Multiple Imputation of Data Below Limits of Detection in Environmental Exposure Mixtures: Using Bayesian Multivariate Tobit Regression

Man Luo[1], Anne E. Nigra[2], Ana Navas-Acien[2], Yutao Liu[1], Bob Carpenter[3], Andrew Gelman[3], Qixuan Chen[1]

[1]Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY USA;

[2]Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, New York, NY, USA;

[3]Department of Statistics, Columbia University, New York, NY, USA;

Abstract (word limit: 250; current: 259)

Environmental health research is often challenged by environmental exposures with data below limits of detection (LODs). The commonly used method is to impute the data below the LOD with a constant or with predictions from a linear regression model. Such approaches work well when missing data are sparse, but can lead to bias in statistical inference when the proportion of data below the LOD is moderate to high. Moreover, environmental exposures often occur in complex mixtures, and thus imputation methods need to account for correlations between co-exposures. In this paper, we developed a Bayesian multivariate Tobit regression multiple imputation (MI) method for data below LODs in exposure mixtures, which assumes the data are correlated and follow a multivariate normal distribution, and utilizes individual-level predictors of mixture components to improve the imputation. Using high-quality, speciated urinary arsenic data measured in the Strong Heart Family Study (inorganic arsenic, monomethylarsonate, dimethylarsinate, and arsenobetaine, which have very low missingness), we conducted simulation studies to compare the performance of the developed multivariate Tobit regression method with alternative methods. Simulations indicated that the data imputed using Bayesian multivariate Tobit regression yield more accurate and precise estimates with 95% CI more likely to cover the true values in descriptive statistics and regression coefficients in linear models. Finally, we applied the proposed method to urinary arsenic species data from the 2013-2014 National Health and Nutrition Examination Survey, which have higher LODs and a moderate amount of missingness. MI using Bayesian multivariate Tobit regression is recommended for exposure mixtures with moderate to high LODs in environmental health research.

Keyword: arsenic; Bayesian modeling; limit of detection; MICE; missing data; multiple imputation; multivariate Tobit regression; Stan.

1. Introduction (word limit: 5000, current: 5568)

Epidemiologic analyses evaluating environmental exposures often rely on concentrations of environmental exposures measured in human biospecimens. These data often contain chemical concentrations measured below the analytical limit of detection (LOD), which is defined as the lowest value that can be reliably distinguished from a blank sample with a stated confidence level. Values below the LOD cannot be precisely measured, but can still contain important information. Simply excluding values below the LOD leads to substantially biased statistical inference.[1]

Researchers use different methods to deal with data that is missing due to LODs. Common techniques include data substitution with a constant, direct estimation of statistics, and multiple imputations.[2–5] Although simple substitution of data below the LOD with a single constant is commonly performed (e.g. LOD/√2 or LOD/2),[6] this approach can result in large bias in both descriptive statistics and analytical inference.[3,7] The standardized data quality assessment guideline of U.S. Environmental Protection Agency (EPA) recommends against this simple substitution method when the proportion of missing data due to the LOD is larger than 15%. [8] Direct estimation of descriptive statistics by assuming a certain distribution is another common way to handle data below LOD.[4,5,9] A limitation of the direct estimation approach is that individual exposure values are unavailable, so more complex analyses using the complete data are not possible. Alternatively, multiple imputation can be used to impute data below the LOD.[10] The resulted completed data set can then be used for any analytic inference (e.g. regression) in addition to summary statistics.

Although MI methods for data below LOD have been well studied, few have considered the LOD issue in exposure mixtures. Some studied exposures to multiple environmental

chemicals but only imputed one environmental exposure variable at a time, ignoring the possible correlations among the mixtures.[11,12] Others assumed bivariate or multivariate normal distribution, but either made the ignorable missingness assumption for data below LODs or did not incorporate possible predictors in the imputations of data below LODs.[13–16] However, the missing data caused by LODs are nonignorable, because the reason of missingness depends on the values of the exposure mixtures themselves.[17] Consequently, imputation methods that fail to account for this nonignorable missing mechanism could lead to poor imputations of data below LODs and thus biased statistical inference of the exposure mixtures. Further, including important predictors of environmental exposures can largely improve the imputations of data below LODs.

In this paper, we developed a Bayesian multivariate Tobit regression imputation method for exposure mixtures with missing data due to LODs, assuming a multivariate normal distribution for exposure mixtures. Our method accounts for nonignorable missing data mechanism via Tobit regression and models possible correlations among exposure mixtures via multivariate normal distribution. We compared the performance of the proposed model to alternative methods in estimating descriptive statistics and regression coefficients in linear regression using simulations with fully observed speciated urinary arsenic data from the Strong Heart Family Study, an epidemiologic cohort of American Indian adults. Arsenic is a particularly interesting chemical to study, as different exposure sources (mainly water and food) contribute different inorganic and organic species, some of which undergo additional biotransformation in the body, resulting in a complex mixture of arsenic species if urine that all contribute to total arsenic. We also applied the proposed method to the estimation of population means and quantiles of speciated urinary arsenic data using the 2013-2014 National Health and Nutrition Examination Survey (NHANES), where there are moderate proportions of arsenic data below

LODs. We provide a user-friendly R package ‘mvtobit’ for the easy implementation of the proposed method.

2. Methods

2.1 Study populations

In this paper, we used speciated urinary arsenic datasets from two epidemiologic cohorts to demonstrate the application of the proposed imputation method and to compare various methods of handling data below the LOD. We used speciated urinary arsenic measurements for participants in the Strong Heart Family Study (SHFS) as the study population for our simulations because very few urinry arsenic measurements were below the LODs. We illustrated the application of the proposed method using speciated urinary arsenic measurements for participants in the 2013-2014 National Health and Nutrition Examination Survey (NHANES), which has moderate LODs for several urinary arsenic species.

2.1.1 The Strong Heart Family Study (SHFS)

The Strong Heart Study (SHS) is a population-based prospective cohort study of 4,549 American Indian (AI) adults living in Arizona, Oklahoma, and North and South Dakota who are followed primarily for the study of cardiometabolic diseases. All eligible adults aged 45-74 were invited to participate at the SHS baseline exam (Phase I, 1989-1991, median age 55 years), and were re-examined during follow-up visits at Phase II (1993-1995) and Phase III (1998-1999). The participation rates for Arizona, Oklahoma, and North/South Dakota were 72%, 62%, and 55%, respectively, and 62% overall.[18] The SHFS is a multigenerational family-based extension of the SHS. Families with a core sibship consisting of at least 3 original SHS participants and at least 5 additional living family members were eligible to be included in SHFS.[19,20] SHS family members 15 years of age and older were recruited during the baseline SHFS examination (Phase IV, 2001-2003). We did not evaluate data for participants living in one community that later revoked

consent to participate. Urinary arsenic species were measured in SHFS baseline (2001-2003) spot urine samples collected from participants who were free from diabetes and who were also re-examined at Phase V (2006-2009) (N=1,949). The follow arsenic species were measured in urine: dimethylarsinate (DMA), monomethylarsonate (MMA), inorganic arsenic (iAs, sum of arsenate As^{+5} and arsenite As^{+3}), arsenobetaine (a nontoxic, seafood-derived organic arsenical), and total arsenic. The LOD and the percent of participants with values measured below the LOD were 0.1 μg/L (0%) for DMA, 0.1 μg/L (3.5%) for MMA, 0.1 μg/L (11.1%) for iAs, 0.1 μg/L (0.8%) for arsenobetaine, and 0.1 μg/L (0%) for total arsenic.[21] To account for hydration status and urine dilution, creatinine (mg/dL) was also measured in spot urine samples. During the examination, centrally trained SHS staff also collected data on participants' sex (male vs female), age (years), and body mass index (BMI, kg/m^2). Self-reported daily intakes of rice, juice, and wine/beer (grams) was derived from Block 119-item food frequency questionnaires, as previously described in detail.[22] Arsenobetaine below the LOD is replacted by LOD/$\sqrt{2}$. After excluding participants with MMA, iAs below the LOD and those missing urine creatinine (N=13), BMI (N=21), or self-reported dietary intake variables (N=116), our final sample size for the simulation study was 1,602 participants with complete measurements for all urinary arsenic species and covariates.

2.1.2 National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is a nationally representative survey of the non-institutionalized, general US population conducted in two-year cycles by the National Center for Health Statistics (NCHS). Our study used the 2013-2014 NHANES survey with data collected from the demographic questionnaire, dietary recall, clinical examination, and

laboratory examination. Similar to the SHFS, we studied measurements of urinary arsenic species. The LOD and the percent of participants with values measured below the LOD were 1.91 μg/L (24.1%) for DMA, 0.2 μg/L (29.7%) for MMA, 0.12 μg/L (27.8%) for As^{+3}, 0.79 μg/L (98.2%) for As^{+5}, 1.16 μg/L (56.3%) for arsenobetaine, and 0.26 μg/L (0%) for total arsenic, respectively.[21] Urine As^{+3} and As^{+5} were measured together in urine samples from the SHFS, but measured separately in urine samples from NHANES. To account for these differences, we pre-processed the NHANES As^{+5} data before analysis and summed the measured As^{+3} and As^{+5} measurements to generate iAs concentrations. Because all iAs in urine tends to be either As^{+3} or As^{+5} depending on pH, participants with low- to moderate- iAs exposure typically have either As^{+3} or As^{+5} present in urine.[16] If an individual had As^{+3} measured above the LOD but As^{+5} measured below the LOD, we replaced the missing As^{+5} with a value of 0. We evaluated the same covariates as in the SHFS, including urinary creatinine (mg/dL), BMI (kg/m^2), and self-reported past 24-hour intake of rice, juice, and wine/beer (grams), as previously described.[23] Urine total and speciated arsenic was measured in a one-third subset of NHANES 2013/2014 participants (N=2,755). We excluded participants who were missing urinary total arsenic (N=93) or urinary arsenobetaine concentrations (N=21), who were pregnant (N=20), who were missing body mass index (N=26), urinary creatinine (N=1), or education (N=21), or who had an unreliable dietary recall as determined by NHANES (N=447). The final sample included 2,126 participants from NHANES 2013-2014 cycles. We replaced data below LOD in the arsenobetaine with LOD/ $\sqrt{2}$ (56.3%). The final sample included 2,126 participants from NHANES 2013-2014 cycles.

2.2 Multiple Imputation (MI) for Data Below LOD

MI can be used to handle data below the LODs.[24] Different from single imputation, MI accounts for imputation uncertainty and yields valid statistical inference. Because of the right censoring, imputation models using ordinary least squares (OLS) regression can lead to biased statistical inference. Instead, censored regression that account for observations both above and below LODs can be used. In this section, we consider MI using censored regression with Tobit models.

2.2.1 Univariate Tobit Regression (UTR)

The Tobit model was first introduced by Arthur Goldberger in reference to James Tobin (1958). Defining y_i to be the i^{th} observation of an environmental exposure, y_i^* to be its corresponding true value, and c to be the censoring point (i.e. the LOD value), the univariate Tobit model can be written as:[25]

$$y_i = \begin{cases} y_i^*, if\ y_i^* \geq c \\ c, otherwise \end{cases}, \qquad i = 1, \ldots, n,$$

where n is the number of observations. The corresponding true value Y_i^* is then modeled as

$$y_i^* = x_i'\beta + \varepsilon_i \text{ , with } \varepsilon_i \sim N(0, \sigma^2),$$

where x_i is the $k \times 1$vector of covariates, β is the $k \times 1$vector of regression coefficients, and $\varepsilon_i$is the normally distributed error term.

The probability density function of Y^*has a form:

$$f_{Y^*}(y^*) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y^* - x_i'\beta)^2}{2\sigma^2}\right).$$

For uncensored observations, the contribution to the likelihood is $L^O = f_{Y^*}(y)$, while the contribution of censored observations is $L^C = \int_{-\infty}^{c} f_{Y^*}(z)\,dz = 1 - \Phi\left(\frac{z - x_i'\beta}{\sigma}\right)$, where Φ is the

cumulative distribution function of a standard normal distribution. Thus, we can write the likelihood function for the univariate Tobit model as:

$$L\left(\beta,\sigma^2 \vee y^{¿}\right)=\prod_{y^{¿}\geq c} L^O \prod_{y^{¿}<c} L^C.$$

The likelihood function of the Tobit model reflects both observed and censored data and thus accounts for the non-ignorable missingness mechanism. The parameters (β,σ^2¿ can be obtained by maximizing the likelihood function or using Bayesian statistics by specifying some prior distributions for the parameters. Once we obtained the estimated parameters ($\hat{\beta},\hat{\sigma}^2$¿, the imputation of the unobserved $y_i^{¿}$ for $y_i^{¿}<c$ can be simulated from a truncated normal distribution with mean $x_i^{'}\hat{\beta}$, variance $\hat{\sigma}^2$, and left-truncation at c.

In the presence of exposure mixtures with multiple censored variables, the univariate Tobit regression can be used together with the Multiple Imputation by Chained Equations (MICE) algorithm.[26] MICE is also called “fully conditional specification” or “sequential regression multiple imputation”. It imputes one variable at a time conditioning on all the other variables. We can apply univariate Tobit regression to impute one exposure at a time conditioning on all the other exposures and other predictors. Once an exposure is imputed, it is used as a predictor in the imputation of other exposures. The imputation continues until all data below the LOD are imputed. This procedure is repeated several cycles. Through each cycle, the imputation is updated taking into consideration the former cycle's information. The final imputed dataset is the set of imputation at the end of the cycles. The entire imputation process is repeated M times to obtain M complete datasets each from the cycles of the chained equation process. The Tobit regression is not available in any existing statistical software implementing the MICE algorithm, but the “mice” package in R allows users to select and write their own imputation model. We wrote a new “tobit” function, which can be used within the “mice” package. The

"tobit" function and a sample code to call "tobit" function from the "mice" package in R is in Appendix A.

2.2.2 Multivariate Tobit Regression (BMTR)

Multiple imputation using joint modeling is another approach to impute multivariate censored data. We consider multivariate Tobit regression, which allows non-ignorable missingness mechanism while accounting for the correlations among mixtures and associations between mixtures and possible predictors via regression. Defining y_{ij} to be the i^{th} observation on the j^{th} environmental exposure, $y_{ij}^{¿}$ to be its corresponding true value, and c_j to be the censoring point (LOD) of the j^{th} environmental exposure, $j=1,\ldots,J$. To simplify the explanation, here we use $J=3$ as an example. The multivariate Tobit model with three dependent variables can be written as:

$$y_{ij}=\begin{cases} y_{ij}^{¿}, if\ y_{ij}^{¿}>c_j \\ c_j, otherwise, \end{cases} \quad i=1,\ldots,n, j=1,2,3,$$

$$y_i^{¿}=\begin{pmatrix} y_{i1}^{¿} \\ y_{i2}^{¿} \\ y_{i3}^{¿} \end{pmatrix} \sim N_3\left[\begin{pmatrix} \beta_1^T x_i \\ \beta_2^T x_i \\ \beta_3^T x_i \end{pmatrix}, \Sigma\right], \text{ with } \Sigma=\begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}, \quad (1)$$

where Σ is the variance-covariance matrix with $(\sigma ¿¿ 1^2, \sigma_2^2, \sigma_3^2)¿$ being residual variances and $(\rho ¿¿ 12, \rho_{13}, \rho_{23})¿$ being correlations between any two exposures. The variance-covariance matrix Σ can also be re-written as

$$\Sigma=\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}.$$

The probability density function of $(Y¿¿1^{¿}, Y¿¿2^{¿}, Y_3^{¿})¿¿$ is

$$f_{Y_1^{¿}, Y_2^{¿}, Y_3^{¿}}(y_1^{¿}, y_2^{¿}, y_3^{¿}) = f_{Y^{¿}}(y^{¿}) = \frac{1}{\sqrt{(2\pi)^3|\Sigma|}} \exp\left(\frac{-1}{2}(y^{¿} - x^T\beta)^T \Sigma^{-1}(y^{¿} - x^T\beta)\right),$$

where $\beta = (\beta_1, \beta_2, \beta_3)$ with β_j being a $k \times 1$vector of regression coefficients for exposure j, and $x = (x_1, x_2, \ldots, x_n)$ with x_i being a $k \times 1$vector of covariates for subject i.

The likelihood function depends on eight possible data patterns determined by LOD conditions of the three exposures. Observed data contribute to the likelihood with their probability density functions but censored data contribute with cumulative distribution functions. Specifically, the contribution of each observation from the eight possible data patterns is as follows,

1) All three exposures are observed: $L^{OOO} = f_{Y^{¿}}(y)$
2) $Y_1^{¿}$ and $Y_2^{¿}$ are observed, $Y_3^{¿}$ is censored: $L^{OOC} = f_{Y_1^{¿}, Y_2^{¿}}(y_1, y_2) \int_{-\infty}^{c_3} f_{Y_3^{¿} \vee Y_1^{¿}, Y_2^{¿}}(z)\, dz$
3) $Y_1^{¿}$ and $Y_3^{¿}$ are observed, $Y_2^{¿}$ is censored: $L^{OCO} = f_{Y_1^{¿}, Y_3^{¿}}(y_1, y_2) \int_{-\infty}^{c_2} f_{Y_2^{¿} \vee Y_1^{¿}, Y_3^{¿}}(z)\, dz$
4) $Y_2^{¿}$ and $Y_3^{¿}$ are observed, $Y_1^{¿}$ is censored: $L^{COO} = f_{Y_2^{¿}, Y_3^{¿}}(y_1, y_2) \int_{-\infty}^{c_1} f_{Y_1^{¿} \vee Y_2^{¿}, Y_3^{¿}}(z)\, dz$
5) $Y_1^{¿}$ is observed, $Y_2^{¿}$ and $Y_3^{¿}$ are censored: $L^{OCC} = f_{Y_1^{¿}}(y_1) \int_{-\infty}^{c_2} \int_{-\infty}^{c_3} f_{Y_2^{¿}, Y_3^{¿} \vee Y_1^{¿}}(z_2, z_3)\, dz_2\, dz_3$
6) $Y_2^{¿}$ is observed, $Y_1^{¿}$ and $Y_3^{¿}$ are censored: $L^{COC} = f_{Y_2^{¿}}(y_2) \int_{-\infty}^{c_1} \int_{-\infty}^{c_3} f_{Y_1^{¿}, Y_3^{¿} \vee Y_2^{¿}}(z_1, z_3)\, dz_1\, dz_3$
7) $Y_3^{¿}$ is observed, $Y_1^{¿}$ and $Y_2^{¿}$ are censored: $L^{CCO} = f_{Y_3^{¿}}(y_3) \int_{-\infty}^{c_1} \int_{-\infty}^{c_2} f_{Y_1^{¿}, Y_2^{¿} \vee Y_3^{¿}}(z_1, z_2)\, dz_1\, dz_2$
8) All three exposures are censored: $L^{CCC} = \int_{-\infty}^{c_1} \int_{-\infty}^{c_2} \int_{-\infty}^{c_3} f_{Y_1^{¿}, Y_2^{¿}, Y_3^{¿}}(z_1, z_2, z_3)\, dz_1\, dz_2\, dz_3.$

In the above equations, $f_{Y_1^{¿}, Y_2^{¿}}(y_1, y_2)$ is used to defined the marginal distribution of $(Y¿¿1^{¿}, Y_2^{¿})¿$, and $f_{Y_3^{¿} \vee Y_1^{¿}, Y_2^{¿}}$ is used to denote the conditional distribution of $Y_3^{¿}$ given $Y_1^{¿}$ and $Y_2^{¿}$. Other distributions are defined similarly. Given the joint distribution of $¿¿, Y_3^{¿}¿$ in model (1), any marginal and conditional distributions are normal and can be easily obtained. Thus, we can write the likelihood function for the multivariate Tobit model as:

$$L(\beta, \Sigma \vee y^{¿}) = \prod_{\substack{y_1^{¿} \geq c_1 \\ y_2^{¿} \geq c_2 \\ y_3^{¿} \geq c_3}} L^{OOO} \prod_{\substack{y_1^{¿} \geq c_1 \\ y_2^{¿} \geq c_2 \\ y_3^{¿} < c_3}} L^{OOC} \prod_{\substack{y_1^{¿} \geq c_1 \\ y_2^{¿} < c_2 \\ y_3^{¿} \geq c_3}} L^{OCO} \prod_{\substack{y_1^{¿} < c_1 \\ y_2^{¿} \geq c_2 \\ y_3^{¿} \geq c_3}} L^{COO} \prod_{\substack{y_1^{¿} \geq c_1 \\ y_2^{¿} < c_2 \\ y_3^{¿} < c_3}} L^{OCC} \prod_{\substack{y_1^{¿} < c_1 \\ y_2^{¿} \geq c_2 \\ y_3^{¿} < c_3}} L^{COC} \prod_{\substack{y_1^{¿} < c_1 \\ y_2^{¿} < c_2 \\ y_3^{¿} \geq c_3}} L^{CCO} \prod_{\substack{y_1^{¿} < c_1 \\ y_2^{¿} < c_2 \\ y_3^{¿} < c_3}} L^{CCC}.$$

To obtain the inference of the parameters in model (1) and thus imputations of censored variables, we use Bayesian statistics by specifying independent weak priors for all parameters. We assume normal prior distributions for β_j and half-normal prior distributions for σ_j, $j = 1,2,3$. For the correlation matrix, we place an LKJ prior distribution on the Cholesky factor L of the correlation matrix with $L L^T = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$.[27] We use Hamiltonian Monte Carlo (HMC) sampler, via RStan,[28] for the Bayesian computation. HMC is a Markov chain Monte Carlo (MCMC) method that uses the derivatives of the density function being sampled to generate efficient transitions spanning the posterior. To check for model convergence, we consider trace plots, the effective sample size, and the Gelman-Rubin diagnostic statistic $\hat{R}$ [29]. After the model converges, draws of $\hat{\beta}^{(d)}$ and $\hat{\Sigma}^{(d)}$ as well as imputations for censored variables $y_i^{¿(d)}$ are then obtained from their posterior distributions, $d = 1, \ldots, D$, with D usually taking a value of 5-10. An R-package "bmvTobit" for fitting Bayesian multivariate Tobit model and imputing censored data below LODs is available on GitHub. A sample code to use "bmvTobit" package in R is in Appendix B.

2.2.3 Analysis of Multiply-Imputed data

The MI generates multiple complete datasets, which are then used for statistical analysis. A complete data method can first be applied to each imputed dataset to generate descriptive statistics (e.g. mean and quantiles) and conduct analytic inference (e.g. regression). The estimates from all the D imputed datasets are then combined using Rubin's Rule.[24] Let $\hat{\theta}_d$ and $\hat{V}_d$ be the estimate and the associated variance for θ (e.g. mean) using the dth imputation $(d=1,\ldots,D)$. The estimate of θ across all D datasets is the average $\bar{\theta}=\frac{1}{D}\sum_{d=1}^{D}\hat{\theta}_d$. The variance of $\bar{\theta}$ is $T=\bar{V}+(1+\frac{1}{D})B$, where $\bar{V}=\frac{1}{D}\sum_{d=1}^{D}\hat{V}_d$ denotes the average within-imputation variance and $B=\frac{1}{D-1}\sum_{d=1}^{D}(\hat{\theta}¿¿d-\bar{\theta})^2¿$ represents the between-imputation variance. When sample size is large, the 95% confidence interval (CI) for θ is $\bar{\theta}\pm t_{.975,\upsilon}\sqrt{T}$, with the degrees of freedom $\upsilon=(D-1)(1+\frac{1}{1+D^{-1}}\frac{\bar{V}}{B})^2$.

2.3 Simulation Study Using the SHFS Data

We conducted a simulation study to compare the performance of MI using the multivariate Tobit regression to the MI using univariate Tobit regression together with MICE and other alternative methods in handling data below the LODs in exposure mixtures. The simulation used the SHFS dataset (n = 1,602) that contains complete measures on DMA, MMA, iAs, and all covariates. Natual log-transformation was conducted on DMA, MMA, iAs, total arsenic, and arsenobetaine. Figure 1(A) gives the bivariate correlation of any two variables in the SHFS, with blue color

denoting positive correlation and red color denoting negative correlation; the darker the color is, the stronger the correlation is. The three arsenic measurements are highly correlated, with correlations ranging from 0.67 to 0.82. Figure 1(B) shows the density plots of DMA, MMA, and iAs after natural-log transformation. The histograms on the top and the right edge of the plot show the marginal distributions of the compounds on the x-axis and y-axis, respectively. Inside the plot is the contour density plot between the two compounds. The arsenic measurements seem to be normally distributed after log-transformation.

We conducted 500 replicates of simulations. In each replicate of simulation, we first randomly selected 500 subjects from the full SHFS dataset (n=1,602). We then created censored data in DMA, MMA, and iAs by using pre-defined LOD levels. Specifically, we considered the following two simulation scenarios with LOD levels for DMA, MMA, and iAs defined using population percentiles of the full SHFS dataset:

Scenario S1: approximately 10% DMA, 20% MMA, 30% iAs are below LOD, with LOD levels defined using the population 10^{th}, 20^{th}, and 30^{th} percentile of DMA, MMA, and iAs, respectively.

Scenario S2: approximately 10% DMA, 40% MMA, 60% iAs are below LOD, with LOD levels defined using the population 10^{th}, 40^{th}, and 60^{th} percentile of DMA, MMA, and iAs, respectively.

In each dataset with censored observations, we applied the following four methods to handle the data below LODs, including

(1) substitution with LOD/$\sqrt{2}$.

(2) **MICE-OLS**: MI using MICE, implemented using IVEware package called from R.[30] Ordinary least squares (OLS) regression was used in the imputation of each arsenic compound with imputed values truncated by LODs.

(3) **MICE-UTR**: MI using MICE, implemented using the "mice" package in R with our new "tobit" function. Univariate Tobit regression (UTR) was used in the imputation of each arsenic compound with imputed values truncated by LODs.

(4) **BMTR**: MI using the Bayesian multivariate Tobit regression (BMTR), implemented using the "bmvTobit" package in R with imputed values below LODs.

The imputation models in (2)-(4) regressed on 9 covariates, including age, gender, BMI, urinary creatinine, urinary arsenobetaine, urinary total arsenic, daily rice intake, daily juice intake, and daily wine/beer intake, and each model generated 50 imputations for data below the LODs.

The imputed data sets were then used to estimate 5^{th} percentile, 1^{st} quartile, median (only for Scenario S2), and geometric mean of MMA, DMA, and iAs. An ordinary least squares regression on age was also conducted for each log-transformed arsenic specie. We used violin plots to show the estimates of each method across the 500 simulations and compare to the estimates using the full SHFS dataset of 1,602 observations (the benchmarck measure if no LOD issues). A violin plot is similar to a box plot, but shows the full distribution of all the estimates from the 500 simulations. Wider sections of the violin plot represent a higher probability that estimates take on the given value. We also calculated the coverage rate of 95% CI associated with each method, with the coverage rate defined as proportion of the replicates of simulations with the 95% CI covers the benmarck measures calculated using the full SHFS dataset.

2.4 Application Using the NHANES 2013-2014

We applied the BMTR method to the NHANES 2013-2014 data for estimating 5^{th} percentile, 1^{st} quartile, and geometric mean of urinary DMA, MMA, and iAs, and compared to the methods using LOD/$\sqrt{2}$, MICE-OLS, and MICE-UTR. The point estimate and 95% CI of the regression

slope of log-transformed urinary arsenic measure on age were also computed for each arsenic specie. All the statisyical analyses accounted for the complex survey design in NHANES.

3. Results

3.1 Simulation Results

Figures 2 and 3 show the violin plots of the estimates and the bart charts of 95% CI coverage rates, respectively, for Scenario S1, in estimating 5th percentile, 1st quartile, geometric mean, and regression slope on age. In scenario S1, where the proportion of censored observations are small for all three arsenics, MICE-UTR and MTR perform similarly and both yield estimates that are centered around the true values, except for 5th percentile and 1st quartile of iAs (Figure 2). The LOD/$\sqrt{2}$ leads to a constant estimate for 5th percentile of DMA and MMA, and for both 5th percentile and the 1st quartile of iAs. The MICE-OLS tends to over-estimate mean, regression slope, and lower percentiles, but yields less biased estimate of 1st quartile of iAs than MICE-UTR and MTR, where the quantile of interest is close to the LOD level. The over-estimate of MICE-OLS in lower percentiles and mean can be explained by the limitations of the MICE-OLS in only using data above LODs for estimating the imputation models. The 95% CIs using LOD/$\sqrt{2}$ or MICE-OLS to fill in data below LODs yield poor coverage in estimating 5th percentile of all three arsenics and 1st quartile or iAs (Figure 3). The LOD/$\sqrt{2}$ aproach also yields lower than nominal level coverage rate in estimating slope. The 95% CIs of the MICE-UTR and BMTR yield good coverage in estimating 1st quartile, mean, and slope, but is slightly under-coveraged in estimating the 5th percentile of DMA and iAs.

In scenario S2, the proportion of censored observations for DMA is the same as scenario S1, but the proportion of censored observations increases from 20% to 40% for MMA and from

30% to 60% for iAs. Figures 4 and 5 show that now MICE-UTR performs poorly with most of the quantities for MMA and iAs, with coverage rates of the 95% CIs being close to 0 in estimating quantiles and means of MMA and iAs. BMTR performs best among all the methods in all the estimation. However, the median estimate of iAs is negatively biased although the 95% CI still covers the true value for most of the simulations. In addition, the 5^{th} percentile estimate of iAs is slightly over-estimated with the 95% CI coverage rate lower than the nominal level. This is consistent with the findings in Scenario S1 where BMTR performs well with small bias except for quantiles that are close to the LOD levels and the 95% CI can have slightly lower coverage rate in estimating 5^{th} percentile. Even given this limitations, the BMTR still outperforms the alternative methods.

3.2 Analysis Results of the NHANES 2013-2014

Figure 6 shows missing patterns, correlation matrix, and histograms and contour plots of log-transformed DMA, MMA, and iAs. The NHANES 2013-2014 data had higher proportion of data below the LODs for all three arsenic compounds than SHFS. Specifically, 24.3%, 30.4%, and 28% of individuals had DMA, MMA, and iAs measures below LOD, respectively. Among the 2,126 observations, 1189 (56%) had complete measurements in all three arsenic compounds, 397 (19%) had one compound measurement below LOD, 261 (12%) had two compound measurements below LOD, and 279 (13%) had all three compounds with measurements below LOD. The correlations between DMA, MMA, and iAs among measurements above the LODs were lower in NHANES 2013-2014 as compared to SHFS, ranging from 0.45 to 0.55. Since arsenobetaine and total arsenic had 56.3% and 0% of measurements below the LOD, and their correlation was as high as 0.96 among those with arsenobetaine measurements above the LOD,

we excluded arsenobetaine from all the imputation models. Finally, the histograms are not symmetric even after the log-transformation. The non-symmetric histograms could be explained by the higher proportions of data below the LODs in NHANES.

Figure 7 shows the estimates and 95% CIs of 5th percentile, 1st quartile, geometric mean, and regression slope for DMA, MMA, and iAs. Substitution using LOD/$\sqrt{2}$ yields a constant estimate of 5th percentile and a very wide 95% CI of 1st quartile for DMA, and a constant estimate of 5th percentile and 1st quartile for both MMA and iAs. The MICE-OLS results in larger estimates of 5th percentile, 1st quartile, and geometric mean than MICE-UTR and MTR, especially in the 5th percentile. This is because the MICE-OLS only uses the data above the LODs for estimating parameters in the imputation models. As a consequence, the data below the LODs are imputed with larger values and narrower 95% CIs using MICE-OLS. MICE-UTR and BMTR yield similar estimates in all quantities, although the BMTR tends to have slightly smaller estimates in the percentiles and means than MICE-UTR. This is not surprising, as the proportions of data below LODs range from 24% to 30% and MICE-UTR and MTR tend to perform similarly when proportions of data below LODs are not high as shown in the simulation Scenario S1. The regression slope estimates are similar across methods, except for iAs where the MICE-OLS results in a larger slope estimate than LOD/$\sqrt{2}$, followed by MICE-UTR and BMTR.

4. Discussion

We consider multiple imputation (MI) for data below LODs in exposure mixtures. To account for the correlations between exposure mixtures and the nonignorable missing data mechanism for data below the LOD in the imputation models, we propose a MI approach using Bayesian multivariate Tobit regression (BMTR). In addition, we extended the existing MICE algorithm in

statistical software, such as R, to allow the use of univariate Tobit regression as the imputation model for one exposure at a time in a sequential manner. We named this MICE-UTR.

Using simulation studies, we showed that the BMTR and MICE-UTR perform similarly when proportions of data below the LODs are small to moderate, and both outperform the substitution approach using LOD/$\sqrt{2}$ and the MICE-OLS approach that fits ordinary least squares regression in the MICE iterations by assuming data are missing at random. However, when proportions of data below the LODs are large, the simulation shows that the imputations using BMTR lead to better inference of percentiles, means and regression slopes than the imputations using MICE-UTR. The improvement of BMTR over the alternative methods becomes more pronounced for the exposure with a larger proportion of data below the LOD. Therefore, BMTR imputation method is recommended for the imputation of data below the LODs in exposure mixtures with moderate or high proportions of censored observations. Although BMTR performs better than the other methods in general, its estimation for quantiles that are close to the LOD levels can be biased even though the 95% CI still provides adequate confidence coverage.

The BMTR and the MICE-UTR methods need to be used with cautions. The BMRT assumes a multivariate normal distribution for the exposure mixtures conditional on covariates and the MICE-UTR assumes a univariate normal distribution for each exposure given the other exposures and covariates. Density plots and contour plots for any two exposures can be checked on the model residuals to assess the assumptions. Transformation on the exposure mixtures may be needed before model fitting to make the assumptions become more reasonable.

The BMTR is more computational intensive than the MICE-UTR and the other methods. The computation time could depend on sample size, number of exposure mixtures, proportions of data below the LODs for each exposure, and missing patterns across all exposure mixtures.

Reference

1. Hornung RW, Reed LD. Estimation of Average Concentration in the Presence of Nondetectable Values. *Applied Occupational and Environmental Hygiene*. 1990;5(1):46-51. doi:10.1080/1047322X.1990.10389587

2. Antweiler RC, Taylor HE. Evaluation of Statistical Treatments of Left-Censored Environmental Data using Coincident Uncensored Data Sets: I. Summary Statistics. *Environ Sci Technol*. 2008;42(10):3732-3738. doi:10.1021/es071301c

3. Helsel. Much ado about next to nothing: incorporating nondetects in science. *Ann Occup Hyg*. 2010;54(3):257-262. doi:10.1093/annhyg/mep092

4. Helsel. *Statistics for Censored Environmental Data Using Minitab and R*. Vol 77. John Wiley & Sons; 2011.

5. Huybrechts T, Thas O, Dewulf J, Van Langenhove H. How to estimate moments and quantiles of environmental data sets with non-detected observations? A case study on volatile organic compounds in marine water samples. *Journal of Chromatography A*. 2002;975(1):123-133. doi:10.1016/S0021-9673(02)01327-4

6. Helsel, Cohn TA. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour Res*. 1988;24(12):1997-2004. doi:10.1029/WR024i012p01997

7. Helsel. Less than obvious - statistical treatment of data below the detection limit. *Environ Sci Technol*. 1990;24(12):1766-1774. doi:10.1021/es00082a001

8. US EPA O. Guidance for Data Quality Assessment. US EPA. Published June 17, 2015. Accessed March 9, 2021. https://www.epa.gov/quality/guidance-data-quality-assessment

9. Gillespie BW, Chen Q, Reichert H, et al. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. *Epidemiology*. 2010;21 Suppl 4:S64-70. doi:10.1097/EDE.0b013e3181ce9f08

10. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581-592.

11. Lockwood JR, Schervish MJ, Gurian PL, Small MJ. Analysis of Contaminant Co-Occurrence in Community Water Systems. *Journal of the American Statistical Association*. 2004;99(465):45-56. doi:10.1198/016214504000000061

12. Lockwood JR, Schervish MJ. MCMC Strategies for Computing Bayesian Predictive Densities for Censored Multivariate Data. *Journal of Computational and Graphical Statistics*. 2005;14(2):395-414. doi:10.1198/106186005X47967

13. Chen H, Quandt SA, Grzywacz JG, Arcury TA. A distribution-based multiple imputation method for handling bivariate pesticide data with values below the limit of detection. *Environ Health Perspect*. 2011;119(3):351-356. doi:10.1289/ehp.1002124

14. Francis RA, Small MJ, VanBriesen JM. Multivariate distributions of disinfection by-products in chlorinated drinking water. *Water Research*. 2009;43(14):3453-3468. doi:10.1016/j.watres.2009.05.008

15. Hopke PK, Liu C, Rubin DB. Multiple Imputation for Multivariate Data with Missing and Below-Threshold Measurements: Time-Series Concentrations of Pollutants in the Arctic. *Biometrics*. 2001;57(1):22-33. doi:10.1111/j.0006-341X.2001.00022.x

16. Navas-Acien Ana, Umans Jason G., Howard Barbara V., et al. Urine Arsenic Concentrations and Species Excretion Patterns in American Indian Communities Over a 10-year Period: The Strong Heart Study. *Environmental Health Perspectives*. 2009;117(9):1428-1433. doi:10.1289/ehp.0800509

17. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Vol 793. John Wiley & Sons; 2019.

18. Stoddard RA, Gee JE, Wilkins PP, McCaustland K, Hoffmaster AR. Detection of pathogenic Leptospira spp. through TaqMan polymerase chain reaction targeting the LipL32 gene. *Diagn Microbiol Infect Dis*. 2009;64(3):247-255. doi:10.1016/j.diagmicrobio.2009.03.014

19. Lee ET, Welty TK, Fabsitz R, et al. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *Am J Epidemiol*. 1990;132(6):1141-1155. doi:10.1093/oxfordjournals.aje.a115757

20. North KE, MacCluer JW, Devereux RB, et al. Heritability of carotid artery structure and function: the Strong Heart Family Study. *Arterioscler Thromb Vasc Biol*. 2002;22(10):1698-1703. doi:10.1161/01.atv.0000032656.91352.5e

21. Scheer J, Findenig S, Goessler W, et al. Arsenic species and selected metals in human urine: validation of HPLC/ICPMS and ICPMS procedures for a long-term population-based epidemiological study. *Anal Methods*. 2012;4(2):406-413. doi:10.1039/C2AY05638K

22. Nigra AE, Olmedo P, Grau-Perez M, et al. Dietary determinants of inorganic arsenic exposure in the Strong Heart Family Study. *Environ Res*. 2019;177:108616. doi:10.1016/j.envres.2019.108616

23. Nigra AE, Nachman KE, Love DC, Grau-Perez M, Navas-Acien A. Poultry Consumption and Arsenic Exposure in the U.S. Population. *Environ Health Perspect*. 2017;125(3):370-377. doi:10.1289/EHP351

24. Rubin DB. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In: *Proceedings of the Survey Research Methods Section of the American Statistical Association*. Vol 1. American Statistical Association; 1978:20-34.

25. Blundell R, Meghir C. Bivariate alternatives to the Tobit model. *Journal of Econometrics*. 1987;34(1-2):179-200. doi:10.1016/0304-4076(87)90072-8

26. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 2011;45(1):1-67. doi:10.18637/jss.v045.i03

27. Lewandowski D, Kurowicka D, Joe H. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*. 2009;100(9):1989-2001. doi:10.1016/j.jmva.2009.04.008

28. Carpenter B, Gelman A, Hoffman MD, et al. Stan: A Probabilistic Programming Language. *Journal of Statistical Software*. 2017;76(1):1-32. doi:10.18637/jss.v076.i01

29. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 0 ed. Chapman and Hall/CRC; 2013. doi:10.1201/b16018

30. Raghunathan T, Berglund PA, Solenberger PW. *Multiple Imputation in Practice: With Examples Using IVEware*. CRC Press; 2018.

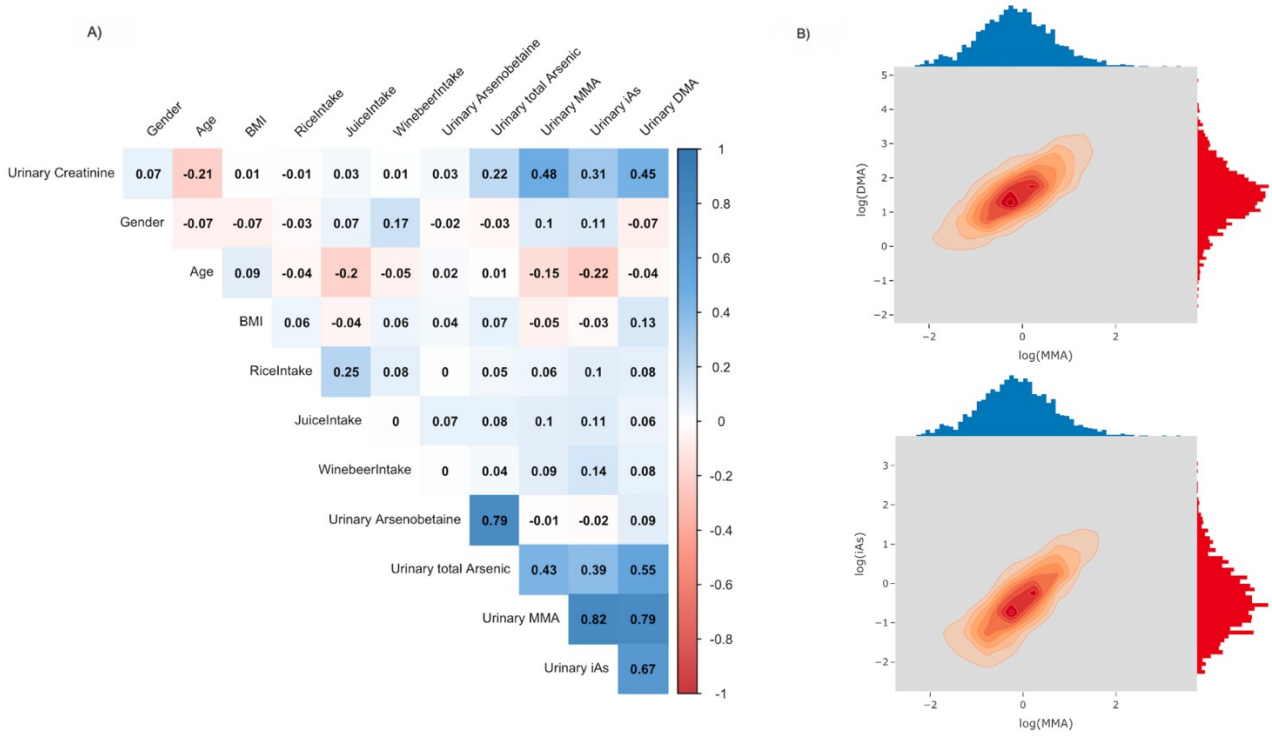

Figure 1. **Strong Heart Family Study Arsenic data information.** A) Correlation matrix between covariates: red color for negative correlations and blue color for positive correlation; the darker the color is, the stronger the correlation is. B) Density plots of DMA, MMA, and iAs after log-transformation: The marginal histogram plot shows the marginal density of each arsenic variable on the top and on the right edge of the plot. Inside the plot is the contour density plot between the two variables.

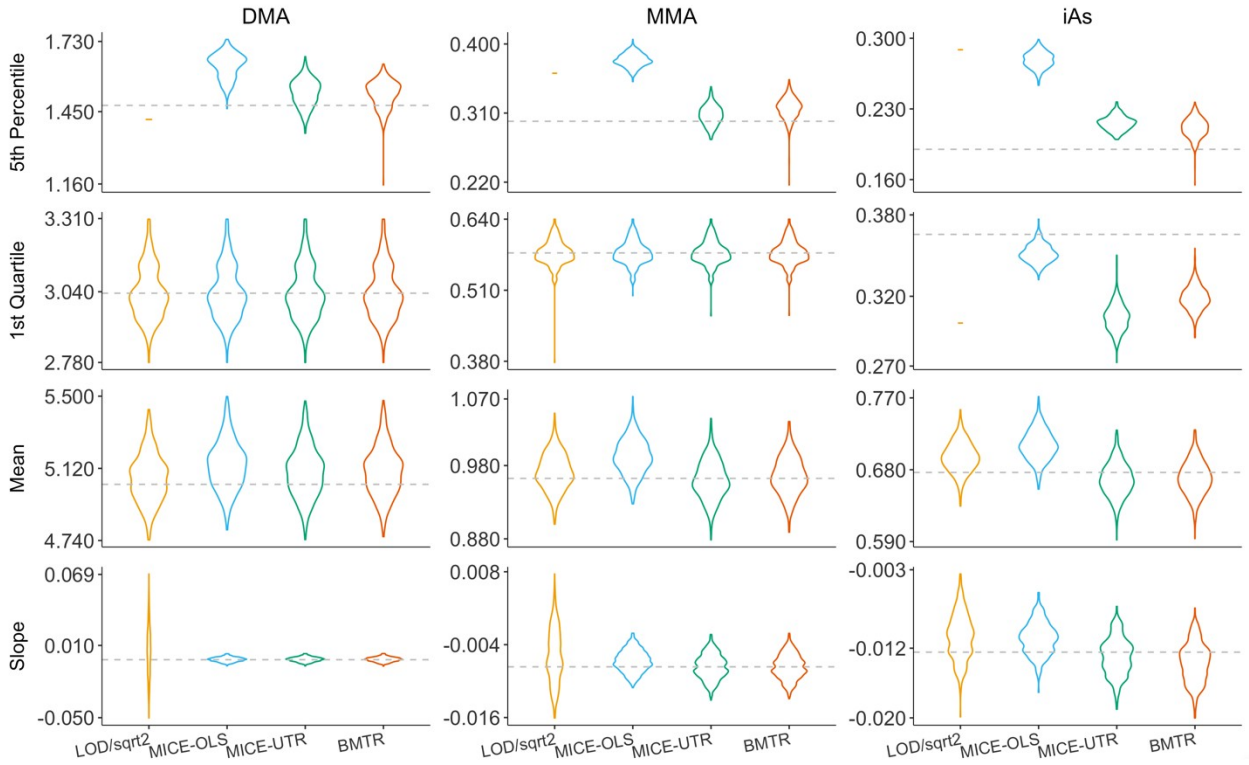

Figure 2. **Simulation results in Scenario S1**: 10% missingness in DMA, 20% missingness in MMA, 30% missingness in iAs; Each plot shows the violin plots of the estimates from the 500 replicates of simulation using LOD/$\sqrt{2}$, MICE-OLS, MICE-UTR, and BMTR. The dash line represents the true parameter using the entire sample of 1,602 subjects.

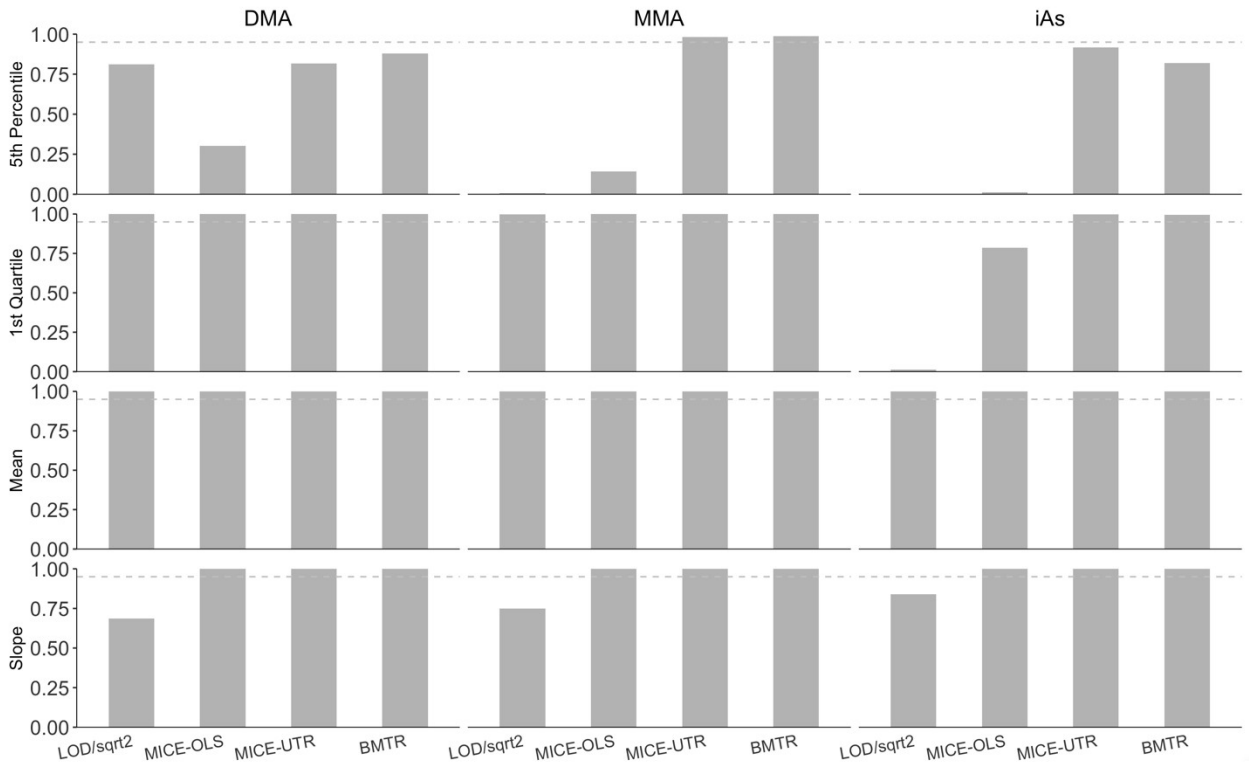

Figure 3. **Simulation results in Scenario 1 coverage rate of 95% CI**: 10% missingness in DMA, 20% missingness in MMA, 30% missingness in iAs;

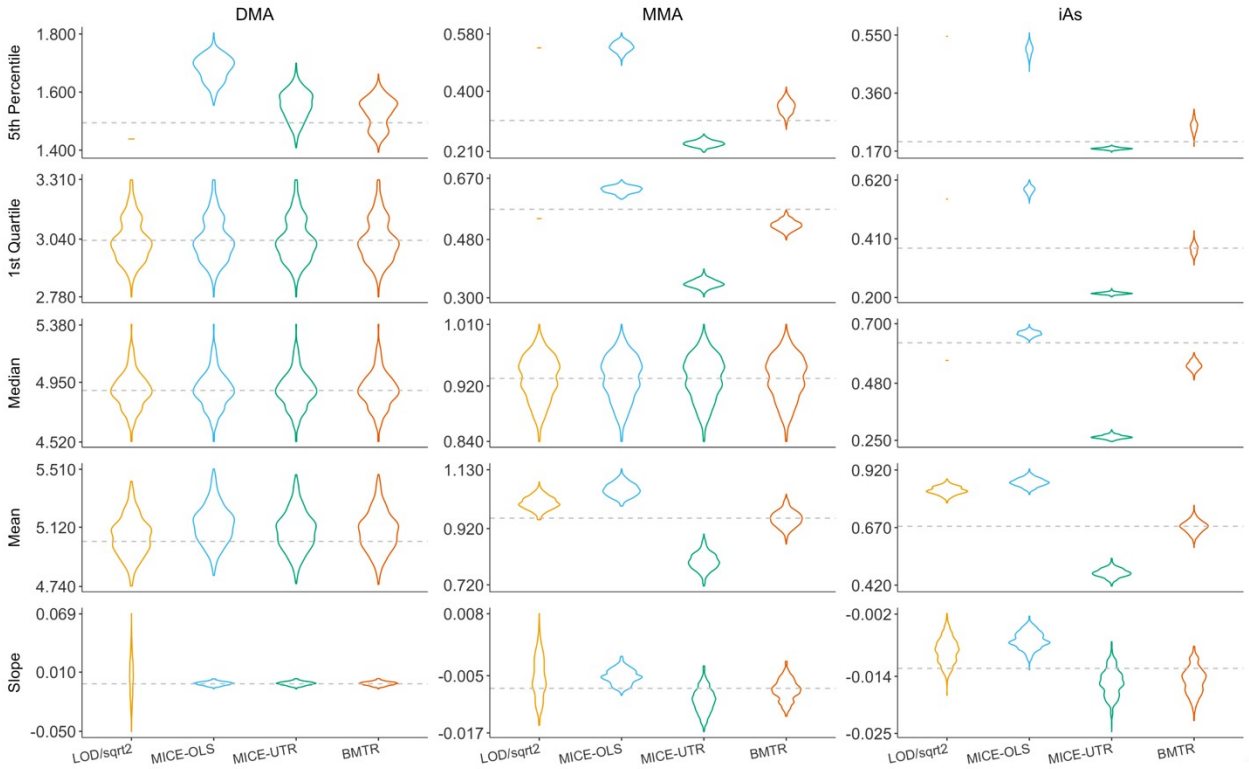

Figure 4. **Simulation results in Scenario S2**: 10% missingness in DMA, 40% missingness in MMA, 60% missingness in iAs; Each plot shows the violin plots of the estimates from the 500 replicates of simulation using LOD/$\sqrt{2}$, MICE-OLS, MICE-UTR, and BMTR. The dash line represents the true parameter using the entire sample of 1,602 subjects.

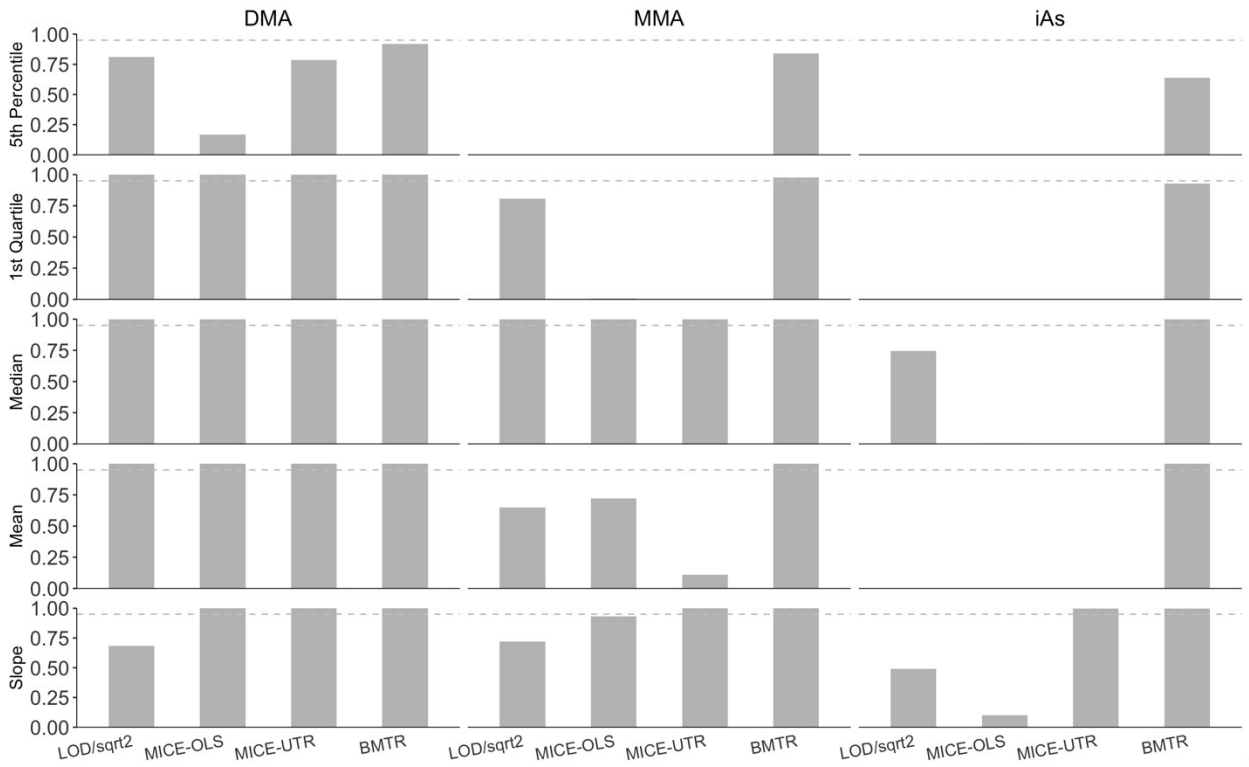

Figure 5. **Simulation results in Scenario 2 coverage rate of 95% CI**: 10% missingness in DMA, 40% missingness in MMA, 60% missingness in iAs;

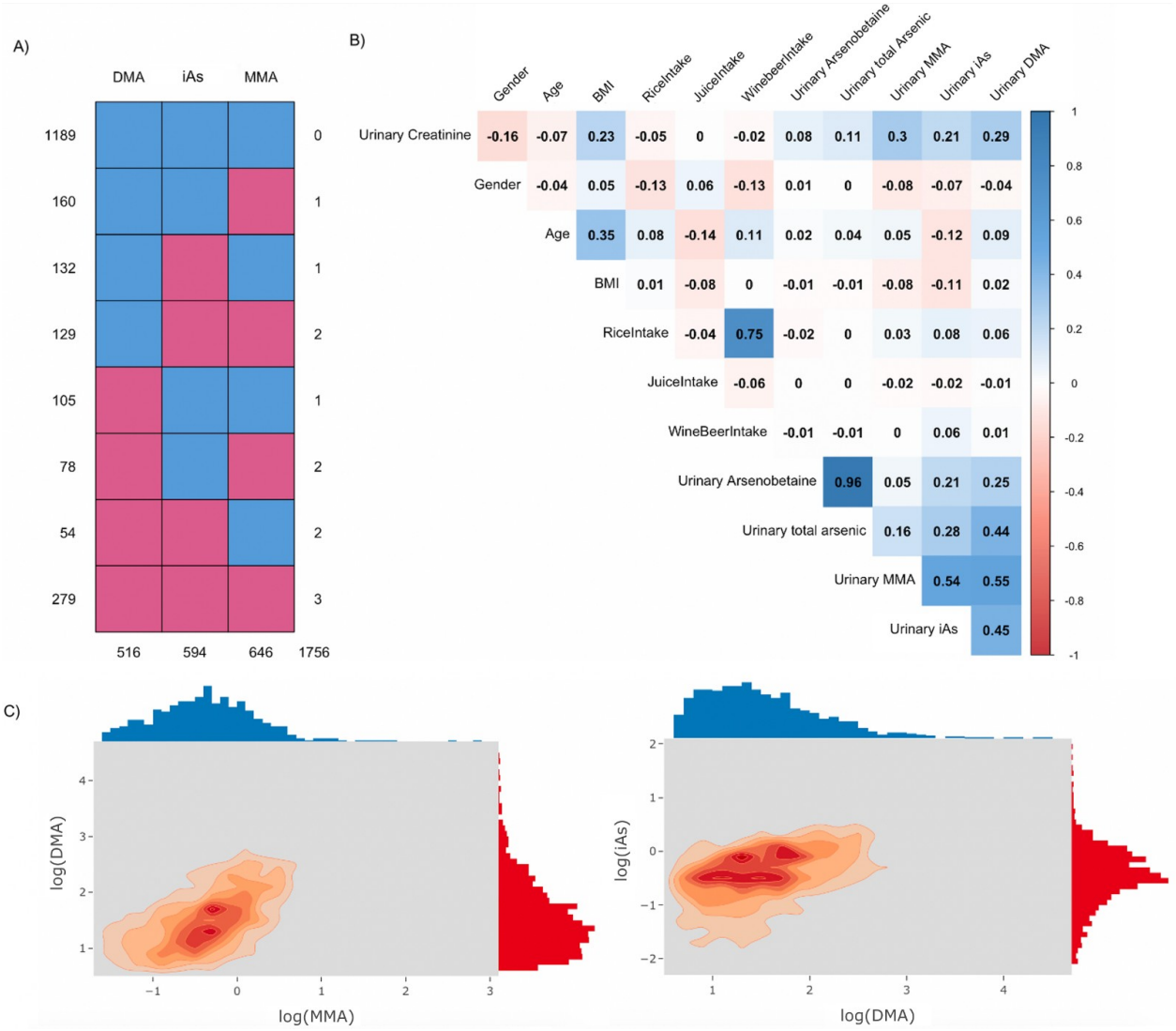

Figure 6. **NHANES 2013-2014 Arsenic application.** A) Missing patterns: each row represents one missing pattern with blue denoting observed and red denoting missing. B) Correlation matrix between covariates: red color for negative correlations and blue color for positive correlation; the darker the color is, the stronger the correlation is. C) Density plots of DMA, MMA, and iAs after log-transformation: The marginal histogram plot shows the marginal density of each arsenic variable on the top and on the right edge of the plot. Inside the plot is the contour density plot between the two variables.

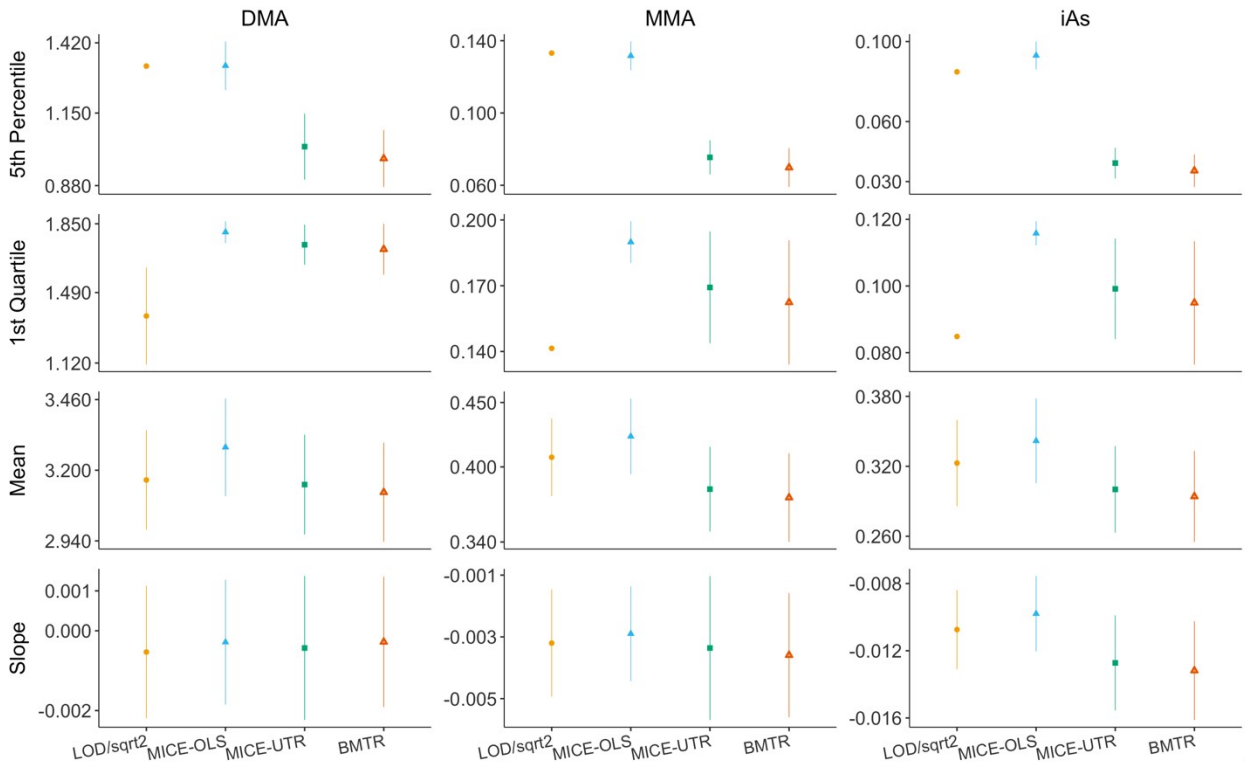

Figure 7. **NHANES 2013-2014 Arsenic data application results**: Each plot shows the estimates and 95% CIs of LOD/$\sqrt{2}$, MICE-OLS, MICE-UTR, and BMTR. The slope is the regression coefficient of the log transformed arsenic variable on age.