Annals of Epidemiology 38 (2019) 57–64

Contents lists available at ScienceDirect

Annals of Epidemiology

Original article

The missing indicator approach for censored covariates subject to limit of detection in logistic regression models

Sy Han Chiou, PhD [a], Rebecca A. Betensky, PhD [a], Raji Balasubramanian, ScD [b, *]

[a] Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA
[b] Department of Biostatistics and Epidemiology, University of Massachusetts – Amherst, Amherst, MA

ARTICLE INFO

Article history:
Received 20 July 2018
Accepted 24 July 2019
Available online 13 August 2019

Keywords:
Limit of detection
Logistics regression
Matched design
Metabolomics

ABSTRACT

Purpose: In several biomedical studies, one or more exposures of interest may be subject to nonrandom missingness because of the failure of the measurement assay at levels below its limit of detection. This issue is commonly encountered in studies of the metabolome using tandem mass spectrometry–based technologies. Owing to a large number of metabolites measured in these studies, preserving statistical power is of utmost interest. In this article, we evaluate the small sample properties of the missing indicator approach in logistic and conditional logistic regression models.
Methods: For nested case-control or matched case control study designs, we evaluate the bias, power, and type I error associated with the missing indicator method using simulation. We compare the missing indicator approach to complete case analysis and several imputation approaches.
Results: We show that under a variety of settings, the missing indicator approach outperforms complete case analysis and other imputation approaches with regard to bias, mean squared error, and power.
Conclusions: For nested case-control and matched study designs of modest sample sizes, the missing indicator model minimizes loss of information and thus provides an attractive alternative to the oft-used complete case analysis and other imputation approaches.

We consider the setting in which one or more covariates of interest may be subject to a limit of detection associated with the measurement assay. This issue arises in high-throughput `omics' technologies, such as metabolomics that involve the measurement of several hundred metabolites per specimen. Several methods are commonly used for handling covariates that are subject to limit of detection. These include complete case (CC) analysis, models including a missing indicator, ad hoc substitution methods, parametric, and likelihood- and imputation-based approaches [1–3]. Here, in the presence of a binary outcome, we compare the performance of the simple missing indicator model [4–6] to CC analysis and other imputation approaches to handle covariates with nonrandom missingness and show that this approach provides an attractive alternative that is especially useful in high-dimensional data settings.

Metabolomic technologies are characterized by detection limits that affect the measurement of low abundance metabolites. In many clinical applications, low abundance metabolites are also likely to be implicated in disease. One widely used treatment of censored covariates is to discard subjects who have censored covariates or the CC analysis. This approach is inefficient under moderate to heavy censoring. Ad hoc substitution methods, likelihood-based methods under the assumption of a parametric distribution for the covariate, and several imputation techniques have been proposed [1–3,7–14]. Parametric and imputation approaches offer considerable improvements but are computationally intensive and/or require stringent assumptions. A recently published comprehensive article compared 31 imputation frameworks for handling missing values in metabolomics data, in which they concluded that multiple imputation using predictive mean matching and K-nearest neighbors had optimal performance [15]. Additional studies have considered linear and time-to-event regression and have obtained similar findings [16–21]. However, to the best of our knowledge, none of these works have evaluated the performance of missing indicator approach to handle covariates subject to nonrandom missingness due to the limit of detection of the measurement technique.

A simple approach for handling missing covariates is the missing data indicator (MDI) model, in which an indicator variable for whether the explanatory variable is observed is included as a covariate in the model, along with the continuous measurements of

No conflicts of interests to declare.

* Corresponding Author: 405 Arnold House, University of Massachusetts – Amherst, Amherst, MA 01003. Tel.: 413-577-0277.
E-mail address: rbalasub@umass.edu (R. Balasubramanian).

https://doi.org/10.1016/j.annepidem.2019.07.014
1047-2797/

the explanatory variable of interest for those for whom it is observed [4–6]. The theoretical properties of the missing indicator model were examined in the context of linear regression for continuous outcomes under various mechanisms of missingness, including when missingness depends on the true value of the covariate as in settings affected by limit of detection [4]. In a MDI model including a completely observed covariate and a censored covariate, the corresponding regression coefficients are unbiased when the covariates are uncorrelated [4]. However, in general, the asymptotic bias of the regression coefficient associated with the censored covariate increases with increasing magnitude of the correlation between the two covariates and the proportion of the censored covariate that falls below the limit of detection. Moreover, these theoretical results do not apply directly to small sample settings, to models with binary outcomes, or to matched studies [4].

Here, in a numerical analysis, we extend the results presented by Jones, M. P. (1996) [4] to the analysis of binary outcomes with a focus on studies of modest sample size ($n < 400$) when missing values of covariates are due to a nonrandom process resulting from the limit of detection of the measurement technique. Through simulations, we compare the MDI approach to a CC analysis and other imputation approaches (i.e. substitution, imputation using predictive mean matching, iterative random forests–based imputation) previously studied in logistic regression models. For each method, we evaluate the bias and mean squared error (MSE) associated with the estimation of the regression parameter of interest and power/type 1 error associated with the corresponding hypothesis tests. We also consider the setting of matched studies, in which a conditional logistic regression is used. We apply the MDI and CC approaches to a cardiovascular disease biomarker study and compare the results. These results could provide useful guidance to investigators involved in the analyses of covariate sets that may be subject to missingness due to limit of detection associated with the measurement technology.

Methods

Let Y denote the outcome of interest and $\tilde{\boldsymbol{X}}$ denote a p-dimensional covariate vector, where each component of $\tilde{\boldsymbol{X}}$ is subject to a different level of left-censoring due to limit of detection. In addition, we assume there exists a q-dimensional covariate vector, **U**, that is fully observed and included in the model.

For simplicity, we set p = q = 1 and assume the generalized linear model:

$$g[E(Y)] = \beta_0 + \beta_1\tilde{X}_1 + \beta_2 U_1 ,$$

where g (.) is a link function and $(\beta_0,\ \beta_1,\ \beta_2)$ are regression coefficients. The parameter of interest is β_1, the regression coefficient reflecting the main effect of $\tilde{X}_1$. In the presence of limit of detection, the observed data are $(Y_{\{i\}}, X_{\{1i\}}, U_{\{1i\}}, \Delta_{\{1i\}})$, for i = 1,…, n, where $X_{\{1i\}} = \max(\tilde{X}_{\{1i\}},\ \alpha_1)$, $\Delta_{\{1i\}} = I(\tilde{X}_{\{1i\}} \le \alpha_1)$, α_1 is the limit of detection for $\tilde{X}_1$, and I (.) is the indicator function. With the missing indicator Δ_1, the CC model can be expressed as a modification of the aforementioned model as follows:

$$g[E(Y)] = \beta_{c0}(1-\Delta_1) + \beta_{c1}X_1(1-\Delta_1) + \beta_{c2}U_1\ (1-\Delta_1)$$

It can be shown that the CC estimators are consistent estimators for true parameters [21].

Despite its desirable asymptotic properties, when the censoring rate is high, the CC approach likely suffers from loss of information. As an alternate approach, we consider the MDI model defined as follows:

$$g[E(Y)] = \beta_{m0} + \beta_{m1}X_1(1-\Delta_1) + \beta_{m2}U_1\ + \beta_{m3}\Delta_1 .$$

In the context of linear regression, the least squares estimators of $(\beta_{m0},\ \beta_{m1},\ \beta_{m2})$ are asymptotically unbiased for $(\beta_0,\ \beta_1,\ \beta_2)$ in Equation (1), if X_1 and U_1 are uncorrelated [4].

Considering the setting in which there are two predictors subject to limits of detection, denoted $\tilde{X}_1$ and $\tilde{X}_2$, and a fully observed predictor U_1, the true model is assumed to follow:

$$g[E(Y)] = \beta_0 + \beta_1\tilde{X}_1 + \beta_2\tilde{X}_2 + \beta_3 U_1 \tag{1}$$

For this setting, the MDI model is given by

$$g[E(Y)] = \beta_{m0} + \beta_{m1}X_1(1-\Delta_1) + \beta_{m2}X_2(1-\Delta_2) + \beta_{m3}U_1 + \beta_{m4}\,\Delta_1 + \beta_{m5}\,\Delta_2, \tag{2}$$

where $(\Delta_1,\ \Delta_2)$ are the missing indicators for the predictors $\tilde{X}_1$ and $\tilde{X}_2$, respectively. The estimates of $(\beta_{m0},\ \beta_{m1},\ \beta_{m2},\ \beta_{m3})$ in Equation (2) are used to estimate the parameters $(\beta_0,\ \beta_1,\ \beta_2,\ \beta_3)$ in Equation (1).

In addition, we consider the expanded MDI model with interactions between the fully observed covariate and the MDIs as follows:

$$g[E(Y)] = \beta_{m0} + \beta_{m1}X_1(1-\Delta_1) + \beta_{m2}X_2(1-\Delta_2) + \beta_{m3}U_1 + \beta_{m4}\,\Delta_1 + \beta_{m5}\,\Delta_2\ + \beta_{m6}\,\Delta_1U_1\ + \beta_{m7}\,\Delta_2U_1 . \tag{3}$$

Model (3) is useful in settings in which the censored covariates have interaction effects with the fully observed covariates.

Simulation

We present results from simulation to assess the performance of the MDI approach incorporated in logistic and conditional logistic regression models. The latter is appropriate for the matched case-control study design of the cardiovascular disease biomarker study. In all simulation studies, we considered three possibly correlated covariates, denoted $\tilde{X}_1,\ \tilde{X}_2, U_1$ as specified in Equation (1), where U_1 is fully observed, and $\tilde{X}_1,\ \tilde{X}_2$ are left-censored. Details of the simulation are included in Supplement.

Using results obtained in the setting in the absence of censoring as the gold standard (M1), we compared the performance of six approaches to handling missing data. These include

- M2: complete case (CC) analysis;
- M3: the MDI model in Equation (2);
- M4: the expanded MDI model in Equation (3);
- M5: substitution of the missing value by one-half the observed minimum;
- M6: imputation of the missing value using the predictive mean matching (PMM) algorithm implemented in the R package *mice*. Here, imputed values are selected randomly from among the five observed values of the covariate whose regression-predicted values are closest to the regression-predicted value of the missing covariate [14,22,23]. This procedure produces imputed covariate values that lie in the range of the observed covariate values; and
- M7: imputation of the missing value using the missForest algorithm implemented in the R package *missForest* [13]. In this case, the missing values are directly predicted using a random forests model that is trained on the observed parts of the data set. This approach does not make distributional assumptions inherent in the PMM algorithm in M6. Default parameter values of 100 trees and mtry = 1 were assumed.

Logistic regression

Assuming the model in Equation (1), we implemented two simulation settings corresponding to (1) multivariate normal and

(2) non-normal distribution for ($\tilde{X}_1$, $\tilde{X}_2$, U_1). In both settings, Spearman's rho (ρ) was used to specify the strength of dependence between $\tilde{X}_1$, $\tilde{X}_2$, and U_1. We simulate data by setting the coefficients $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$ in the model given in Equation (1). We compare the bias and MSE associated with the maximum likelihood estimates (MLEs) of β_1 and β_2. Total sample sizes of $n = 100$ and 200 with equal numbers of cases and controls were considered. All models incorporated the bias reduction method for reduction in finite sample bias as implemented in the *brglm* R package [24]. The simulation was repeated 100,000 times and results averaged.

Bias and MSE: Figure 1 (Table 2 in Supplement) summarizes the average bias and MSE associated with the MLEs of β_1, β_2 in Equation (1), when $n = 100$ (50 cases). The MDI (M3) and expanded MDI (M4) approaches had among the smallest bias across all settings considered for logistic regression (Fig. 1, Table 2 in Supplement). These general trends persisted for a larger sample size of $n = 200$ (100 cases) (Supplemental Table 3).

The MDI approach (M3) has a clear advantage over the CC approach (M2) for small and moderate sample sizes, with respect to MSE. The MDI and CC approaches had generally equivalent bias across all settings considered.

Estimates of bias were largest for imputation using *mice* (M6) and *missForest* (M7). The distributions of the difference between the true and imputed covariate values from *mice* (M6) and *missForest* (M7) were examined for X_1 and X_2 under each simulation setting. The imputed values were uniformly larger than the corresponding true covariate values that fall below the limit of detection—see Figure 1 in Supplement for a representative distribution. Imputed values from *mice* (M6) are always randomly selected from among a set of observed values of the covariate whose regression-predicted values are closest to the regression-predicted value of the missing covariate. Imputed values from *missForest* (M7) are based on predicted values from a random forests classifier trained on the observed parts of the data set, resulting in imputed values that were in the range of the observed covariate values. Despite large bias, the MSE of the imputation approaches using *mice* and *missForest* were comparable and sometimes smaller than that of MDI and expanded MDI.

The imputation approach based on one half the minimum observed value (M5) had larger bias when compared with the MDI approaches overall; however, the bias associated with this ad hoc approach was substantially lower in the non-normal setting. This is driven by the fact that the conditional expectation, E (X | X < limit of detection), is well approximated by one half the minimum observed for the non-normal setting, but not in the normal setting (Table 1 in the Supplement).

Figure 1 (Table 2 in Supplement) results also suggest that higher ρ is generally associated with larger MSE in CC (M2), MDI (M3), and expanded MDI (M4) models. However, the effect of ρ on the magnitude of the bias is less noticeable. As expected, because the censoring rate is higher for $\tilde{X}_2$ when compared with $\tilde{X}_1$, the corresponding regression coefficient estimate is associated with larger bias and MSE. The non-normal distribution setting was observed to be associated with lower bias and MSE when compared with the gaussian setting.

Bias and MSE estimates for all approaches are presented for a larger sample size of $n = 200$ (100 cases) subjects in Supplemental Table 3.

Type I error and power corresponding to the individual hypothesis tests for $H_0: \beta_1 = 0$, $H_0: \beta_2 = 0$ are summarized in Table 1. We estimated power and type-I error by computing the proportion (of 100,000 replicates) of *P*-values ≤ 0.05. For the MDI approaches (M3, M4), the *P*-values were obtained from a likelihood ratio test of the composite null hypothesis of the coefficients associated with X_k, Δ_k for $k = 1, 2$. For the CC approach (M2) and other imputation models (M5, M6, M7), the *P*-values were computed based on a Wald test for each of the parameters of

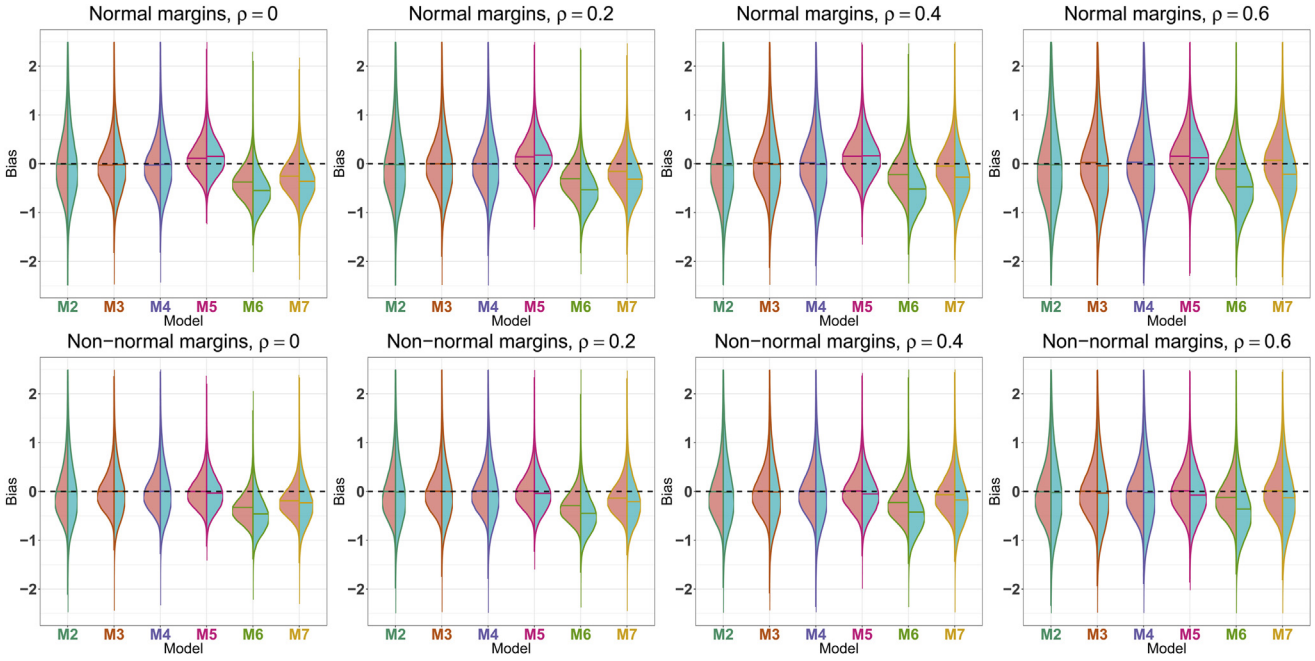

Fig. 1. *Bias and MSE associated with estimates of regression coefficients for logistic regression models.* The sample size considered here is $n = 100$ (50 cases). Results are based on 100,000 converged replications. Approaches with mean bias reduction to estimate the true regression coefficients β_1, β_2 are as follows: M2 denotes the complete case analysis; M3 denotes the missing data indicator (MDI) model; M4 denotes the expanded missing data indicator (MDI) model; M5 denotes imputation using one half the observed minimum value; M6 denotes predictive mean matching (PMM) imputation implemented in R package *mice*; and M7 denotes the missForest algorithm implemented in the R package *missForest*.

Table 1
Summary of power and type-I error for logistic regression models

ρ		M1	M2	M3	M4	M5	M6	M7
		Normal margins						
		Power						
0	β_1	0.969	0.284	0.931	0.897	0.944	0.472	0.536
	β_2	0.970	0.179	0.907	0.868	0.926	0.188	0.225
0.2	β_1	0.940	0.261	0.898	0.854	0.919	0.506	0.582
	β_2	0.941	0.155	0.861	0.808	0.887	0.174	0.211
0.4	β_1	0.867	0.236	0.819	0.757	0.850	0.502	0.579
	β_2	0.866	0.131	0.751	0.684	0.793	0.145	0.180
0.6	β_1	0.707	0.203	0.664	0.581	0.703	0.467	0.525
	β_2	0.711	0.108	0.571	0.501	0.615	0.114	0.134
		Type-I error						
0	β_1	0.043	0.015	0.045	0.038	0.034	0.037	0.039
	β_2	0.042	0.015	0.053	0.048	0.031	0.036	0.041
0.2	β_1	0.044	0.017	0.046	0.039	0.036	0.037	0.039
	β_2	0.042	0.016	0.053	0.047	0.031	0.036	0.040
0.4	β_1	0.043	0.019	0.047	0.038	0.036	0.038	0.039
	β_2	0.043	0.018	0.054	0.049	0.034	0.037	0.038
0.6	β_1	0.042	0.021	0.048	0.039	0.038	0.038	0.039
	β_2	0.042	0.019	0.053	0.049	0.035	0.035	0.035
		Power						
0	β_1	0.970	0.460	0.950	0.923	0.958	0.734	0.810
	β_2	0.832	0.241	0.759	0.701	0.790	0.259	0.334
0.2	β_1	0.944	0.477	0.920	0.886	0.929	0.729	0.804
	β_2	0.777	0.224	0.698	0.633	0.731	0.240	0.302
0.4	β_1	0.877	0.457	0.844	0.794	0.853	0.692	0.762
	β_2	0.665	0.201	0.578	0.515	0.613	0.203	0.250
0.6	β_1	0.719	0.397	0.680	0.619	0.686	0.610	0.648
	β_2	0.490	0.163	0.412	0.363	0.433	0.161	0.188
		Type-I error						
0	β_1	0.034	0.015	0.045	0.037	0.033	0.034	0.035
	β_2	0.037	0.015	0.053	0.046	0.032	0.034	0.039
0.2	β_1	0.035	0.016	0.045	0.036	0.034	0.033	0.034
	β_2	0.038	0.017	0.053	0.046	0.034	0.035	0.038
0.4	β_1	0.033	0.017	0.045	0.036	0.032	0.032	0.033
	β_2	0.038	0.018	0.052	0.047	0.033	0.033	0.036
0.6	β_1	0.033	0.018	0.048	0.038	0.031	0.031	0.030
	β2	0.038	0.018	0.054	0.048	0.034	0.033	0.032

Each entry represents the proportion of *P*-value less than or equal to 0.05 of 100,000 replicates. The sample size is $n = 100$ (50 cases). M1 denotes the true model before censoring; M2 denotes the complete case analysis; M3 denotes the missing data indicator (MDI) model; M4 denotes the expanded missing data indicator (MDI) model; M5 denotes imputation using one half the observed minimum value; M6 denotes predictive mean matching (PMM) imputation implemented in R package *mice*; and M7 denotes the missForest algorithm implemented in the R package *missForest*.

interest. To obtain type-I error, we set $\beta_1 = \beta_2 = \beta_3 = 0$ in the data generating model.

For all approaches, larger correlation (ρ) results in lower power, mirrored by the corresponding increase in MSE seen in Figure 1 (Supplemental Table 2). Imputation using one-half the minimum observed value (M5) was associated with the highest power—however, this advantage is offset by the relatively large bias in the normal distribution setting when the imputed value is not a good estimate of E (X | X < limit of detection). The power associated with MDI (M3) was among the highest, whereas that of the expanded MDI model (M4) was somewhat lower. This is expected as the expanded MDI model includes two additional parameters, resulting in a corresponding loss of power. The power associated with imputation using *mice* (M6) and *missForest* (M7) was considerably lower than that of the two MDI approaches. Similar trends were observed for both the normal margins and the non-normal margins.

These observations demonstrate a clear advantage of the MDI approaches over the CC approach and other imputation approaches in small to modest sample size settings. The type-I error associated with the MDI model is in good agreement with the nominal value of 0.05.

Power and type 1 error rate estimates for all approaches are presented for a larger sample size of $n = 200$ (100 cases) subjects in Supplemental Table 4.

Conditional logistic regression

We considered two settings with respect to the joint distribution of $\tilde{X}_1, \tilde{X}_2, U_1$: (1) multivariate normal distribution and (2) non-normal distribution; the Spearman's rho, ρ, was used to specify the strength of dependence between $\tilde{X}_1, \tilde{X}_2, U_1$. To mimic a matched study design, for each subject, an unobserved variable $\varepsilon \sim N(1, \sigma = 1.5)$ was simulated. The deciles of the distribution of ε were used to determine the matching stratum *g* for every subject. For each subject, the covariates $\tilde{X}_1, \tilde{X}_2, U_1$ were simulated according to a multivariate distribution with stratum-specific parameters. In the multivariate normal setting, both mean and variance parameters were assumed to depend on matching stratum. In the non-normal setting, the rate parameters of the exponential and Weibull distributed covariates were assumed to depend on matching stratum. Details of the simulations are presented in Supplement.

The binary outcome was simulated according to

$$P(Y=1) = \frac{e^{\beta_0 + \beta_1\tilde{X}_1 + \beta_2\tilde{X}_2 + \beta_3 U_1 + \beta_4\varepsilon}}{1 + e^{\beta_0 + \beta_1\tilde{X}_1 + \beta_2\tilde{X}_2 + \beta_3 U_1 + \beta_4\varepsilon}},$$

where $\beta_0 = -3$, $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 1$.

A matched data set was generated by selecting *m* cases and *m* matching controls, where the matching was done within group *g*.

Bias and MSE corresponding to MLEs of β_1, β_2 in Equation (1) are shown in Figure 2 (Table 5 of the Supplement) for $n = 136$ (68 matched pairs). The MDI approach (M3) shows a clear advantage over the CC (M2) and *mice* imputation (M6) approaches in terms of bias and MSE. The expanded MDI model (M4) has a consistently larger bias and MSE when compared with the MDI approach (M3). The bias and MSE reduction in the MDI model is more substantial when compared with the CC approach in the context of a matched study because the CC approach drops study pairs when at least one of its pairs is missing. For the multivariate normal distribution setting, the substitution with one-half the minimum value (M5) achieves comparable bias and MSE to that of the MDI approach (M3), when $n = 136$ (Fig. 2, Table 5 of Supplement). However, the trends reverse in favor of the MDI approach for larger sample sizes such as $n = 400$ (Supplemental Table 8). Imputation with *missForest* (M7) appears to have a somewhat larger bias when compared with the MDI (M3) in settings of low ρ—however, this trend is reversed in favor of *missForest* imputation (M7) when ρ is large.

Higher ρ, reflecting higher correlation between covariates is associated with larger MSE for the MDI model (M3); as in logistic regression, the effect of ρ on bias is modest. Because the censoring percentage is higher for $\tilde{X}_2$ when compared to $\tilde{X}_1$, the estimate of β_2 has larger bias and MSE when compared with β_1 in the MDI model. In the CC approach, the bias associated with estimates of β_1, β_2 are comparable because the study subjects are discarded at the cluster level; however, the MSE corresponding to the estimate of β_2 is still observed to be larger than that for β_1. In most cases, similar trends are observed when comparing the various approaches in the non-normal distribution setting. An exception was in the case of substitution with one half the minimum (M5)—under non-normal covariate distributions, this approach achieves minimum bias and MSE, with a substantial advantage over other approaches when ρ is large. This is driven by the fact that the imputed values are good approximations to the expectation E (X | X < limit of detection) (See Table 1 in Supplement).

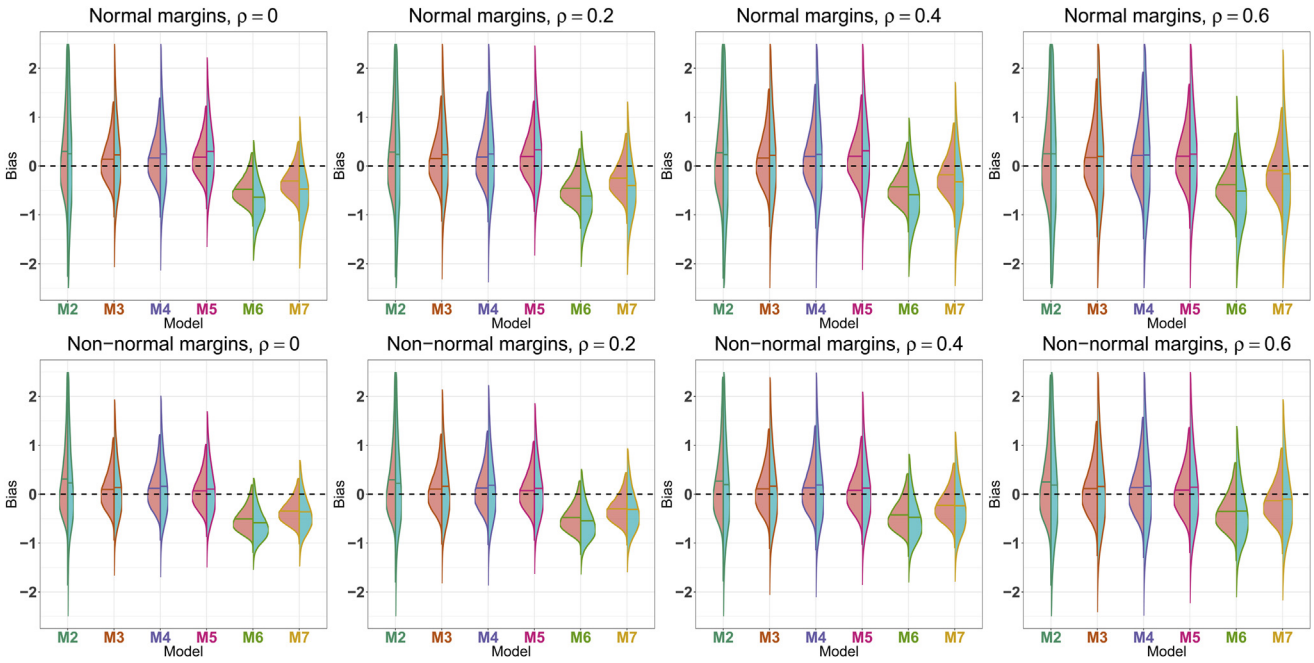

Fig. 2. *Bias and MSE associated with estimates of regression coefficients for conditional logistic regression models. The sample size considered here is n = 136 (68 matched pairs). Results are based on 100,000 converged replications.* Approaches with mean bias reduction to estimate the true regression coefficients β_1, β_2 are as follows: M2 denotes the complete case analysis; M3 denotes the missing data indicator (MDI) model; M4 denotes the expanded missing data indicator (MDI) model; M5 denotes imputation using one half the observed minimum value; M6 denotes predictive mean matching (PMM) imputation implemented in R package *mice*; and M7 denotes the missForest algorithm implemented in the R package *missForest*.

Bias and MSE estimates for all approaches are presented for larger sample sizes of $n = 200$ (100 matched pairs) and $n = 400$ (200 matched pairs) subjects in Supplemental Tables 6 and 8.

Type I error and power are summarized in Table 2 corresponding to the hypothesis tests $H_0: \beta_1 = 0$, $H_0: \beta_2 = 0$. Trends observed here were similar to that for logistic regression—the MDI approach (M3) yields substantial higher power when compared with the CC (M2) and imputation using *mice* (M6) and *missForest* (M7) approaches, in all scenarios considered. The expanded MDI (M4) and substitution using one half the observed minimum (M5) had comparable but lower power than the MDI approach.

Power and type 1 error rate estimates for all approaches are presented for larger sample sizes of $n = 200$ (100 matched pairs) and $n = 400$ (200 matched pairs) subjects in Supplemental Tables 7 and 9.

Cardiovascular disease biomarker study

This matched case-control study was conducted by the High-Risk Plaque Initiative (BG Medicine Inc., Waltham, MA, and other partners) to discover prognostic biomarkers in blood plasma for near-term cardiovascular events. Matched cases and controls were selected from the CATHGEN study, in which peripheral blood samples were collected from consenting research subjects undergoing cardiac catheterization at Duke University Medical Center from 2001 through 2011[25]. 68 cases were selected from among individuals who had a major adverse cardiac event (MACE) within two years following the time of their sample collection. 68 controls were selected from individuals who were MACE-free for the two years following sample collection and were matched to cases on age, gender, race/ethnicity, and severity of coronary artery disease. High-content mass spectrometry based techniques were used to quantify 472 metabolites from each subject's serum specimen [26,27]. The identities of the measured metabolites and proteins are masked because of a data confidentiality agreement.

Of the 472 quantified metabolites, 99 metabolites have at least one missing value. Among these 99 metabolites, the median number of pairs missing at least one measurement was 16. Each metabolite was analyzed in a MDI model:

$$g[E(Y)] = \beta_0 + \beta_1 X(1-\Delta) + \beta_2 \Delta\,,$$

where X is the metabolite that is subject to limit of detection, and Δ is the missing indicator defined as $\Delta = \mathrm{I}\,(X \leq \alpha)$ for some limit of detection threshold α. For comparison, we also analyzed the data using the CC approach. We used a Wald test to test for the significance of the main effect of metabolite in the CC model and a likelihood ratio test to test for the joint significance of both the main effect and the indicator term in the MDI model. Because the observed X is highly correlated with missing indicator, Δ, we also fit a separate model for each metabolite including the missing indicator as the only covariate, referred to as the Δ-model below. The Δ-model enables us to test whether the metabolite level falling below the limit of detection is associated with the outcome, MACE.

At the 0.05 level of significance, our analysis identified 15 metabolites for which the CC and MDI models both converged and yielded discordant results based on a *P* value threshold of 0.05; that is, one model has a *P*-value less than 0.05, whereas the other has a *P*-value greater than 0.05. There was one additional metabolite for which the CC model did not converge, but the MDI model converged. This metabolite had extensive censoring where at least one member of 81% of the matched pairs had an undetectable (missing) value. The results for the 15 metabolites with discordant results when comparing the CC and MDI models are presented in Table 3.

As expected, when the censoring percent is <10% of the matched pairs, the point estimates and the 95% confidence intervals from the

Table 2
Summary of power and type-I error for conditional logistic regression models

ρ		M1	M2	M3	M4	M5	M6	M7
		Normal margins						
		Power						
0	β_1	0.889	0.423	0.906	0.873	0.880	0.431	0.500
	β_2	0.302	0.104	0.320	0.290	0.270	0.053	0.064
0.2	β_1	0.793	0.412	0.839	0.799	0.793	0.385	0.448
	β_2	0.234	0.105	0.279	0.252	0.234	0.057	0.055
0.4	β_1	0.652	0.351	0.721	0.670	0.652	0.306	0.365
	β_2	0.173	0.106	0.222	0.207	0.173	0.048	0.049
0.6	β_1	0.419	0.373	0.526	0.489	0.419	0.217	0.261
	β_2	0.115	0.105	0.167	0.167	0.115	0.043	0.044
		Type-I error						
0	β_1	0.043	0.004	0.048	0.072	0.041	0.038	0.041
	β_2	0.040	0.004	0.047	0.064	0.038	0.042	0.039
0.2	β_1	0.044	0.006	0.043	0.071	0.044	0.045	0.046
	β_2	0.039	0.005	0.042	0.066	0.039	0.043	0.040
0.4	β_1	0.045	0.010	0.047	0.071	0.045	0.042	0.044
	β_2	0.045	0.008	0.048	0.068	0.045	0.038	0.042
0.6	β_1	0.045	0.013	0.041	0.074	0.045	0.046	0.044
	β_2	0.042	0.013	0.047	0.065	0.042	0.043	0.041
		Power						
0	β_1	0.874	0.547	0.923	0.885	0.873	0.545	0.651
	β_2	0.255	0.279	0.386	0.344	0.253	0.105	0.121
0.2	β_1	0.807	0.564	0.872	0.833	0.807	0.509	0.594
	β_2	0.196	0.240	0.307	0.277	0.196	0.093	0.095
0.4	β_1	0.668	0.530	0.766	0.708	0.668	0.433	0.500
	β_2	0.151	0.211	0.254	0.230	0.151	0.078	0.082
0.6	β_1	0.461	0.460	0.589	0.533	0.461	0.323	0.362
	β_2	0.091	0.190	0.171	0.165	0.091	0.060	0.056
		Type-I error						
0	β_1	0.038	0.031	0.048	0.066	0.038	0.040	0.037
	β_2	0.033	0.011	0.045	0.065	0.032	0.042	0.037
0.2	β_1	0.037	0.032	0.043	0.067	0.037	0.038	0.038
	β_2	0.038	0.012	0.041	0.067	0.038	0.032	0.035
0.4	β_1	0.037	0.034	0.048	0.072	0.037	0.037	0.037
	β_2	0.037	0.013	0.047	0.064	0.037	0.040	0.034
0.6	β_1	0.042	0.039	0.043	0.078	0.042	0.039	0.040
	β_2	0.039	0.018	0.047	0.067	0.039	0.037	0.036

Each entry represents the proportion of *P*-value less than or equal to 0.05 of 100,000 replicates. The sample size is $n = 68$ matched pairs. M1 denotes the true model before censoring; M2 denotes the complete case analysis; M3 denotes the missing data indicator (MDI) model; M4 denotes the expanded missing data indicator (MDI) model; M5 denotes imputation using one half the observed minimum value; M6 denotes predictive mean matching (PMM) imputation implemented in R package mice; and M7 denotes the missForest algorithm implemented in the R package *missForest*.

CC and MDI models are almost identical. As the censoring proportion increases, the results from the two models begin to diverge, due to increasing differences in sample sizes. Of the 15 metabolites shown in Table 3, four metabolites (Mx 1 to Mx 4) have similar results from the two models; that is, one model has a significant association with $P < 0.05$ and the other model has a marginally significant association ($0.05 < P < 0.1$). Six other metabolites (Mx 5 to Mx 10) have statistically significant *P*-values ($P < 0.05$) resulting from the MDI model but nonsignificant *P*-values ($P > 0.1$) from the CC model. These six metabolites have censoring levels ranging from 4% (Mx 5) to 29% (Mx 10). For each of these six metabolites, the CC coefficient estimate is comparable with that from the MDI model. In addition, four of these metabolites (Mx 7 to Mx 10) have a statistically significant association in the Δ-model. In the case of metabolites Mx 5 and Mx 6, the Δ-model did not converge due to perfect separation. For metabolites Mx 5 to Mx 10, the missing indicator Δ is strongly associated with outcome. These results are consistent with the observations in the simulations that indicate that the more efficient use of information in the MDI model yields increased power. Three metabolites (Mx 11 to Mx 13) had statistically significant ($P < 0.05$) associations in the CC model but insignificant *P*-values in the MDI model. These metabolites had censoring levels between 24% and 37%—in all three cases, the Δ-model results are insignificant with *P*-values exceeding 0.5, thereby resulting in insignificant *P*-values from the MDI approach. Finally, two metabolites (Mx 14 and Mx 15) with heavy censoring levels of approximately 80%–81% have significant associations ($P < 0.05$) in the CC model but nonsignificant associations ($P > 0.1$) in the MDI model. Because these metabolites have extreme levels of censoring, it is difficult to determine which model is appropriate (if any).

Discussion

Missing covariates are commonly encountered in many biomedical investigations. In studies using high-throughput metabolomic technologies, missing values can arise because of a combination of limit of detection issues and random instrument failure. In these settings, investigators often consider discarding data from subjects with incomplete covariate measurements and/or various imputation techniques. Previous literature has shown that ad hoc imputation techniques can result in severe bias [1,2]. More complex likelihood-based methods rely on stringent parametric assumptions and can be computationally intensive. Discarding subjects with partially missing values in a CC analysis can result in a dramatic reduction in power. In these settings, the MDI or MDI approach is an attractive alternative as all available information remains in the analysis to maintain statistical power. In this article, for settings of moderate sample size, a binary outcome, and where the missingness in the covariate is nonrandom and is a result of falling below the limit of detection, we evaluated the bias, MSE, and statistical power associated with the MDI model when compared with the CC analysis and three different imputation approaches.

In our simulation study, the MDI approach shows a clear advantage over the CC approach when there are two censored covariates. Imputation using two other strategies, namely the missForest algorithm and the PMM algorithm implemented in the *mice* R package did not show improved performance when compared with MDI approach. In particular, these approaches resulted in imputed values in the range of the observed covariate distributions, resulting in large bias due to the nonrandom censoring mechanism (Fig. 1 in Supplement). The advantage of the MDI approach was preserved in settings of normal and non-normal covariate distributions and in the presence of modest correlation between the censored covariates. The advantage of the MDI model over the CC approach was more substantial for the analysis of matched case-control studies in conditional logistic regression models. The MDI approach can be easily adopted in multivariable models that include several censored covariates jointly, in which setting the CC approach would be severely impacted as it discards the union of subjects with at least one missing value.

In our study, the performance of logistic regression and conditional logistic regression models were evaluated solely in the context of missing covariate values that arise due to limit of detection limitations of the assay. The MDI approach can be used in data sets where missingness is due to other mechanisms, including data that are missing completely at random. In these settings, we also expect approaches such as the PMM-based imputation (*mice*) and missForest algorithms to have better performance. However, a comparative evaluation of the MDI approach relative to PMM and missForest algorithms for more general missing data mechanisms was not studied in this article.

For the logistic regression analyses, the imputation approach based on one half the minimum observed value had larger bias relative to the MDI model in the normal distribution setting; however, the bias associated with this ad hoc approach was

Table 3
Cardiovascular disease biomarker study

Metabolites	Cen %	Complete case (CC) model			Missing indicator (MDI) model			Δ-model*
		EST	95% CI	*P*-value		95% CI	*P*-value	*P*-value
				X	EST		(*X*, Δ)	Δ
Mx 1	0.02	0.43	(0.19, 0.97)	.03	0.43	(0.19, 0.97)	.05	—
Mx 2	0.10	2.32	(1.03, 5.26)	.03	2.08	(0.96, 4.55)	.08	.27
Mx 3	0.10	1.97	(0.96, 4.03)	.05	1.87	(0.94, 3.73)	.09	.27
Mx 4	0.81	0.18	(0.01, 2.73)	.09	0.17	(0.03, 1.02)	.05	.72
Mx 5	0.04	0.68	(0.40, 1.16)	.14	0.68	(0.40, 1.16)	.04	—
Mx 6	0.09	0.59	(0.01, 25.61)	.78	0.59	(0.01, 25.61)	.02	—
Mx 7	0.18	1.71	(0.45, 6.45)	.43	1.50	(0.41, 5.52)	.01	.02
Mx 8	0.18	1.63	(0.46, 5.79)	.45	1.45	(0.42, 5.02)	.01	.02
Mx 9	0.22	1.21	(0.92, 1.59)	.17	1.17	(0.90, 1.52)	.01	.01
Mx 10	0.29	0.79	(0.27, 2.30)	.67	0.77	(0.29, 2.08)	.03	.02
Mx 11	0.24	2.52	(1.07, 5.97)	.02	2.18	(0.98, 4.86)	.12	.62
Mx 12	0.24	4.29	(0.99, 18.54)	.04	3.50	(0.95, 12.92)	.13	.62
Mx 13	0.37	0.68	(0.48, 0.98)	.02	0.83	(0.63, 1.08)	.34	.68
Mx 14	0.79	0.29	(0.06, 1.49)	.05	0.76	(0.41, 1.42)	.63	.72
Mx 15	0.81	0.09	(0.01, 1.67)	.02	0.41	(0.12, 1.45)	.29	.72

Summary of results for 15 metabolites with discordant results from the complete case (CC) and missing indicator (MDI) models. We use cen % to denote the censoring proportion at cluster level; a cluster is considered censored if at least one of its element is censored. EST denotes the estimate of the log odds ratio. The *P*-values presented under the CC model and the Δ-model were obtained by Wald tests. The *P*-values presented under the MDI model was obtained by a likelihood ratio test of the joint association of (X, Δ)

* The Δ-model did not converge for Mx 1, Mx 5, and Mx 6 because of perfect separation.

substantially lower in the non-normal setting. For the simulation settings considered in this article, the imputed value based on by one half the minimum observed value closely approximated the conditional expectation, E (X | X < limit of detection), in the non-normal setting, but not in the normal setting (Table 1 in the Supplement). For the conditional logistic regression analyses, the imputation approach based on one half the minimum observed value did as well or better than the MDI model. As discussed in Cole, S. R. et. al. (2009) [2], the observed bias is a function of the imputed value. Thus, when the ad hoc substitution approach closely approximates the conditional expectation of the covariate, the resulting bias is low.

To the best of our knowledge, theoretical asymptotic properties for the missing indicator model have been derived only for linear regression and are not yet available for more general settings [4]. We acknowledge that the MDI approach could be asymptotically biased for general link functions. Here, we assumed that the censoring of the covariates is independent of the outcome when conditioning on the covariates. The performance of the missing indicator approach under informative censoring warrants further investigation. It would also be useful to evaluate the properties of the MDI approach in models for count data and survival outcomes.

Acknowledgment

This work was supported in part by the Harvard NeuroDiscovery Center, National Institutes of Health grants T32NS048005 and R01HL122241.

Authors' contributions: All authors have participated in conception and design or analysis and interpretation of the data; drafting the article or revising it critically for important intellectual content; and approval of the final version.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.annepidem.2019.07.014.

References

[1] Schisterman EF, Vexler A, Whitcomb BW, Liu A. The limitations due to exposure detection limits for regression models. Am J Epidemiol 2006;163(4):374–83.
[2] Cole SR, Chu H, Nie L, Schisterman EF. Estimating the odds ratio when exposure has a limit of detection. Int J Epidemiol 2009;38(6):1674–80.
[3] Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Multiple Imputation of a Randomly Censored Covariate Improves Logistic Regression Analysis. J Appl Stat 2016;43(15):2886–96.
[4] Jones MP. Indicator and stratification methods for missing explanatory variables in multiple linear regression. J Am Stat Assoc 1996;91(433):222–30.
[5] Cohen J, Cohen P. Applied Multiple Regression Correlation Analysis for the Behavioral Sciences. New York: John Wiley; 1975.
[6] Miettinen OS. Theoretical Epidemiology: principles of occurrence research. New York: John Wiley and Sons; 1985.
[7] Wei R, Wang J, Su M, Jia E, Chen S, Chen T, et al. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. Sci Rep 2018;8(1):663.
[8] Wei R, Wang J, Jia E, Chen T, Ni Y, Jia W. GSimp: A Gibbs sampler based left-censored missing value imputation approach for metabolomics studies. PLoS Comput Biol 2018;14(1):e1005973.
[9] Trainor PJ, DeFilippis AP, Rai SN. Evaluation of Classifier Performance for Multiclass Phenotype Discrimination in Untargeted Metabolomics. Metabolites 2017;7(2):30.
[10] Armitage EG, Godzien J, Alonso-Herranz V, Lopez-Gonzalvez A, Barbas C. Missing value imputation strategies for metabolomics data. Electrophoresis 2015;36(24):3050–60.
[11] Gromski PS, Xu Y, Kotze HL, Correa E, Ellis DI, Armitage EG, et al. Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites 2014;4(2):433–52.
[12] Jin Z, Kang J, Yu T. Missing value imputation for LC-MS metabolomics data by incorporating metabolic network and adduct ion relations. Bioinformatics 2018;34(9):1555–61.
[13] Stekhoven DJ, Buhlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28(1):112–8.
[14] Buuren SV, Groothuis-Oudshoorn K. {mice}: Multivariate Imputation by Chained Equations in R. J Stat Softw 2011;45(3):1–67.
[15] Do KT, Wahl S, Raffler J, Molnos S, Laimighofer M, Adamski J, et al. Characterization of missing values in untargeted MS-based metabolomics data and evaluation of missing data handling strategies. Metabolomics 2018;14(10):128.
[16] Ma Y, Yin G. Censored quantile regression with covariate measurement errors. Stat Sinica 2011;21:949–71.
[17] Wang JH, Stefanski LA, Zhu Z. Corrected-loss estimation for quantile regression with covariate measurement errors. Biometrika 2012;99(2):405–21.
[18] D'Angelo G, Weissfeld L, I.M.S.I. Gen. An index approach for the Cox model with left censored covariates. Stat Med 2008;27(22):4502–14.
[19] Wang HJ, Feng X. Multiple imputation for M-regression with censored covariates. J Am Stat Assoc 2012;107(497):194–204.
[20] Sattar A, Sinha SK, Morris NJ. A parametric survival model when a covariate is subject to left-censoring}. J Biom Biostat 2012;2.

[21] Bernhardt PW, Wang HJ, Zhang D. Flexible modeling of survival data with covariates subject tos detection limits via multiple imputation. Comput Stat Data Anal 2014;69:81–91.
[22] Heitjan DF, Little RJ. Multiple imputation for the fatal accident reporting system. J R Stat Soc Ser C (Appl Stat) 1991;40(1):13–29.
[23] Schenker N, Taylor JM. Partially parametric techniques for multiple imputation. Comput Stat Data Anal 1996;22(4):425–46.
[24] Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27–38.
[25] Shah SH, Granger CB, Hauser ER, Krauss WE, Sun JL, Pieper K, et al. Reclassification of cardiovascular risk using integrated clinical and molecular biosignatures: Design of and rationale for the Measurement to Understand the Reclassification of Disease of Cabarrus and Kannapolis (MURDOCK) Horizon 1 Cardiovascular Disease Study. Am Heart J 2010;160(3):371–379 e2.
[26] Guo Y, Graber A, McBurney RN, Balasubramanian R. Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms. BMC Bioinformatics 2010;11:447.
[27] Kraus WE, Granger CB, Sketch MH, Donahue MP, Ginsberg GS, Hauser ER, et al. A Guide for a Cardiovascular Genomics Biorepository: the CATHGEN Experience. J Cardiovasc Transl Res 2015;8(8):449–57.