1,j}$, respectively. Thus, after the imputation we have complete data, denoted $O^\text{imp}_m$ on everyone in phase-2: $O^\text{imp}_m$ for those in phase-3 and $O^\text{imp}_m$ for those in phase-2 but not phase-3.
Similar steps are taken to impute fully validated data for those who only had EHR data (ie, \( R_1 = 0 \)). Specifically, we now fit models for \( Y_{m,1}^{\text{comp}}, X_{1,m}^{\text{comp}} \), and \( X_{2,m}^{\text{comp}} \) among those in phase-2 (\( R_1 = 1 \)); these models can be written as \( g_y(Y_{m}^{\text{comp}}|O^s, X_3) \), \( g_{x_1}(X_{1,m}^{\text{comp}}|O^s, X_3, Y_{m}^{\text{comp}}) \), and \( g_{x_2}(X_{2,m}^{\text{comp}}|O^s, X_3, Y_{m}^{\text{comp}}, X_{1,m}^{\text{comp}}) \), respectively. Again, in our application to the Kenyan study we used logistic and multinomial logistic regression models, as appropriate, and instead of conditioning on data during the entire follow-up period, we only conditioned on data during the previous, current, and subsequent months:

\[
g_y(Y_{m}^{\text{comp}}|O^s, X_3) = f_y(Y_{m}^{\text{comp}}|O_{m-1}^s, O_{m}^s, O_{m+1}^s, X_3),
\]

\[
g_{x_1}(X_{1,m}^{\text{comp}}|O^s, X_3, Y_{m}^{\text{comp}}) = f_{x_1}(X_{1,m}^{\text{comp}}|O_{m-1}^s, O_{m}^s, O_{m+1}^s, X_3, Y_{m}^{\text{comp}}, Y_{m-1}^{\text{comp}}, Y_{m+1}^{\text{comp}}),
\]

and

\[
g_{x_2}(X_{2,m}^{\text{comp}}|O^s, X_3, Y_{m}^{\text{comp}}, X_{1,m}^{\text{comp}}) = f_{x_2}(X_{2,m}^{\text{comp}}|O_{m-1}^s, O_{m}^s, O_{m+1}^s, X_3, Y_{m}^{\text{comp}}, Y_{m-1}^{\text{comp}}, Y_{m+1}^{\text{comp}}, X_{1,m}^{\text{comp}}, X_{1,m+1}^{\text{comp}}).
\]

Validated data for those not in phase-2 (ie, \( R_1 = 0 \)) are then imputed from the fitted models of \( g_y(\cdot) \), \( g_{x_1}(\cdot) \), and \( g_{x_2}(\cdot) \) in a manner similar to that described above. First, parameter coefficients are randomly selected from the multivariate normal sampling distribution of the fitted coefficients; and second, \( Y_m, X_{1,m} \), and \( X_{2,m} \) are imputed from the models using the sampled parameter coefficients. Denote the imputed values as \( Y_{m}^{\text{imp}}, X_{1,m}^{\text{imp}} \), and \( X_{2,m}^{\text{imp}} \). Note that to impute from \( g_y(\cdot) \), \( Y_j^{\text{imp}} \) is used in place of \( Y_j^{\text{comp}} \), and similarly, to impute from \( g_{x_1}(\cdot) \), \( Y_j^{\text{imp}} \) and \( X_{1,j}^{\text{imp}} \) are used in place of \( Y_j^{\text{comp}} \) and \( X_{1,j}^{\text{comp}} \), respectively. Thus, we have complete data, denoted \( O_{m}^{\text{comp}} \) on everyone in phase-1.

Finally, a Poisson regression model with \( X_{1,m}^{\text{comp}}, X_{2,m}^{\text{comp}} \) and their interaction, is used to estimate the impact of efavirenz-containing ART vs nevirapine-containing ART among those on implant contraception on pregnancy, while adjusting for the error-free covariate \( X_3 \) and including follow-up time as an offset. This process of imputing outcome and exposures and analyzing the fully imputed dataset is repeated \( B \) times, leading to a total of \( B \) estimates for the parameter of interest. Final parameter estimates are obtained by averaging them. Because the imputation and analysis models are uncongenial, an estimator for the variance is obtained using a three-phase adaption of the approximation proposed by Robins and Wang and implemented in Giganti and Shepherd.

For purposes of comparison, we also consider a two-phase MI procedure that is similar to that of three-phase MI except that the chart review data are ignored and the imputation is directly from phase-3 data to phase-1. Details of this two-phase MI procedure are written out in the Appendix A.

A few remarks on the three-phase MI estimator: First, notice that the three-phase MI procedure uses all error-prone variables in constructing the complete data and is thus efficient, provided that the imputation and analysis models are correctly specified. Second, there are of course other ways that one could specify the imputation models. For example, machine learning or other less parametric imputation models may result in better model fit and hence, better imputations. With that said, deriving the variance of estimators with uncongenial non-parametric models is not straightforward. In earlier work in a similar setting with uncongenial imputation and analysis models, the application of Rubin’s rules resulted in very conservative confidence intervals (ie, 95% confidence intervals having coverage > 99%). The Robins and Wang approach for calculating the variance of MI estimators with uncongenial models requires information from the imputation model that is often not easily extracted from machine learning methods. In contrast, we know how to compute the variance of estimators using our simpler parametric models, which seemed to fit the data well in our analyses. Third, there are other imputation strategies that could have been employed. For example, another approach might be to impute \( \hat{O} \) based on \( O^s \) for everyone not in phase-2, and then to impute \( O \) based on \( \hat{O} \) and \( O^s \) for everyone not in phase-3 using imputed \( \hat{O} \) for those in phase-1 but not phase-2. Such an approach is reasonable, but does not provide any obvious advantages (or disadvantages) over the imputation strategy we employed. Certainly, there is room for more research in this general area of model selection and inference with multiple imputation. But we turn our attention to another type of estimator.

### 3.2 GR estimators

In this section we extend GR estimators to three-phase studies, so that all error-prone data, from both phase-1 and phase-2, are used in the analysis. Unlike MI, GR does not require discretizing data into monthly intervals and fitting imputation
models, and it can be applied using readily available software. GR estimators are built upon IPW estimators; thus, to fully described them and introduce our new three-phase estimator, we first review IPW and two-phase GR estimators.

The Horvitz-Thompson\(^{27}\) or IPW estimator is a popular technique for handling incomplete data. It works by performing a weighted regression on the fully validated dataset where the weights are the inverse of the probability of being sampled for full validation. The probability for being selected for the chart review and the probability for being further selected for the telephone interview given the chart review data are denoted by \(\pi_{1i}\) and \(\pi_{2i}\), respectively, such that \(\pi_{i} = \pi_{1i}\pi_{2i}\) is the probability of subject \(i = 1, \ldots, n_1\) being selected for full validation. Let \(S_i(\beta) = \partial \log \{Pr(Y_i = 1 | X_{1i}, X_{2i}, X_{3i}; \beta)\} / \partial \beta\) be the score function associated with the regression parameter \(\beta\). The IPW estimator, \(\hat{\beta}_{IPW}\), is obtained by solving the estimating equation \(U_w(\beta) = \sum_i U_w,i(\beta) = \sum_i R_i d_i S_i(\beta) = 0\) for \(\beta\), where \(d_i = 1/\pi_i\) is the design weight for the \(i\)th participant being selected for full validation. The IPW estimator assumes data are missing at random and requires that every subject has a positive probability of being sampled, that is, \(\pi_{1i} > 0\) and \(\pi_{2i} > 0\) for all \(i = 1, \ldots, n_1\). Under these assumptions, the resulting IPW estimator is asymptotically normal with variance estimated by \(I(\hat{\beta}_{IPW})^{-1} \{ \sum_{k,l} U_{w,k}^t (\hat{\beta}_{IPW}) U_{w,l}(\hat{\beta}_{IPW}) \} I(\hat{\beta}_{IPW})^{-1}\), where \(I(\hat{\beta}_{IPW}) = \partial U_w(\hat{\beta}_{IPW}) / \partial \hat{\beta}_{IPW}\).

Although consistent, the IPW estimator is inefficient because it does not make full use of the dataset. It discards information on any subject that has not been fully validated. More efficient methods that make better use of data have been proposed for two-phase sampling, for example, GR estimators, and they can be extended to settings with three-phase sampling to incorporate partially validated data.

GR estimators are robust approaches that build upon IPW.\(^8\) They use auxiliary variables available in all records to calibrate the design weights \(d_i = 1/\pi_i\), so that the weighted total of the auxiliary variable in the validation sample (eg, phase-3 sample) equals the known total of the auxiliary variable in the larger sample (eg, phase-1 sample). This is called the calibration equation and is written \(\sum_i a_i = \sum_i R_i d_i a_i\), where \(a_i\) is the auxiliary variable, \(w_i = g_i d_i\), and \(g_i\) is obtained by minimizing the distance between the calibrated and design weights, \(D(w_i, d_i)\), for some distance function \(D(.)\).

GR estimators of \(\beta\) are more efficient than IPW estimators, while still benefitting from the same robustness properties.\(^{13}\) In particular, GR estimators remain consistent under the same assumptions as IPW estimators regardless of the choice of auxiliary variable used for calibration.\(^{10}\) However, gains in efficiency can be obtained by carefully selecting the auxiliary variable used in the calibration equations. The goal is to find an auxiliary variable that is correlated with the true, unobserved influence function for the parameter of interest: higher correlation leads to higher efficiency. Breslow et al\(^{10}\) showed that the expectation of the influence function obtained from the validation dataset, given error-prone records, is the optimal auxiliary variable. This expectation is typically unknown, but may be well-approximated using the influence function for the parameter of interest based on the unvalidated data.

Using the influence function as auxiliary variables helps connect GR and augmented inverse probability weighting (AIPW) estimators. AIPW builds upon IPW by adding a weighted average of the outcome model into the estimating equation,\(^{11}\) leading to an efficient and doubly-robust estimator. That is, the AIPW estimator is consistent if either the probability weights or the outcome model are correctly specified; if both are correctly specified, the AIPW estimator is semiparametrically efficient.\(^{11}\) At first glance, AIPW and GR estimators seem very different: while AIPW focuses on estimating the mean via regression modeling, GR estimators were first derived among survey statisticians as a regression estimator for a population total.\(^{39}\) However, if we instead set up the calibration problem to estimate the population total of the influence functions of regression parameters, then one can show that GR and AIPW estimators are asymptotically equivalent.\(^{12}\) GR estimators have, therefore, the same desirable and well-known properties of AIPW estimators: high statistical efficiency and double-robustness. This is particularly appealing when the data are missing by design, as in our motivating example.

When applied to three-phase studies, GR estimators as just described will lead to inefficient estimates. This is because GR estimators calibrate the design weights \(d_i\) using information from the EHR, ignoring the chart review data. We extend GR estimators to fully use all error-prone data, leading to more efficient estimators. We focus on a single component of \(\beta\), say \(\beta_k\). The auxiliary variables are the efficient influence functions associated with this parameter. Let \(d_{1i} = 1/\pi_{1i}\) and \(d_{2i} = 1/\pi_{2i}\) be the design weights for selecting the \(i\)th individual for chart review and further telephone interview, respectively. Similarly, let \(w_{1i} = g_{1i} d_{1i}\) and \(w_{2i} = g_{2i} d_{2i}\) denote the calibrated weights. The quantity \(g_{ij}\), \(j = 1, 2\), is obtained by minimizing the distance between the calibrated and design weights under the revised constraints (i) \(a_{i\text{total}}^* = \sum_i a_i^* = \sum_i R_{1i} w_{1i} a_i^*\), which calibrates the design weights using the auxiliary variable observed in the EHR; and (ii) \(\tilde{a}_{i\text{total}}^* = \sum_i R_{1i} \tilde{a}_{i} = \sum_i R_{1i} w_{2i} \tilde{a}_{i}\), which calibrates the weights using the auxiliary variables observed in the chart review. The auxiliary variables \(a_i^*\) and \(\tilde{a}_{i}\) correspond to the influence functions obtained from the error-prone EHR.
and chart review datasets, respectively. The efficient influence functions for \( \beta \) are unknown, but a natural choice is to use the observed, error-prone values. Specifically, we use the influence function derived from the EHR and chart review datasets to approximate the true ones. For the Poisson model used in our analyses, these influence functions are 
\[
\alpha(X, Y; \beta) = \left\{ X^t \mu(X, \beta) X \right\}^{-1} X^t \{ Y - \mu(X, \beta) \},
\]
where \( X = (X_1, X_2, X_3) \) and \( \mu(X, \beta) = \exp \{ X \beta + \log(\text{offset}) \} \). Thus, \( a^* = \alpha(X, Y^*; \beta^*) \) and \( \tilde{a} = \alpha(\tilde{X}, \tilde{Y}; \beta) \), where \( \beta^* \) is the estimate of \( \beta \) that uses only error-prone data and \( \beta \) is the estimate of \( \beta \) using only chart review data, but properly accounting for the sampling probabilities via IPW.

There are several types of distance functions, \( D(w, d) \), each of which gives rise to a different calibration estimator.\(^8\) Consider, for example, the distance function \( D(w_i, d_i) = (w_i - d_i)^2 / 2d_i \). Using Lagrange multipliers \( \lambda_1 \) and \( \lambda_2 \) to minimize \( \sum_k D(w_{ki}, d_{ki}) \), for \( k = 1, 2 \), under constraints (1) and (2), we have \( g_{1i} = 1 - \lambda_1^2 a_i \) and \( g_{2i} = 1 - \lambda_2^2 \bar{a}_i \), where \( \lambda_1 = (\sum_{i=1}^n R_{i1} d_{i1} a_i a_i^t)^{-1} (\sum_{i=1}^n R_{i1} d_{i1} a_i - a^*_{\text{total}}) \) and \( \lambda_2 = (\sum_{i=1}^n R_{i2} d_{i2} \bar{a}_i)^{-1} (\sum_{i=1}^n R_{i2} d_{i2} \bar{a}_i - \bar{a}_{\text{total}}) \).

GR estimators are obtained by solving \( U_r(\beta) = \sum_{i=1}^n U_{ri}(\beta) = \sum_{i=1}^n R_{wi} w_{2i} S_i(\beta) = 0 \) with respect to \( \beta \). The variance of the resulting GR estimator, \( \hat{\beta}_r \), can be obtained via linearization.\(^13\) Applying the delta method to \( \sum_i U_r(\hat{\beta}_r) = 0 \), we have that \( \text{Var}(\hat{\beta}_r) \approx I(\hat{\beta}_r)^{-1} \{ \sum_{k=1}^n U_{rk}(\hat{\beta}_r, \hat{\beta}_r) \} I(\hat{\beta}_r)^{-1} \), where \( I(\hat{\beta}_r) = \partial U_r(\hat{\beta}_r) / \partial \hat{\beta}_r \). Asymptotic consistency and normality of this GR estimator follow from van der Vaart.\(^28\)

It is important to notice that, even though we used the distance function \( D(w, d) = (w - d)^2 / 2d \) to reach a closed-form expression for the variance of \( \hat{\beta}_r \), the asymptotic distribution of \( \hat{\beta}_r \) is independent of the choice of \( D(\cdot) \).\(^8,10\) Also, since \( D(w, d) = (w - d)^2 / 2d \) may lead to negative weights, making its interpretation harder, for the remainder of this paper we will use the Poisson deviance \( D(w, d) = w \{ \log(w) - \log(d) \} + (d - w) \). This distance function guarantees nonnegative weights and leads to an estimator that is equivalent to the classical raking adjustment in the special case of discrete auxiliary variables.\(^12\)

### 3.3 GR with multiple imputation

Here we introduce a final estimation approach that combines the GR and MI techniques discussed thus far. Recall that the two-phase GR estimator calibrates the product of the sampling weights using the influence functions obtained from the phase-1 error-prone data, while the three-phase GR estimator uses the product of the calibrated weights obtained by calibrating each sampling weight individually using the influence functions obtained in the previous phases. Both the two-phase and three-phase estimators use the error-prone influence functions. The error-prone influence functions are good auxiliary variables if they are highly correlated to the true, unknown influence functions, but depending on the degree of errors in the phase-1 data, they can also be poor auxiliary variables.

A possible way to improve the correlation between the influence functions used as auxiliary variables and the true influence functions is via imputation. Note that the error-prone influence functions used in GR are obtained at phase-1 and do not make any use of the validated data. Instead, one could potentially use the validated data to multiply impute influence functions that are hopefully closer (ie, high correlation) to the true influence functions. This can be done by combining GR and MI as follows. Instead of using the error-prone phase-1 and phase-2 data to calculate the influence functions, we can follow the imputation steps discussed in Section 3.1 to create imputed datasets at the phase-1 and phase-2 levels. With the imputed datasets we then compute the influence functions associated with the parameters of interest. If we impute \( B \) datasets, we will have \( B \) (imputed) influence functions. For each subject, we can take the average influence functions across all \( B \) imputations, and these average (imputed) influence functions, which make use of the validated data, are expected to be closer to the true influence functions than those based on the error-prone data. These average (imputed) influence functions are then used as auxiliary variables in GR. Note that for this procedure we use fully imputed data, ie, we even impute \( O_i \) for those with \( R_i = 1 \) and we use these imputed values of \( O_i \).

The rationale is that by using a multiply imputed influence function we can estimate the expected value of the desired influence function given the observed data, leading to auxiliary variables that are closer to the optimal ones and thus in theory resulting in estimators with smaller variance than our GR estimators while being more robust to model misspecification than our MI estimators.\(^21,22\) An incorrect imputation model will result in suboptimal efficiency (although typically still more efficient than IPW) but the estimator will remain consistent.\(^10\) So the advantage of this approach is potential gains in efficiency over GR using the error-prone influence function. The disadvantage is that this approach requires building an imputation model, thus adding analytical and computational complexity.
4 | SIMULATIONS

4.1 | Data generation

We studied the performance of our MI and GR estimators with simulated data. The simulated data roughly followed our motivating example, with a binary error-prone outcome $Y$, a time-varying error-prone exposure $X_1$, and a time-varying error-free covariate $X_2$. Data were generated for $i = 1, \ldots , n_1 = 15000$ subjects, for a follow-up of up to 18 months. We simulated three different values for $X_2$ over the entire follow-up, with constant periods of up to 6 months, while $X_1$ varied monthly. Data for the $i$th patient can be written as $(Y_{ijk}, X_{1,ijk}, X_{2,ijk})$, where $j = 1, 2, 3$ denotes the periods in which $X_2$ varied and $k = 1, \ldots , 6$ denotes the period of time (months) in which $X_1$ varied. The length of periods $j = 1, 2, 3$ were defined as the min$(t, 6)$, where $t$ is the time to event (eg, pregnancy). The rationale was the following: if an event occurred at that time, for example, if the $i$th woman becomes pregnant during the fourth follow-up month (so that $Y_{i14} = 1$), her follow-up stops and only returns when she is no longer pregnant. At this stage her $X_2$ status may have changed, so a new value is assigned to her. If we further assume that this woman did not get pregnant again during her follow-up, her data are written as $(Y_i, X_{1i}, X_{2i})$, where $Y_i = (Y_{i11}, \ldots , Y_{i14}, Y_{i21}, \ldots , Y_{i26}, Y_{i31}, \ldots , Y_{i36})$, $X_1 = (X_{1i11}, \ldots , X_{1i14}, X_{1i21}, \ldots , X_{1i26}, X_{1i31}, \ldots , X_{1i36})$ and $X_2 = (X_{2i1}, X_{2i2}, X_{2i3})$. The time to event $T$ was generated from a Weibull distribution with shape parameter equal to 0.5 and scale equal to 2.5 $\times$ $10^{-3} \exp \{- (Y_{1i,ijk} + Y_{2i,ijk})\}$. The outcome $Y_{ijk}$ was defined as 0 if no event occurred or 1 otherwise. Values of the time-varying exposure and covariate $X_{1,ijk}$ and $X_{2,ij}$, respectively, were independently drawn from a standard normal distribution. We set $g_1 = g_2 = 0.5$, leading to a prevalence for $Y$ of approximately 5%.

Both the outcome $Y_{ijk}$ and covariate $X_{1,ijk}$ were assumed to be measured with errors, with $(Y_{ijk}^*, X_{1,ijk}^*)$ and $(\tilde{Y}_{ijk}, \tilde{X}_{1,ijk})$ denoting the observed values of $(Y_{ijk}, X_{1,ijk})$ at phase-1 and phase-2, respectively. We assumed that the correctly recorded variables $(Y, X)$ were only available at phase-3. The error-prone variables $Y_{ijk}^*$ and $\tilde{Y}_{ijk}$ were generated from separate Bernoulli distributions $\Pr(Y_{ijk}^* = 1|Y_{ijk}) = \exp\{-5 + 4.5Y_{ijk} + X_{2,ijk}\}$ and $\Pr(\tilde{Y}_{ijk} = 1|Y_{ijk}) = \exp\{-7 + 6.75Y_{ijk} + 0.5X_{2,ijk}\}$. That way, both $Y_{ijk}^*$ and $\tilde{Y}_{ijk}$ will correctly classify a case in about 61% and 78% of times, while $Y_{ijk}$ and $\tilde{Y}_{ijk}$ will incorrectly identify a case in about less than 1% of times. The error-prone variables $X_{1,ijk}^*$ and $\tilde{X}_{1,ijk}$ were equal to $X_{1,ijk} = X_{1,ijk} + 0.25Y_{ijk} + 0.1X_{2,ijk} + U^*$ and $\tilde{X}_{1,ijk} = X_{1,ijk} + 0.25Y_{ijk} + 0.1X_{2,ijk} \tilde{U}$, where $U^*$ and $\tilde{U}$ were generated from mean zero normal distributions with SDs 1 and 0.75, respectively. Notice that this generates data such that the chart-reviewed records $(\tilde{Y}, \tilde{X}_1)$ are generally closer to the truth than the EHR-data $(Y^*, X_1^*)$.

Let $E^*$ denote 1 of the groups of patients that had at least 1 event recorded in their EHR (phase-1) during the follow-up period and $E^* = 0$ otherwise; $\tilde{E}$ represents the same in the chart review (phase-2). A total of $n_2 = 1500$ subjects were randomly sampled for chart review from each group, defined by $E^*$, wherein $\tilde{Y}$ and $\tilde{X}$ were obtained. A subsample of size $n_3$ were then sampled from each strata into phase-3 (telephone interview) for further validation. We considered two simulation settings for phase-3: (1) sampling equally from strata defined by $E^*$ alone, or (2) sampling equally from the strata defined by the combinations of $(E^*, \tilde{E})$. Notice that simulation setting (1) ignores variables validated at phase-2, using only those observed at phase-1. We ran 10000 Monte Carlo simulations with $B = 50$ imputations for the MI approaches. Estimates were obtained by fitting the logistic model $\logit(\Pr(Y_{ijk} = 1|X_{1,ijk}, X_{2,ij})) = \alpha + \beta_1 X_{1,ijk} + \beta_2 X_{2,ij}$ on the imputed data for all MI procedures or on the weighted fully validated data for all design-based estimators. With rare events and short follow-up periods, $\beta \approx \gamma$, so the estimates obtained from the logistic model can be used for inference.

We computed the empirical bias, variance, and mean squared error (MSE) for the following methods: IPW, two-phase and three-phase GR estimators, two- and three-phase MI, and two- and three-phase combinations of GR and MI. For the GR estimators, we used the influence functions associated with the two parameters of interest and their interaction with the stratification variable $E^*$ or $\tilde{E}$. Since each patient contributed more than one row in the dataset, in both the simulated and real data analysis, influence functions for parameters will vary by row within patients. This means that the sampling weights are calibrated by row, so that each unique row for a patient may have varying calibrated weights. An alternative, although less efficient approach is to summarize influence functions for each patient (eg, total or mean influence function across rows), so that each patient has a single influence function for each parameter, and then to use these influence functions as auxiliary variables. Although we did not use this latter approach, it could be useful if one wants to guarantee constant calibrated weights across rows within the same patient.
4.2 | Results

Figure 2 shows the MSE for all methods for the two simulation settings and varied phase-3 sample sizes, $n_3 = (250, 500, 750)$ from each strata. The results for empirical bias and variance are presented in Tables S1 and S2, respectively, in the Appendix B. Consider first simulation setting (1) (Figure 2A), where phase-3 sampling depended only on $E^*$. All estimators were approximately unbiased so differences in MSE reflect differences in the variance of the estimators. As expected, the IPW estimator was least efficient, followed by the two-phase GR estimators that ignored the partially validated data. The three-phase GR estimators were substantially more efficient than the regular two-phase estimators. These gains in efficiency were more pronounced with smaller numbers of patients selected for phase-3, leading to estimates for $\beta_1$ that were about 20% more efficient than the two-phase GR estimators and 30% more efficient than the IPW estimators. For both two-phase and three-phase GR estimators, the MSE was slightly smaller when weights were calibrated with the multiply imputed influence function versus the naive influence function. Interestingly, the three-phase GR estimator was more efficient than the two-phase MI when estimating $\beta_2$ and it had comparable performance when
estimating $\beta_1$. This highlights the potential gains in efficiency by considering all stages of sampling. Finally, and not surprisingly, the three-phase MI was the most efficient estimator.

Results were similar under simulation setting (2) (Figure 2B), when the sampling probability depended on $(E^*, \tilde{E})$, except for the two-phase MI estimators. The two-phase MI estimator completely ignored the phase-2 data so its missing at random assumption did not hold, resulting in a biased estimator. The three-phase MI procedure, on the other hand, correctly accounted for all sampling stages, so the final estimates were unbiased and again more efficient than all other estimators.

We performed an additional set of simulations, simulation setting 3, with higher phase-1 error rates. Specifically, the error-prone variable $\tilde{Y}_{ijk}$ was generated from a Bernoulli distribution with $\Pr(\tilde{Y}_{ijk} = 1|Y_{ijk}) = \exp[-8 + 7.75Y_{ijk}]$, so that the phase-2 data is even close to truth. Similarly, we reduced the standard error of $\tilde{X}_{1,ijk}$ from 0.75 to 0.1. Subjects were selected for phase-2 and phase-3 based only on $E^*$, as in sampling scenario 1) above. Results are displayed in Figure 3, and Tables S1 and S2 of Appendix B. The two-phase GR and MI estimators were greatly affected by this higher error rate because the unvalidated EHR data were much less correlated with the truth and contained little information to calibrate or impute. In contrast, the three-phase GR and MI estimators resulted in efficiency gains because they effectively incorporated the phase-2 chart review data which was highly correlated with the truth.

In the previous settings, all models were correctly specified, so that when the missing at random assumption was satisfied, the three-phase MI estimators were most efficient. We performed an additional simulation (simulation setting 4) to investigate the impact of a misspecified imputation model. Data were generated as in simulation simulation settings (1), except that $X_{1,ijk} = \log(Z_{ijk} + 0.5X_{2,ijk} + 0.13X_{2,ijk}^2)$ with $Z_{ijk}$ generated from a gamma distribution with shape equal to 10 and rate equal to 1, and with $\Pr(Y^*_{ijk} = 1|Y_{ijk}) = \exp[-4 + 3.5Y_{ijk} + X_{2,ijk}]$ and $\Pr(\tilde{Y}_{ijk} = 1|Y_{ijk}) = \exp[-5 + 4.75Y_{ijk} + 0.5X_{2,ijk}]$. However, a normal linear model as described in Section 4.1, without the quadratic term, was still used to impute the fully validated $X_1$ into phase-2 and phase-1. Details for the imputation model are given in Appendix C. The remaining steps followed as before, and results are displayed in Figure 4. Both MI estimators were strongly affected by misspecification of the imputation model, leading to large bias and thus large MSE. The GR estimators combined with MI were only slightly affected, if at all, by model misspecification. The GR estimators yielded the smallest MSE when estimating $\beta_1$.

Finally, it is important to highlight that depending on the degree of model missclassification and sample size, the phase-3 data may end up with very few events or with co-linear variables. Although it would not affect IPW nor GR (assuming that we have enough events to fit the main outcome model), it may affect the MI approach, as the imputation model may become unstable, with high variability. This highlights the importance of model checking in the imputation.
step. In the Data S1 we have included simulations with simpler missclassification or error models. The results are generally similar to those presented here, but illustrate that two-phase MI may be more efficient than three-phase GR in some settings.

5 | REVISITING THE PREGNANCY STUDY

We next reevaluated the pregnancy dataset from Patel et al\(^3\) described in Section 2, but now applying our methods to account for both validation stages. Recall that the objective was to verify whether efavirenz-containing ART decreased the effectiveness of contraceptive implants among women with HIV. The EHR contained data from \(n_1 = 83,671\) women living with HIV, of whom \(n_2 = 4732\) had their medical charts reviewed to validate their pregnancy status, ART regimen, and contraceptive type. A total of \(n_3 = 1210\) women from the chart review were also contacted by telephone to have the same variables further validated. The IRR for the effectiveness of efavirenz-containing ART with contraceptive implant, compared to nevirapine-containing ART and contraceptive implant as the reference level, was calculated via a Poisson regression model; pregnancy status was set as the outcome, with contraceptive type, ART regimen, as well as their interactions, as exposures of interest, while further adjusting for age at baseline and study site.

IRRs obtained for all estimators are presented in Table 1. Both two- and three-phase MI estimators used \(B = 50\) imputations. We present point estimates and the corresponding 95% confidence intervals. In general, nearly all methods led to point estimates that were very similar to each other with respect to the parameter of interest. Conclusions were, in general, similar to the naïve analysis discussed in Section 2. All methods suggested that efavirenz is indeed associated with a reduction in efficacy of the contraceptive implant, leading to higher risks of becoming pregnant. The IRR ranged from 3.06 to 4.09 for all methods, except for three-phase MI. The three-phase MI estimator showed a substantially lower estimated IRR, 1.81. This smaller risk, compared to all other estimators, may be due to misspecification of the imputation models. Recall that the three-phase MI imputes missing data in two steps: first into the chart review data and later into the EHR. If any of the models is incorrectly specified, the final estimates may be biased, as shown in the simulations.

In terms of efficiency, two-phase GR resulted in slightly smaller variance for the parameter of interest when compared to IPW, but greater variance than the two-phase raking with MI and two-phase MI only. The methods that were constructed to use all three phases of data (EHR, chart, telephone interview) led to smaller variances compared to their two-phase counterparts. This highlights the importance of including the extra information available in the chart review. There were also some gains in efficiency by calibrating weights using the multiply imputed influence function (ie, GR+MI)
rather than the error-prone influence function (ie, GR) with two- and three-phase GR. Similar to our simulations, the fully parametric MI estimators had smaller variance than the semiparametric GR and GR+MI estimators.

6 | CONCLUSION

In this paper, we introduced methods for the analysis of three-phase validation studies that efficiently use error-prone, partially validated, and fully validated data: a time-discretized multiple imputation approach, a GR approach, and an estimator that combines the two methods. Via simulations, we illustrated the substantial efficiency gains of our proposed methods over existing methods that discard the intermediate, partially validated phase-2 data. These methods were also used to re-analyze a large three-phase validation study of women living with HIV from western Kenya. Results were generally consistent across methods, but the estimated variances for the parameter of interest were much smaller using our new methods that incorporated data from all three phases into the analyses.

Under correct modeling assumptions, our three-phase MI estimators were more efficient than our three-phase GR estimators. However, three-phase MI estimators require correct specification of the imputation models, which may be challenging in practice, to be unbiased. As mentioned in Section 3.1, more sophisticated imputation models (eg, machine learning approaches) can be employed, but it is often difficult to properly quantify uncertainty after fitting these types of models. Calculating proper confidence intervals in settings with ungenial imputation and analysis models is fairly complicated even after fitting parametric imputation models.\textsuperscript{25,26} The use of machine learning methods with multiple imputation in two-/three-phases is an interesting potential area for future research. If proper model specification feels unattainable, the less-efficient three-phase GR estimator may be preferable because it is consistent under fewer assumptions. The GR estimator is also quite easy to implement across different settings using the \texttt{survey} package\textsuperscript{30} in R software.\textsuperscript{31} The GR estimator can also be obtained using the \texttt{svycal} command in Stata\textsuperscript{32} or a SAS macro developed by An.\textsuperscript{33} We chose to use the three-phase GR estimator for the analyses reported in Patel et al\textsuperscript{8} because of its fewer assumptions and its implementation simplicity. With that said, we recognize that there may be settings when the efficiency loss of the three-phase GR estimator may favor the use of the three-phase MI estimator. We also believe it can be appropriate, and often preferable, to show results using both GR and MI estimators.

Our three-phase estimator that combined GR and MI—more specifically, that used the average of multiply imputed influence functions to calibrate weights for the three-phase GR procedure—tended to have efficiency that was slightly better than our three-phase GR estimator that calibrated weights using naive influence functions. These results are generally consistent with what we have seen in two-phase analysis settings,\textsuperscript{15,16,34} and they suggest that if one is applying a three-phase GR estimator, it might not be worth the added effort to also fit imputation models. With that said, unlike the three-phase MI estimator, the three-phase GR plus MI estimator does not require correct specification of the imputation model to be unbiased and variance estimation is straightforward, so fitting the three-phase GR plus MI estimator may be less daunting than fitting the three-phase MI estimator. And if one has already fit a three-phase MI estimator and plans to also fit a three-phase GR estimator, one might as well fit the three-phase GR+MI procedure as it involves almost no additional effort and has the potential to improve efficiency over the three-phase GR estimator.

It is important to highlight limitations in the motivating example. Due to differences in terms of follow-up periods and timing of events and exposures between the phase-1 EHR data and both the phase-2 and phase-3 validation datasets, we needed to make some simplifications to employ our methods. In particular, we restricted the follow-up time for each patient to be the intersection of their follow-up times across all three datasets, which resulted in some loss of information. In addition, in the MI approach we discretized the follow-up into monthly intervals. Since data were collected at irregular intervals, this coarsening of time potentially resulted in information loss, although prior two-phase studies that have employed similar discretization have seen little information loss.\textsuperscript{14} Discretizing follow-up time also required sometimes carrying forward values from previous months. For the key variables in our study (ie, ART use, contraceptive use, and pregnancy status), carrying forward values was reasonable; however, with other variables (eg, laboratory measurements), such an approach might be less reasonable. Patients selected for telephone interview were sampled based on convenience; therefore, the missing at random assumption made in all of our analyses may have been violated. Finally, throughout we implicitly assumed that the fully validated data after the telephone interview are correct; this may not be the case, and it is possible that for some patients/variables, the EHR and/or chart review data may be correct but different from the telephone interview data.

With two-phase studies, the choice of which records to validate can have a big impact on the efficiency of results.\textsuperscript{35,36} Three-phase studies are often conducted via stratified random sampling, as in our motivating example. However, it may
be useful to derive optimal three-phase sampling designs that target the parameter of interest. Multi-wave designs may be warranted, as the optimal design may depend on parameters that are unknown without preliminary validation data. We are currently investigating designs of this nature.

R code can be found at https://github.com/gustavodecastro/Three_Phase_Analysis.

ACKNOWLEDGMENTS
This research was funded by the U.S. National Institutes of Health grants R01AI131771, U01AI069911, and K23AI120855, and by the Patient-Centered Outcomes Research Institute grant R-1609-36207. The authors would like to thank investigators in the East Africa IeDEA Consortium.

DATA AVAILABILITY STATEMENT
The data that support the findings in this paper are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID
Gustavo Amorim https://orcid.org/0000-0002-2941-5360
Ran Tao https://orcid.org/0000-0002-1106-2923
Sarah Lotspeich https://orcid.org/0000-0001-5380-2427
Pamela A. Shaw https://orcid.org/0000-0003-1883-8410
Rena C. Patel https://orcid.org/0000-0001-9893-5856

REFERENCES
1. Patel RC, Onono M, Gandhi M, et al. Pregnancy rates in HIV-positive women using contraceptives and efavirenz-based or nevirapine-based antiretroviral therapy in Kenya: a retrospective cohort study. The Lancet HIV. 2015;2:e474-e482.
2. Chappell CA, Lamorde M, Nakalema S, et al. Efavirenz decreases etonogestrel exposure: a pharmacokinetic evaluation of implantable contraception with antiretroviral therapy. AIDS. 2017;31:1965-1972.
3. Patel RC, Amorim G, Jakait B, et al. Pregnancies among women living with hiv using contraceptives and antiretroviral therapy in western Kenya: a retrospective, cohort study. BMC Med. 2021;19:1-11.
4. White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol. 1982;115:119-128.
5. Lawless JF, Kalbfleisch JD, Wild CJ. Semiparametric methods for response-selective and missing data problems in regression. J R Stat Soc Ser B Stat Methodology. 1999;61:413-438.
6. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement Error in Nonlinear Models: a Modern Perspective. New York: Chapman and Hall/CRC; 2006.
7. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006;35:1074-1081.
8. Deville J-C, Särndal C-E. Calibration estimators in survey sampling. J Am Stat Assoc. 1992;87:376-382.
9. Särndal C-E, Swensson B, Wretman J. Model Assisted Survey Sampling. New York: Springer; 2003.
10. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Improved Horvitz-Thompson estimation of model parameters from two-phase stratified samples: applications in epidemiology. Stat Biosci. 2009;1:32-49.
11. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89:846-866.
12. Lumley T, Shaw PA, Dai JY. Connections between survey calibration estimators and semiparametric models for incomplete data. Int Stat Rev. 2011;79:200-220.
13. Lumley T. Complex Surveys: A Guide to Analysis Using R. Hoboken, NJ: John Wiley & Sons, Inc; 2010.
14. Giganti MJ, Shaw PA, Chen G, et al. Accounting for dependent errors in predictors and time-to-event outcomes using electronic health records, validation samples, and multiple imputation. Ann Appl Stat. 2020;14:1045.
15. Oh EJ, Shepherd BE, Lumley T, Shaw PA. Raking and regression calibration: methods to address bias from correlated covariate and time-to-event error. Stat Med. 2021;40:631-649.
16. Shepherd BE, Han K, Chen T, et al. Multiwave validation sampling for error-prone electronic health records. Biometrics. 2023;79:2649-2663.
17. Whittemore AS, Halpern J. Multi-stage sampling in genetic epidemiology. Stat Med. 1997;16:153-167.
18. Pfeiffer RM, Pee D, Landi MT. On combining family and case-control studies. Gene Epidemiol. 2008;32:638-646.
19. Breslow NE, Amorim G, Pettinger MB, Rossouw J. Using the whole cohort in the analysis of case-control data. Stat Biosci. 2013;5:232-249.
20. Zhou H, Elliott MR, Raghunathan TE. Multiple imputation in two-stage cluster samples using the weighted finite population bayesian bootstrap. J Surv Stat Methodol. 2016;4:139-170.
21. Han P. Combining inverse probability weighting and multiple imputation to improve robustness of estimation. Scand J Stat. 2016;43:246-260.
22. Han K, Shaw PA, Lumley T. Combining multiple imputation with raking of weights: An efficient and robust approach in the setting of nearly true models. Stat Med. 2021;40:6777-6791.
23. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. Vol 81. Hoboken, NJ: John Wiley & Sons, Inc; 2004.
24. Meng X-L. Multiple-imputation inferences with unconnected sources of input. *Stat Sci*. 1994;9:538-558. https://doi.org/10.1214/ss/1177010269
25. Robins JM, Wang N. Inference for imputation estimators. *Biometrika*. 2000;87:113-124.
26. Giganti MJ, Shepherd BE. Multiple-imputation variance estimation in studies with missing or misclassified inclusion criteria. *Am J Epidemiol*. 2020;189:1628-1632.
27. Horvitz DG, Thompson JD. A generalization of sampling without replacement from a finite universe. *J Am Stat Assoc*. 1952;47:663-685.
28. van der Vaart AW. Cambridge: Cambridge University Press; 2000.
29. Ngwa JS, Cabral HJ, Cheng DM, et al. A comparison of time dependent cox regression, pooled logistic regression and cross sectional pooling with simulations and an application to the Framingham Heart Study. *BMC Med Res Methodol*. 2016;16:1-12.
30. Lumley T. Survey: analysis of complex survey samples. R Package Version 4.0. 2020.
31. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2021 https://www.R-project.org/
32. Valliant R, Dever JA. *Survey Weights: A Step-by-Step Guide to Calculation*. TX: Stata Press; 2018.
33. An T. A SAS® macro for calibration of survey weights. 2020 https://support.sas.com/resources/papers/proceedings20/4284-2020.pdf. Accessed July 2023.
34. Oh EJ, Shepherd BE, Lumley T, Shaw PA. Improved generalized raking estimators to address dependent covariate and failure-time outcome error. *Biom J*. 2021;63:1006-1027.
35. McIsaac MA, Cook RJ. Response-dependent two-phase sampling designs for biomarker studies. *Can J Stat*. 2014;42:268-284.
36. Amorim G, Tao R, Lotspeich S, Shaw PA, Lumley T, Shepherd BE. Two-phase sampling designs for data validation in settings with covariate measurement error and continuous outcome. *J R Stat Soc A Stat Soc*. 2021;184:1368-1389.
37. McIsaac MA, Cook RJ. Adaptive sampling in two-phase designs: a biomarker study for progression in arthritis. *Stat Med*. 2015;34:2899-2912.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Amorim G, Tao R, Lotspeich S, et al. Three-phase generalized raking and multiple imputation estimators to address error-prone data. *Statistics in Medicine*. 2024;43(2):379-394. doi: 10.1002/sim.9967