ARTICLE

Compare the marginal effects for environmental exposure and biomonitoring data with repeated measurements and values below the limit of detection

I-Chen Chen[1]✉, Stephen J. Bertke[1] and Cheryl Fairfield Estill[1]

BACKGROUND: Environmental exposure and biomonitoring data with repeated measurements from environmental and occupational studies are commonly right-skewed and in the presence of limits of detection (LOD). However, existing model has not been discussed for small-sample properties and highly skewed data with non-detects and repeated measurements.
OBJECTIVE: Marginal modeling provides an alternative to analyzing longitudinal and cluster data, in which the parameter interpretations are with respect to marginal or population-averaged means.
METHODS: We outlined the theories of three marginal models, i.e., generalized estimating equations (GEE), quadratic inference functions (QIF), and generalized method of moments (GMM). With these approaches, we proposed to incorporate the fill-in methods, including single and multiple value imputation techniques, such that any measurements less than the limit of detection are assigned values.
RESULTS: We demonstrated that the GEE method works well in terms of estimating the regression parameters in small sample sizes, while the QIF and GMM outperform in large-sample settings, as parameter estimates are consistent and have relatively smaller mean squared error. No specific fill-in method can be deemed superior as each has its own merits.
IMPACT:

- Marginal modeling is firstly employed to analyze repeated measures data with non-detects, in which only the mean structure needs to be correctly provided to obtain consistent parameter estimates. After replacing non-detects through substitution methods and utilizing small-sample bias corrections, in a simulation study we found that the estimating approaches used in the marginal models have corresponding advantages under a wide range of sample sizes. We also applied the models to longitudinal and cluster working examples.

Keywords: Marginal analysis; Left censoring; Right skewness; Limit of detection; Repeated measures; Environmental exposure

Journal of Exposure Science & Environmental Epidemiology (2024) 34:1018–1027; https://doi.org/10.1038/s41370-024-00640-7

INTRODUCTION

In environmental and occupational studies, repeated concentration measurements not detected or falling below limits of detection (LOD) of laboratory instruments are called left-censored repeated measures data. The unquantified non-detects are generally low-level concentrations with values between zero and LOD. Statistical models continue to arise for industrial hygienists to analyze left-censored environmental exposure and biomonitoring data with repeated measures in cluster and longitudinal studies because the estimation of the effect of exposure on risk of disease and the importance of within- and between-worker variability in occupational exposure have been increasingly acknowledged. Analytical results from laboratories, environmental contaminants, and occupational exposures, e.g., the concentration of an analyte in a biological urine or serum sample, or an environmental hand wipe or personal breathing zone air sample, are often subject to non-detects or left censoring and the data are skewed to the right. These measurements are usually collected from the same subject or the same study site. In such cases, statistical modeling of exposure and biomonitoring data can be complicated when repeated measurements are collected in a cluster or longitudinal study.

Statistical methods have been continuously proposed to analyze left-censored data sets. The substitution or single value imputation method, e.g., assigning a value (LOD/2 or LOD/$\sqrt{2}$) [1, 2] for measurements less than the LOD, is commonly adopted by industrial hygienists. Unfortunately, there is no unique replaced value for this substitution or single value imputation method, and

[1]Division of Field Studies and Engineering, National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, OH, USA.
✉email: okv0@cdc.gov

Received: 14 March 2023 Revised: 3 January 2024 Accepted: 4 January 2024
Published online: 22 January 2024

regression parameter estimation of the substitution method can be biased for high censoring proportion [1]. This substitution approach is also not advisable unless less than 10% of values are below the LOD [3]. Comparatively, the multiple random value imputation technique, e.g., creating imputed values based on throughout scatter of the dataset, has been advocated [3–5]. The use of a maximum likelihood (ML) estimation approach has also been shown to outperform other methods when the working model is well specified and the sample size is large [6–8, 20], but bias and imprecision are expected with the ML approach when sample size is small, i.e., fewer than 50 detectable values, and/or censoring proportion is high, even though the distribution is correctly specified [7, 9, 10]. In addition, the ML method based on lognormal and Weibull distributions generates poor estimates when the true data distribution is mis-specified [11]. Recently, the β-substitution method deriving the calculation of a β factor based on the uncensored data was presented to produce results comparable to the ML method, even when sample sizes were less than 20. Through a simulation study, this method has smaller biases and improved root mean squared errors relative to the LOD/2 and LOD/$\sqrt{2}$ [12]. Another multiple imputation method utilized the actual data distribution to generate the relatively conjunct and ordered values has also been added to the approaches for handling left censoring [13, 14].

The use of linear mixed effects models incorporating ML estimation method for left-censored data with repeated measurements has been discussed for modeling log-normal data in longitudinal infectious disease studies, in which correlation among measures is modeled using a random effect [15–17]. Nevertheless, a correctly specified distribution being assumed for the random effect is practically unknown. Another study recommended the use of mixed effects model for exposure and biomonitoring data with repeated measures while accounting for different levels of censoring [18]. Existing literature also assumes the data with correlated outcomes were log-normally distributed and transformed using a natural logarithm [18, 19]. However, the ML method can result in bias and imprecision in small-sample data even when distributional assumptions are met [20].

In contrast with random effects modeling, marginal modeling can be used as an alternative to analyzing small-sample data with repeated measurements, in which the use of simpler working correlation structures requires less nuisance covariance parameters to be estimated. Generalized estimating equations (GEE) are a special type of the marginal analysis. As long as mean structure is correctly specified, consistent regression parameter estimates can often be obtained under the assumption of a mis-specified working correlation structure [21]. This property also enables marginal modeling to utilize data with highly right-skewed outcome distributions. Furthermore, accurately modeling of the correlation structure can improve estimation efficiency, i.e., smaller standard errors (SEs) of regression parameters [22]. Another method of marginal analysis discussed in this manuscript is the generalized method of moments (GMM) [23], which takes advantage of all estimating equations and has widely developed in econometrics. Additionally, based on GEE and GMM methods, the quadratic inference functions (QIF) method [24] has been proposed to improve efficiency of estimators when the working correlation structure is mis-specified and in large-sample settings. This method rewrites GEE as a linear combination of sets of unbiased estimating equations for representing the inverse of the working correlation matrix. These marginal models are promptly available and regularly used for quantifying repeated measures data without censoring; however, they have not been carried out for data with repeated measurements and non-detects.

In this manuscript, we outline the theories of three marginal models, i.e., GEE, GMM, and QIF, and with these approaches, we propose to incorporate the fill-in methods, including single and multiple value imputation techniques, and β-substitution method such that any measurements less than the LOD are assigned values. We also will consider small-sample bias corrections for the empirical covariance estimators of regression parameter estimates [25–31]. Therefore, the resulting approaches will have the potential to perform better than the existing methods in small-sample settings. Secondly, we conduct a simulation study to compare the proposed methods and evaluate how well they estimate regression parameters for exposure and biomonitoring data with repeated measurements under a range of sample sizes and LOD proportions. Finally, we illustrate the proposed methods using a longitudinal chlorpyrifos exposure dataset and a cluster flame retardant biomonitoring dataset.

METHODS

In this section, we described the marginal models popularly used for analyzing repeated measures data without censoring, the substitution approaches proposed for filling in left-censored data without repeated measures, and the proposed methods for data with repeated measures and non-detects. With these proposed approaches, a simulation study will be presented later to examine the validity of inference and two real-world motivating applications will be demonstrated.

In the longitudinal application, the data were collected by the National Institute for Occupational Safety and Health (NIOSH), in which termite control workers who utilized chlorpyrifos-containing termiticides to commercial and residential structures in North Carolina in 1998 [32]. Thirty-seven male applicators participated in the study (number of independent subjects or N is 37). A total of 184 full-shift breathing zone air samples for determination of exposure levels of chlorpyrifos were measured from each applicator on consecutive days during a five-day workweek (number of repeated measures/time points or M is 5). Only one applicator had four-day air samples (M is 4). Here the correlation between any two time points decreases as the time lag increases. In the cluster application which was also conducted by NIOSH, four nail salons (number of independent subjects) located in the San Francisco area in California in 2016 were recruited [33]. Each nail salon had three workers (number of repeated measures/cluster sizes) and a total of twelve workers on site were asked to participate. Workers were required to provide spot urine samples at the workplace prior to their first-day shift and after their second-day shift for determination of biomonitoring level of diphenyl phosphate (DPhP), a metabolite of triphenyl phosphate (TPhP) which is one of the commonly used organophosphorus flame retardants. Repeated measures from the same company are typically positively correlated, which must be accounted for when performing data analysis. The measures are equally correlated because no ordering occurs with the participants within the company or industry, hence, the outcomes should be equally correlated. Note that unbalanced repeated measures (cluster sizes or time points) are permitted with the discussed study designs.

Marginal models

Generalized estimating equations (GEE) provide consistent regression parameter estimates as long as the mean structure is assumed to be correctly specified. Although the data analyst is required to incorporate a working correlation structure within the GEE, the structure does not need to be correctly clarified. However, accurately modeling this structure has the impact on improving estimation, i.e., reduce SEs of regression parameters [22]. We denote the observed exposure outcome for the ith subject as $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iM}]^T$, which connects to a marginal mean given by $E(\boldsymbol{Y}_i|\boldsymbol{X}_i) = \boldsymbol{\mu}_i$. The marginal mean is linked to independent variables via a function, f, such that $f(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta}$, where $\mathbf{x}_{ij} = [1, \mathrm{x}_{1ij}, \ldots, \mathrm{x}_{pij}]^T$ and $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_p]^T$, $i = 1, \ldots, N$; $j = 1, \ldots, M$. The working covariance matrix for $\boldsymbol{Y}_i$ is $\boldsymbol{V}_i = \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i\boldsymbol{A}_i^{1/2}$, diagonal matrix $\boldsymbol{A}_i = \mathrm{diag}[\phi\upsilon(\mu_{i1}), \ldots, \phi\upsilon(\mu_{iM})]$ represents marginal variances, including a scale parameter, ϕ, assuming common dispersion and a known function, $\upsilon(.)$, and $\boldsymbol{R}_i$ is a symmetric positive definite working correlation matrix. Let $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^T$ and incorporate the GEE approach into marginal model, the estimates of regression parameters, $\hat{\boldsymbol{\beta}}_{\boldsymbol{GEE}}$, can be obtained by iteratively solving

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \quad (1)$$

Generalized Method of Moments (GMM) method constructs a vector, $\mathbf{g}_i(\boldsymbol{\beta})$, consisted of all moment conditions from ith subject, $i = 1, \ldots, N$, corresponding

to the $p+1$ parameters' estimation such that $E[\mathbf{g}_i(\boldsymbol{\beta})] = \mathbf{0}$. The vector is created by stacking all M^2 valid moments according to each parameter so that the maximum length of $\mathbf{g}_i$ $(\boldsymbol{\beta})$ is $M^2 \times (p+1)$. The GMM estimator, $\hat{\boldsymbol{\beta}}_{\boldsymbol{GMM}}$, derived by minimizing the quadratic form, $N\bar{\mathbf{g}}_{\boldsymbol{N}}^{\boldsymbol{T}}(\boldsymbol{\beta})\boldsymbol{C}_{\boldsymbol{N}}^{-1}(\boldsymbol{\beta})\bar{\mathbf{g}}_{\boldsymbol{N}}(\boldsymbol{\beta})$, asymptotically resolves the estimating equations $N\dot{\mathbf{g}}_{\boldsymbol{N}}^{\boldsymbol{T}}(\boldsymbol{\beta})\boldsymbol{C}_{\boldsymbol{N}}^{-1}(\boldsymbol{\beta})\bar{\mathbf{g}}_{\boldsymbol{N}}(\boldsymbol{\beta}) = \mathbf{0}$, where $\dot{\mathbf{g}}_{\boldsymbol{N}}(\boldsymbol{\beta}) = E[\partial\bar{\mathbf{g}}_{\boldsymbol{N}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T]$, $\bar{\mathbf{g}}_{\boldsymbol{N}}(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta})$, and $\boldsymbol{C}_{\boldsymbol{N}}(\boldsymbol{\beta}) = (1/N)\sum_{i=1}^{N}\mathbf{g}_i(\boldsymbol{\beta})\mathbf{g}_i^{\boldsymbol{T}}(\boldsymbol{\beta})$ is an empirical covariance matrix [23].

Quadratic inference functions (QIF) method rewrites GEE as a linear combination of k sets of unbiased estimating equations through correlation structures such that $\boldsymbol{R}_i^{-1} \approx \sum_{r=1}^{k} a_{ri}\boldsymbol{M}_{ri}$. With this approach, $\boldsymbol{M}_{ri}$, $r = 1, \ldots, k$; $i = 1, \ldots, N$, are known basis matrices and a_{ri} are functions of correlation parameters. These estimating equations can then be optimally, linearly stacked through the GMM method [24]. Two basis matrices are typically employed for exchangeable and first-order autoregressive (AR-1) working correlation structures used in cluster and longitudinal studies, correspondingly. For both structures, $\boldsymbol{M}_{1i}$ is an identity matrix. $\mathbf{M}_{2i}$ is a matrix with 0 on the diagonal and 1 elsewhere for exchangeable structure, and with 1 on the sub-diagonal and 0 elsewhere for AR-1 structure.

The robust and empirical sandwich estimator of marginal models increasingly used in practice is because it provides consistency for estimates of regression parameter SEs whether the working correlation structure is correctly specified. However, when the number of subjects, N, is not large, this estimator can be negatively biased because the estimated residuals, $\hat{\boldsymbol{e}}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i$, $i = 1, \ldots, N$, are too small on average [25] and, when estimating $\boldsymbol{\beta}$, the covariance inflation, increasing estimation variability, for $Cov\left(\hat{\boldsymbol{\beta}}\right)$ results from the use of estimated correlation parameters in $\boldsymbol{R}_i$ or estimated covariance parameters in $\boldsymbol{C}_{\boldsymbol{N}}$ [26, 28, 34]. Therefore, corrections were utilized for the bias from the use of residual vectors, such as the corrections of Mancl and DeRouen [25] and Kauermann and Carroll [35]. The two well-known methods, along with other corrections, have been demonstrated to outperform for bias correction. Nonetheless, in practice, either the correction of Kauermann and Carroll [35] and the average of these two corrected methods from Ford and Westgate [28, 31] may be most desirable. Adjustment to any covariance inflation occurring with the use of empirical covariance matrix or correlation matrix for the three estimating equation approaches was also recommended [26–29].

Substitution methods

There are two major imputation techniques that are generally used for environmental monitoring and exposure assessment samples measured less than or below the LOD. We did not advocate direct truncation for the left-censored values because this method might alter the results of central tendency and variability or dispersion, thus decreasing accuracy and

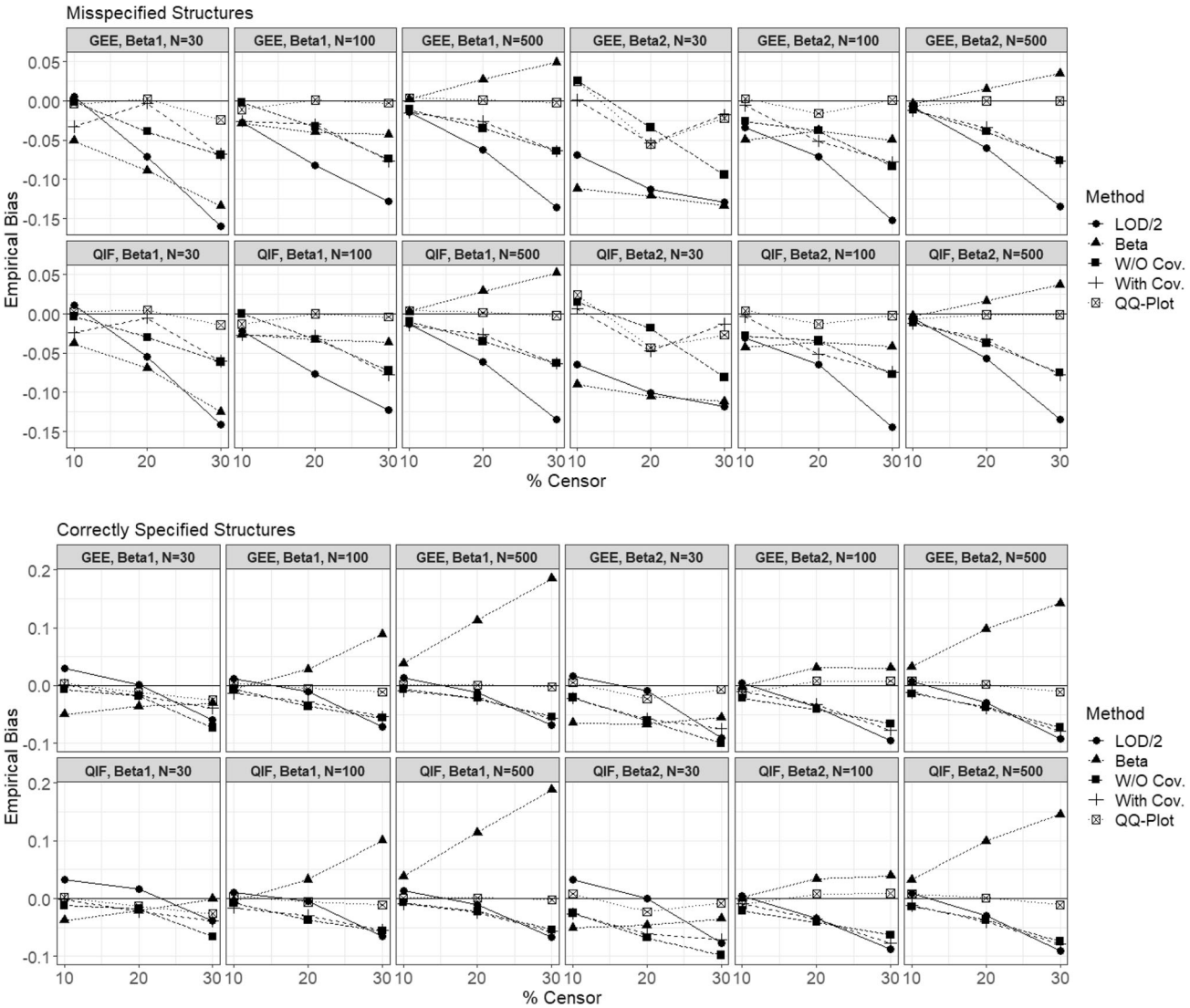

Fig. 1 Comparisons of empirical biases. Simulation results for the marginal model comparing two estimation approaches, GEE and QIF, and five substitution methods, LOD/2 substitution, β-substitution (Beta), multiple random value imputations without (W/O Cov.) and with covariates (With Cov.), and multiple ordered value imputation (QQ-Plot) for log-normal data. Empirical bias is the mean difference between the estimated value and the true parameter value of β_1 or β_2 for 1000 simulations. N is the number of subjects. % Censor is the censoring proportion.

precision. Zero replacement for values below the LOD is also not supportive because the logarithmic zero is undefined if the data distribution is log-transformed.

Single value imputation techniques, such as LOD/2 or LOD/$\sqrt{2}$ [1, 2], and β-substitution method [12], assigning a value to a range between 0 and the LOD is the most popular technique adopted for conducting summary statistics. See Appendix A of Ganser and Hewett [12] for detailed β-substitution method algorithm and Appendix B's spreadsheet for example demonstration. Multiple value imputation techniques provide an increasingly attractive alternative for exposure and biomonitoring data with left censoring [3–5]. They employ a ML estimation and a bootstrap procedure, i.e., randomly sampling with replacement, to estimate distribution parameters based on the uncensored data and the censoring proportion so that a common parametric distribution with the estimated parameters can be used for imputing values for observation below the LOD [36]. With these methods, the imputed values can also be generated using a regression of an exposure measurement on covariate(s) [3]. More recently, multiple order value imputation was proposed by depicting the natural logarithm of the uncensored exposure and biomonitoring concentration levels versus the Z-scores to fit a linear equation regularly presented in a quantile-quantile (QQ) plot [13, 14]. All the substitution methods discussed here were also listed in the manuscript of Pleil [14]. Note that R functions implementing the β-substitution and QQ-plot methods were provided in Supplementary Material.

Proposed methods

Because of the popularity of marginal models and lack of the use of these models in environmental monitor and exposure assessment, we propose to develop marginal models in which the observations of exposure outcome, $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{iM}]^T$, $i = 1, \ldots, N$, below the LOD are substituted with either single value imputation, multiple random value imputation, or multiple ordered value imputation method. This outcome or dependent variable represents the exposure and biomonitoring level measured for the in ith subject at the jth measurement. Substitution is performed for each non-detectable jth, $j = 1, \ldots, M$, observation measured in the ith subject. The proposed marginal models then utilize all estimating equations, and estimations of regression parameter, standard error (SE), and correlation parameter in $\boldsymbol{R}_i$ or $\boldsymbol{R}_i^{-1}$ are carried out in the same manner as with marginal models. When repeated measures are in a cluster study, exchangeable or compound symmetry structure is accommodated to the working correlation matrix, $\boldsymbol{R}_i$, while AR-1 will be used in a longitudinal data set. We note that small-sample corrections of estimated residuals and SE, such as the ones discussed for the existing marginal models, can be applied with our modified marginal models.

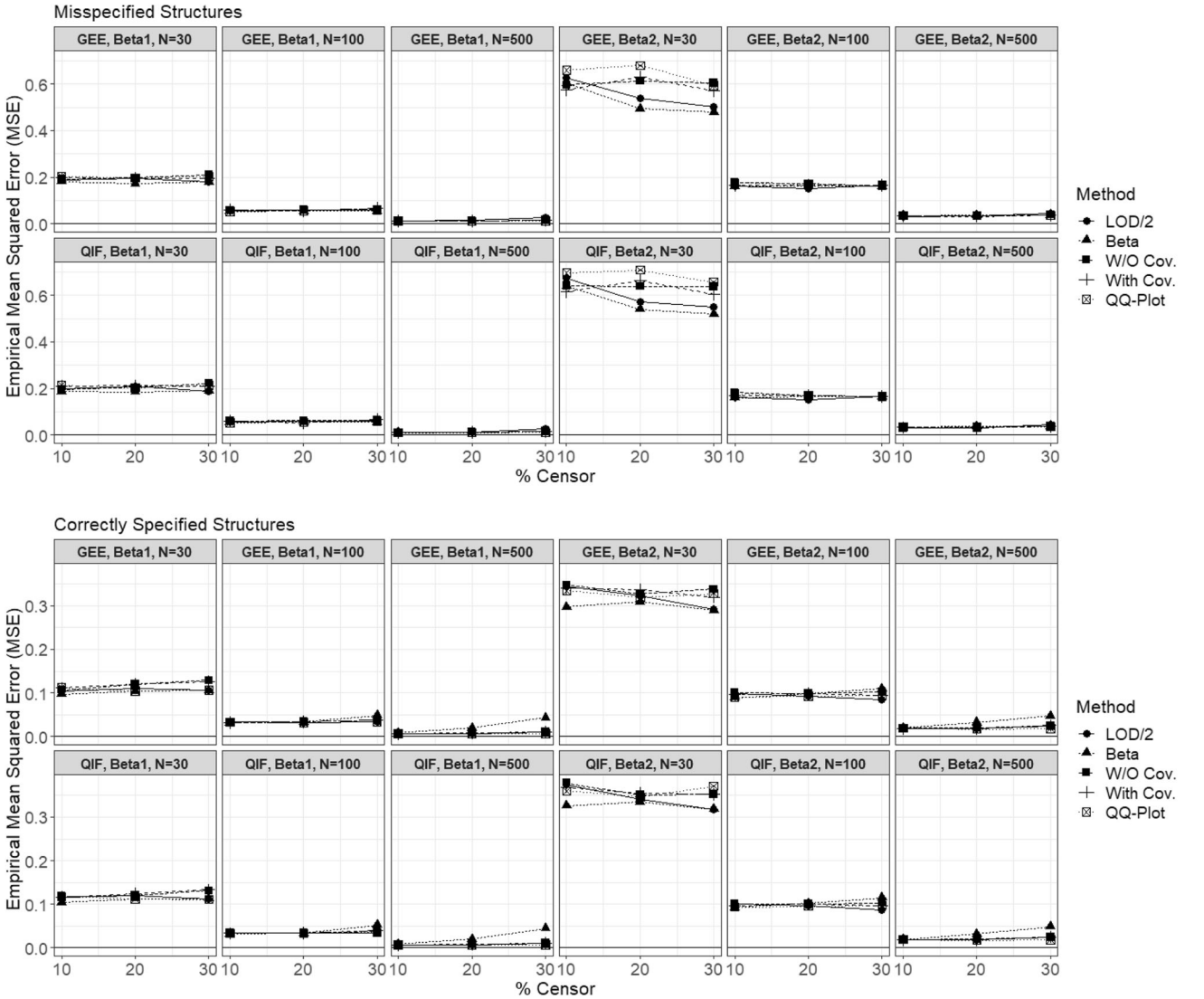

Fig. 2 Comparisons of empirical mean squared errors. Simulation results for the marginal model comparing two estimation approaches, GEE and QIF, and five substitution methods, LOD/2 substitution, β-substitution (Beta), multiple random value imputations without (W/O Cov.) and with covariates (With Cov.), and multiple ordered value imputation (QQ-Plot) for log-normal data. Empirical mean squared error (MSE) is the average squared difference between the estimated value and the true parameter value of β_1 or β_2 for 1000 simulations. N is the number of subjects. % Censor is the censoring proportion.

Simulation study

We now compare the performances of small-sample regression parameter estimation of the proposed approaches featuring combinations of two estimation approaches (GEE and QIF) with an AR-1 working correlation structure and five substitution methods for right-skewed and left-censored exposure and biomonitoring data with correlated exposure outcomes. The substitution methods include [14]:

(1) Single value imputation with the use of LOD/2.
(2) Single value imputation calculating a β factor to adjust each non-detectable value below the LOD, i.e., LOD/β.
(3) Multiple random value imputation using bootstrapping and omitting covariate information that corresponds to exposure outcome [37].
(4) Multiple random value imputation accounting for covariates that contain cluster identification number used in both cluster and longitudinal studies and order of time points within each cluster regularly required in a longitudinal study.
(5) Multiple ordered value imputation using a linear equation fit in a QQ-plot to obtain the imputed values regressed on the calculated Z-scores.

Settings of the simulation study consist of either 30, 100, or 500 subjects (N) presenting small, moderate, and large sample sizes. Each subject contributes three repeated measures (M) that represent the size of cluster or the number of time points. Each setting is conducted through 1000 simulations. All simulations with results provided in Figs. 1–3 and Supplementary Tables S1–S8 for the proposed approaches were carried out using R version 4.2.2 [38]. Moreover, we utilized one data generating model motivated by the literature of parametric repeated measures models [18, 29, 39] with the censoring proportions subjected to 10%, 20%, and 30% censoring. The marginal model is generated from $\log Y_{ij} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma_i + \varepsilon_{ij}$, $i = 1, \ldots, N$; $j = 1, \ldots, M$, where Y_{ij} is the jth measurement for the ith subject, x_{1i} and x_{2i} are independent variables following Bernoulli and uniform distributions of *Bernoulli*(0.5) and *Uniform*(0, 1), respectively, and γ_i and ε_{ij}, are mutually independent random effects, in which two scenarios are accounted for in the model. Scenario 1 assumes that the random effects are normally distributed with mean 0 and variance 1 (GEE in Supplementary Table S1 and QIF in Supplementary Table S2), whereas a true AR-1 correlation structure is formed in scenario 2 (GEE in Supplementary Table S3 and QIF in Supplementary Table S4). Because the working correlation structures employed in this study are AR-1, all structures are mis-specified in scenario 1 (upper panels of Figs. 1–3) but correctly specified in scenario 2 (lower panels of Figs. 1–3). The true

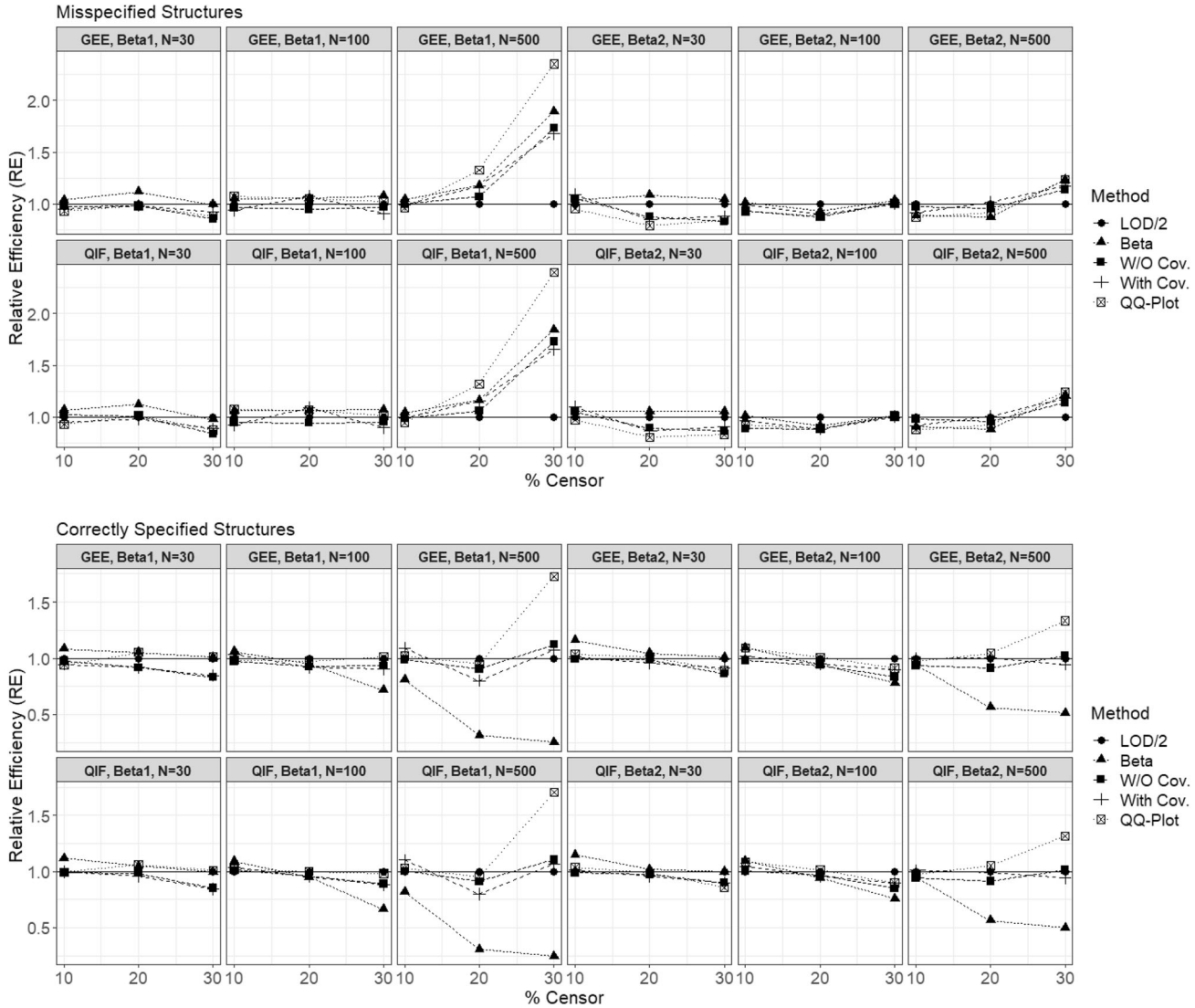

Fig. 3 Comparisons of relative efficiencies. Simulation results for the marginal model comparing two estimation approaches, GEE and QIF, and five substitution methods, LOD/2 substitution, β-substitution (Beta), multiple random value imputations without (W/O Cov.) and with covariates (With Cov.), and multiple ordered value imputation (QQ-Plot) for log-normal data. Relative efficiency (RE) is the ratio that, for each sample size or number of subjects (N) and censoring proportion (% Censor), compare the empirical MSE from the LOD/2 substitution method to the MSE from the other substitution method.

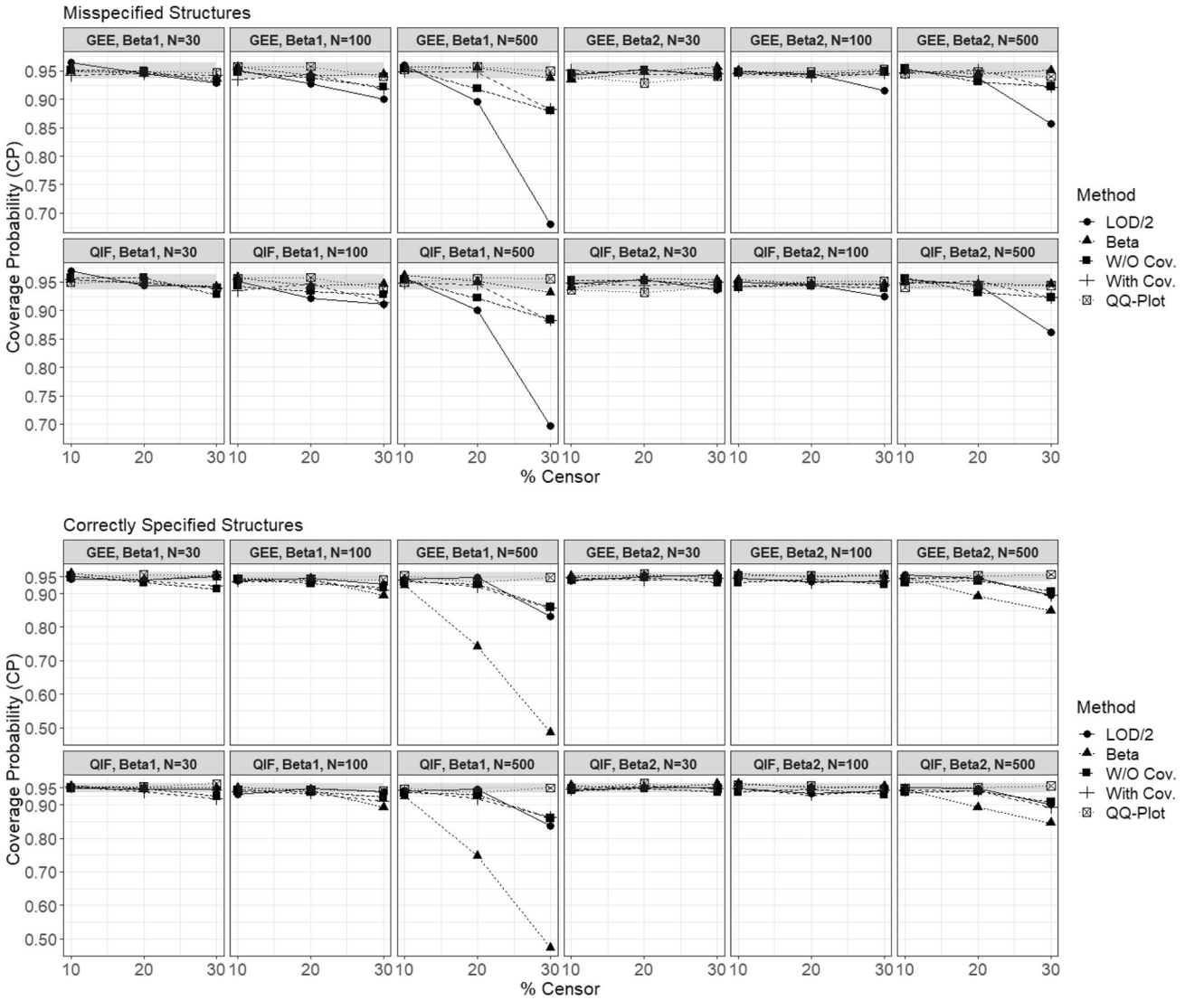

Fig. 4 Comparisons of coverage probabilities. Simulation results for the marginal model comparing two estimation approaches, GEE and QIF, and five substitution methods, LOD/2 substitution, β-substitution (Beta), multiple random value imputations without (W/O Cov.) and with covariates (With Cov.), and multiple ordered value imputation (QQ-Plot) for log-normal data. Coverage probability represents (CP_{AVG}) corresponding 95% confidence intervals which utilize the average of two bias corrections. The coverage probabilities falling within a near-nominal range of 0.936 and 0.964 are shaded. *N* is the number of subjects. % Censor is the censoring proportion.

values of $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$ are corresponded to the marginal intercept and slopes. The two scenarios were also carried out using the GMM approach, as the results were shown in Supplementary Tables S5 and S6. If $Y_{ij} < \text{LOD}_{ij}$, then Y_{ij} is replaced using the listed substitution methods. In addition to assume that the random effects are log-normally distributed, we further account for a highly skewed pattern, i.e., skewed after logarithmic transformation, assuming the random effects follow a chi-squared distribution with two degrees of freedom (d.f.). Same simulation settings as the ones in scenarios 1 and 2 and mis-specified working correlation structures are utilized, and the results are provided in Supplementary Tables S7 and S8. Note that the GMM method performed poorly in terms of small-sample regression parameter estimation and validity of inference. Therefore, we initially do not consider it within this manuscript. Instead, we place its results of estimation and inference in Supplementary Material.

To assess the differences in estimation performances of the two estimation approaches and the five substitution methods, we present empirical biases, empirical mean squared errors (MSEs), and ratios of MSEs from non-intercept parameters, in which we refer to as relative efficiencies (REs), in Figs. 1–3. For any given RE, the numerator is the MSE from the use of LOD/2, and the denominator is the MSE for the other substitution method. The modeling option that performs best therefore has the largest ratio. We also present empirical coverage probabilities (CPs) from corresponding 95% confidence intervals (CIs) which utilize the average of Mancl and DeRouen [25] and Kauermann and Carroll [35] corrections from Ford and Westgate [31], denoted by CP_{AVG} (Fig. 4).

RESULTS

In Fig. 1, the QQ-plot method had smaller empirical biases in either scenario and in either GEE or QIF estimation approach. For scenario 1 with structures incorrectly specified, the GEE approach had relatively lower MSEs compared to the QIF approach in either sample size setting. The REs indicated that the use of β-substitution method performed best overall in terms of regression parameter estimation. However, when sample size is large ($N = 500$), the QIF demonstrated an efficiency advantage over the GEE based on the results of MSEs, and the QQ-plot method outperformed the other substitution methods, especially for censoring = 30% (Figs. 2 and 3). The empirical CPs using the average of corrections were around the nominal value of 0.95 for the marginal models utilizing either the GEE or the QIF approach

Table 1. Summary results of relative efficiencies (RE) and coverage probabilities (CP_{AVG}) in scenarios 1 and 2 of the simulation study.

			Misspecified structures		Correctly specified structures	
Performance	***N***	**% Censoring**	**Estimation approach**	**Substitution method**	**Estimation approach**	**Substitution method**
RE	30	10	GEE	All Methods	GEE	All Methods
		20	GEE	All Methods	GEE	All Methods
		30	GEE	All Methods	GEE	All Methods
	100	10	GEE	All Methods	GEE	All Methods
		20	GEE	All Methods	GEE	All Methods
		30	GEE	All Methods	GEE	All Methods
	500	10	QIF	All Methods	QIF	All Methods
		20	QIF	QQ-Plot	QIF	All But Beta-Substitution
		30	QIF	QQ-Plot	QIF	QQ-Plot
CP_{AVG}	30	10	GEE & QIF	All Methods	GEE & QIF	All Methods
		20	GEE & QIF	All Methods	GEE & QIF	All Methods
		30	GEE & QIF	Beta-Substitution & QQ-Plot	GEE & QIF	QQ-Plot
	100	10	GEE & QIF	All Methods	GEE & QIF	All Methods
		20	GEE & QIF	All Methods	GEE & QIF	All Methods
		30	GEE & QIF	Beta-Substitution & QQ-Plot	GEE & QIF	QQ-Plot
	500	10	GEE & QIF	All Methods	GEE & QIF	All Methods
		20	GEE & QIF	All Methods	GEE & QIF	All But Beta-Substitution
		30	GEE & QIF	Beta-Substitution & QQ-Plot	GEE & QIF	QQ-Plot

N number of subject, *RE* relative efficiency is a ratio that, for each sample size (*N*) and percent censoring, compare the empirical MSE from the LOD/2 substitution method to the MSE from the other substitution method, CP_{AVG} coverage probability from corresponding 95% confidence intervals which utilize the average of two bias corrections.

and incorporating different substitution methods when $N \leq 100$ and censoring ≤20%. When censoring reached 30%, only β-substitution and QQ-plot methods resulted in near-nominal empirical CPs (Fig. 4; Tables S1 and S2).

When the working and true structures were AR-1 in scenario 2, MSE results of the GEE and QIF approaches were similar to the results observed in scenario 1 (Fig. 2). RE results corresponding to the QQ-plot method worked very well among all sample size and censoring settings (Fig. 3). Furthermore, the QQ-plot method maintained near-nominal 95% CPs for censoring ≤30%. Regarding censoring ≤20%, the CPs of the β-substitution method were close to the value of 0.95 in either $N = 30$ or 100, whereas the LOD/2 substitution method had appropriate CPs when $N = 30$ (Fig. 4; Supplementary Tables S3 and S4). The summary results of REs and CPs for either scenario are provided in Table 1.

Additionally, the QQ-plot performed better over the other methods under different censorings and had near-nominal CPs for the GMM when $N = 500$ (Supplementary Tables S5 and S6). When a skewed chi-squared distribution occurred with the exposure outcome data, the GEE approach was recommended for small ($N = 30$) and moderate ($N = 100$) sample sizes, while the QIF was favorable for large sample size ($N = 500$). Based on the simulation results of CPs and REs, both the LOD/2 substitution and β-substitution methods were advocated. The two multiple imputation techniques were also considerable because empirical CPs were near nominal (Supplementary Tables S7 and S8).

Tests Using Real-world Data

We applied the existing and proposed methods to a longitudinal study in which 37 applicators (number of independent subjects or *N*) with 184 breathing zone air samples were measured in four or five time points (number of repeated measures or *M*) [32]. The analytic LODs ranged from 0.05 to 0.2 µg/sample and the maximum chlorpyrifos exposure level was 73 µg/sample. All laboratory air mass data in µg/sample were converted to concentrations in $\mu g/m^3$ by dividing by the air sample volumes. The percentage of chlorpyrifos levels below the LOD was 1.63%, i.e., only three of 184 samples were censored. The distribution of the concentrations was skewed to the right and the exposure data were considered being log-normally distributed based on an examination of quantile-quantile (QQ) plot.

The marginal model we adopt was given by $\mu_{ij} = \beta_0 + \beta_1(x_{1ij} = 1) + \beta_2 x_{2ij}$, where μ_{ij} is the *i*th applicator's mean log-transformed airborne chlorpyrifos concentration during the *j*th 5-day workweek, x_{1ij} is an indicator for enclosed crawl space, and x_{2ij} is minutes of chlorpyrifos application on the sample collected day [32]. The data were analyzed using GEE and QIF estimation approaches, along with five substitution methods, with an AR-1 working correlation structure. The GMM approach was excluded because of its low precision in small-sample estimation. Table 2 provides regression parameter estimates and bias-corrected empirical SE estimates. To explore repeated measures data with higher censoring proportion, the laboratory air mass data were also artificially censored at the 20th percentile (20%). All approaches yield same directions and similar magnitudes for regression parameter estimates when the data were subject to low censoring. Both β-substitution and QQ-plot methods in either GEE or QIF approach produce smaller SE estimates when censoring level reached 20%, revealing the two method's potential for efficiency improvement and being consistent with the simulation results. Specifically, minutes chlorpyrifos applied and whether crawl space was treated were significantly associated with increased chlorpyrifos exposure. The correlation parameter estimates used to construct the AR-1 correlation structure ranged from 0.50 to 0.56, expressing moderate correlation among air samples collected from the same applicator.

The discussed methods were also carried out in a cluster study of NIOSH, which recruited four nail salons (number of independent subjects) and each nail salon had three workers (number of repeated measures) [33]. The analytic LOD was 0.16 µg/L and the

maximum DPhP metabolite level was 2.39 μg/L. All urinary sample data in μg/L were converted in μg/g by adjusting for their creatinine levels in mg/dL. There were three of 12 workers' DPhP pre-shift levels subject to censoring or below the LOD, i.e., 25% censoring. A visual determination of the QQ plot indicated that the data were normal. Therefore, no transformation was applied to the DPhP pre-shift concentrations.

The marginal model was given by $\mu_{ij} = \beta_0 + \beta_1(x_{1ij} = 1)$, where μ_{ij} is the mean concentration collected from the *j*th urine pre-shift sample of the *i*th worker and x_{1ij} is an indicator of whether the worker worked the previous day or worked two or more days ago [33]. We analyze the data using the same methods as in the longitudinal example, but with an exchangeable working correlation structure. Table 3 presents the estimates of regression parameter and empirical SE are presented. Results demonstrated that workers who had last worked two or more days ago had lower DPhP pre-shift concentrations relative to those who worked the previous day. When focusing on the preferable GEE approach for small number of clusters, all substitution methods yield same directions and but have different magnitudes for regression parameter estimates and SE estimates. To choose an appropriate method, we extend the correlation information criteria (CIC) for use in the setting, as it has gained popularity in simultaneously selecting estimation approaches with different working correlation structures [29, 40, 41]. Specifically, the CIC value is calculated by utilizing the trace of $\hat{\boldsymbol{\Sigma}}_I^{-1}\hat{\boldsymbol{\Sigma}}_{BC}$, in which $\hat{\boldsymbol{\Sigma}}_I = \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1} \boldsymbol{D}_i\right)^{-1}$ and $\hat{\boldsymbol{\Sigma}}_{BC}$ denotes small-sample corrected estimate of $Cov\left(\hat{\boldsymbol{\beta}}\right)$ for any candidate method under consideration. Therefore, the method yielding the smallest CIC will result in the least variable regression parameter estimates. In short, the CIC values indicate that the methods of β-substitution and multiple random value imputation with covariate of cluster identification number are preferable for this dataset.

DISCUSSION

Environmental exposure and biomonitoring data with repeated measurements in longitudinal and cluster studies are known to generally be subject to left censoring. Marginal models are appropriate when focusing on inferences about the population average [42] but these models have not been carried out in the literature on exposure and biomonitoring data with repeated measures and non-detects. Therefore, we proposed incorporating available fill-in or substitution methods for utilizing detects below the LOD into three estimating approaches, i.e., GEE, QIF, and GMM, in which consistent regression parameter estimates can be obtained even when a working correlation structure is incorrectly specified [21]. Additionally, we implemented recently developed small-sample corrections to estimators of covariance matrix corresponding to regression parameter estimates. Through bias corrections, the GEE approach is expected to potentially improve upon the QIF performance when the number of subjects (*N*) is small. The QIF approach performed better in terms of estimating the regression parameters in moderate and large sample sizes in the simulation study in consequence of its theoretical efficiency. In Supplementary Material, the GMM approach was shown to have invalid inference if the sample size is not large even when small-sample corrections are implemented, for instance, CPs had notable departures from the nominal 0.95 level. The finding was consistent with the study of Newey and Smith [43], which specified that unreliable inferences for the GMM would result from utilizing many moment conditions, e.g., increasing the number of time points or cluster sizes, relative to the sample size. The GMM was demonstrated to have valid inferences when the sample size is large. However, we do not advocate its use in practice because the validity of inference would be questionable when the number of time points increases.

Table 2. Parameter estimates, bias-corrected standard error estimates (in parentheses), and corresponding correlation parameter estimates resulting from analyses of the chlorpyrifos data.

% Censoring		Covariate	LOD/2 substitution	Beta-substitution	Imputation without covariates	Imputation with covariates	QQ-plot
1.63	GEE	Minutes Applied	0.006 (0.002)	0.006 (0.002)	0.006 (0.002)	0.006 (0.002)	0.006 (0.002)
		Crawl Space Treated	0.612 (0.103)	0.614 (0.102)	0.614 (0.102)	0.613 (0.102)	0.618 (0.102)
		Estimated Correlation	0.546	0.557	0.558	0.555	0.568
	QIF	Minutes Applied	0.008 (0.002)	0.007 (0.002)	0.007 (0.002)	0.007 (0.002)	0.007 (0.002)
		Crawl Space Treated	0.638 (0.105)	0.636 (0.104)	0.635 (0.104)	0.636 (0.104)	0.635 (0.104)
20	GEE	Minutes Applied	0.008 (0.002)	0.007 (0.002)	0.008 (0.002)	0.007 (0.002)	0.005 (0.002)
		Crawl Space Treated	0.713 (0.130)	0.664 (0.116)	0.726 (0.129)	0.549 (0.159)	0.540 (0.090)
		Estimated Correlation	0.556	0.553	0.532	0.501	0.564
	QIF	Minutes Applied	0.009 (0.002)	0.008 (0.002)	0.009 (0.003)	0.007 (0.002)	0.007 (0.002)
		Crawl Space Treated	0.725 (0.137)	0.683 (0.120)	0.734 (0.132)	0.621 (0.143)	0.568 (0.091)

GEE generalized estimating equations, *QIF* quadratic inference function.

Table 3. Parameter estimates, bias-corrected standard error estimates (in parentheses), and correlation parameter estimates resulting from analyses of the flame retardant data.

	Covariate	LOD/2 substitution	Beta-substitution	Imputation without covariates	Imputation with covariates	QQ-plot
GEE	Last Shift Worked					
	Previous Day	Reference	Reference	Reference	Reference	Reference
	Two or More Days Ago	−0.76 (0.17)	−0.72 (0.11)	−0.80 (0.09)	−0.59 (0.12)	−0.36 (0.43)
	Estimated Correlation Parameter in R_i	−0.03	−0.21	−0.03	−0.25	−0.29
	CIC	1.67	0.95	1.35	0.89	1.18
QIF	Last Shift Worked					
	Previous Day	Reference	Reference	Reference	Reference	Reference
	Two or More Days Ago	−0.73 (0.08)	−0.69 (0.08)	−0.59 (0.10)	−0.91 (0.17)	−0.71 (0.21)

GEE generalized estimating equations, *QIF* quadratic inference function, *CIC* correlation information criteria.

Even though we only accounted for parsimonious AR-1 and exchangeable working correlation structure for marginal models in this manuscript, other working matrices are available for the GEE approach, including exponential, Toeplitz, and unstructured working structures [27]. The unstructured form can also be incorporated with the QIF approach [44]. Including an AR-1 working structure to longitudinal exposure and biomonitoring data with repeated measures and non-detects is an additional advantage because it is favored over the other structures in a longitudinal study. Furthermore, based on literature, random effects models used for analyzing left-censored repeated measures exposure and biomonitoring data do not accommodate the use of AR-1 because the covariance matrix of the random error term corresponding to the subject level cannot be clarified [18].

Our simulation study and two application examples conducted marginal models for continuous responses and multiple covariates, i.e., univariable and multivariable analyses. The longitudinal example was demonstrated that subjects with varying numbers of time points are allowable. Similarly, unbalanced cluster sizes can also be implemented to any cluster exposure and biomonitoring data. Note that software such as the SAS procedure GLIMMIX [45], e.g., which can be used for GEE analyses and data with varying numbers of time points or cluster size, can accommodate the user with the two well-known corrections and further average these two corrections. In the simulations and examples, a single LOD or unique censoring proportion and multiple LODs were also permittable. A follow-up topic for future research using longitudinal data is to account for time-varying covariates and to select the type of time-dependency a covariate belongs to [29, 46].

The study has some limitations. The substitution methods we considered as options are generally used for calculating summary statistics for real-world left-censored environmental and biological data. Regardless of the chosen method, all imputed values are not real data but rather estimated values for measurements that are subject to LOD of laboratory instruments and contain unobserved errors. Additionally, the methods of multiple imputation and QQ plot require a common parametric distribution, e.g., log-normal distribution, with the estimated parameters obtained from the uncensored data to impute values for observations below the LOD. These methods might result in biased or unstable statistics when exposure and biomonitoring data are asymmetric after log-transforming [20]. Although the methods of LOD/2 and β-substitution are difficult to perform standard normality tests, they are easier to implement and calculate. Another limitation is that all imputed values for measurements falling below the LOD are assumed to be independent. There is no difference in the estimate of interest when incorrectly ignoring the correlation. However, disregard of the correlation will result in positively biased SE estimates, i.e., incorrect estimates of the sampling variability. In such cases, future work can be developed by imputing values based on truncated multivariate normal distribution for log-normal repeated measures data with non-detects [3] or taking truncated multivariate gamma distribution into account for right-skewed data. However, the desired imputation techniques are still constrained by the distributional assumption. Treating the non-detects as non-random missingness and applying a penalized expectation-maximization algorithm might also be considerable [47].

Because of the increasingly multilevel or hierarchical exposure and biomonitoring data regarding multiple levels of outcomes, future work for other marginal models is needed. Although the primary focus in this manuscript is on statistical inference, improvement of empirical power can be further considered for other topics of interest corresponding to environmental monitoring and exposure assessment data.

CONCLUSIONS

Only working mean and covariance structures are necessary to be specified for the popular estimating approaches, i.e., GEE, QIF, and GMM, used for the marginal analysis of correlated data. Furthermore, only the mean structure needs to be correctly provided to obtain consistent parameter estimates. After incorporating substitution methods for replacing non-detects below the LOD and utilizing small-sample bias corrections, the GEE approach is expected to outperform the others when the number of subjects is small, while the QIF approach performed better in moderate and large sample size. Although the replacement for non-detectable values is warranted, through the simulations, we suggested the uses of β-substitution and QQ-plot methods for log-normal data with repeated measurements and left censoring, and LOD/2 and β-substitution, as well as multiple imputations, were recommended for highly skewed data.

DATA AVAILABILITY

Detailed information of the two working examples can be found in the selected articles [32, 33]. The simulation and application R code and functions for implementing the proposed approaches in this manuscript are presented in Supplementary Material or can be addressed to I-Chen Chen.

REFERENCES

1. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg. 1990;5:46–51.
2. Burstyn I, Teschke K. Studying the determinants of exposure: a review of methods. Am Ind Hyg Assoc J. 1999;60:57–72.
3. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004;112:1691–6.

4. Huybrechts T, Thas O, Dewulf J, Van Langenhov H. How to estimate moments and quantiles of environmental data sets with nondetected observations? A case study on volatile organic compounds in marine water samples. J Chromatogr A. 2002;975:123–33.
5. Baccarelli A, Pfeiffer R, Consonni D, Pesatori AC, Bonzini M, Patterson DG Jr, et al. Handling of dioxin measurement data in the presence of nondetectable values: overview of available methods and their application in the Seveso chloracne study. Chemosphere. 2005;60:898–906.
6. Amemiya T. Regression analysis when the dependent variable is truncated normal. Econometrica. 1973;41:997–1016.
7. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere. 2006;65:2434–9.
8. Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. Ann Occup Hyg. 2007;51:611–32.
9. Gilliom RJ, Helsel DR. Estimation of distributional parameters for censored trace level water quality data 1. estimation techniques. Water Resour Res. 1986;22:135–46.
10. Helsel DR, Cohn TA. Estimation of descriptive statistics for multiply censored water quality data. Water Resour Res. 1988;24:1997–2004.
11. Shoari N, Dubé JS, Chenouri S. Estimating the mean and standard deviation of environmental data with below detection limit observations: Considering highly skewed data and model misspecification. Chemosphere. 2015;138:599–608.
12. Ganser GH, Hewett P. An accurate substitution method for analyzing censored data. J Occup Environ Hyg. 2010;7:233–44.
13. Pleil JD. QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. J Breath Res. 2016;10:035001.
14. Pleil JD. Imputing defensible values for left-censored 'below level of quantitation' (LoQ) biomarker measurements. J Breath Res. 2016;10:045001.
15. Thi´ebaut R, Jacqmin-Gadda H. Mixed models for longitudinal left-censored repeated measures. Comput Methods Prog Biomed. 2004;74:255–60.
16. Thi´ebaut R, Guedj J, Jacqmin-Gadda H, Chenê G, Trimoulet P, Neau D, et al. Estimation of dynamical model parameters taking into account undetectable marker values. BMC Med Res Methodol. 2006;6:38.
17. Vaida F, Liu L. Fast implementation for normal mixed effects models with censored response. J Comput Graph Stat. 2009;18:797–817.
18. Jin Y, Hein MJ, Deddens JA, Hines CJ. Analysis of lognormally distributed exposure data with repeated measures and values below the limit of detection using SAS. Ann Occup Hyg. 2011;55:97–112.
19. Leidel NA, Busch KA, Lynch JR. Occupational exposure sampling strategy manual (DHEW [NIOSH] publication no. 77-173). Cincinnati, OH: National Institute for Occupational Safety and Health; 1977.
20. Helsel DR. Less than obvious: statistical treatment of data below the detection limit. Environ Sci Technol. 1990;24:1766–74.
21. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73:13–22.
22. Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika. 2003;90:29–41.
23. Hansen LP. Large sample properties of generalized method of moments estimators. Econometrica. 1982;50:1029–54.
24. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000;87:823–36.
25. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001;57:126–34.
26. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. Stat Med. 2012;31:4003–22.
27. Westgate PM. A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. Stat Med. 2013;32:2850–8.
28. Westgate PM. A covariance correction that accounts for correlation estimation to improve finite-sample inference with generalized estimating equations: A study on its applicability with structured correlation matrices. J Stat Comput Simul. 2016;86:1891–1900.
29. Chen IC, Westgate PM. Improved methods for the marginal analysis of longitudinal data in the presence of time-dependent covariates. Stat Med. 2017;36:2533–46.
30. Ford WP, Westgate PM. Improved standard error estimator for maintaining the validity of inference in cluster randomized trials with a small number of clusters. Biometrical J. 2017;59:478–95.
31. Ford WP, Westgate PM. A comparison of bias-corrected empirical covariance estimators with generalized estimating equations in small-sample longitudinal study settings. Stat Med. 2018;37:4318–29.
32. Hines CJ, Deddens JA. Determinants of chlorpyrifos exposures and urinary 3,5,6-trichloro-2-pyridinol levels among termiticide applicators. Ann Occup Hyg. 2001;45:309–21.
33. Estill CF, Slone J, Mayer AC, Chen IC, Zhou M, La Guardia MJ, et al. Assessment of Triphenyl Phosphate (TPhP) exposure to nail salon workers by air, hand wipe, and urine analysis. Int J Hyg Environ Health. 2021;231:113630.
34. Windmeijer F. A finite sample correction for the variance of linear efficient two-step GMM estimators. J Econ. 2005;126:25–51.
35. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc. 2001;96:1387–96.
36. Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken, New Jersey: John Wiley and Sons; 2002.
37. Hargarten PM, Wheeler DC. miWQS: Multiple Imputation Using Weighted Quantile Sum Regression. R package version 0.4.4; 2021. https://CRAN.R-roject.org/package=miWQS
38. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2022. https://www.R-project.org
39. Jacqmin-Gadda H, Thi´ebaut R. Analysis of left-censored longitudinal data with application to viral load in HIV infection. Biostatistics. 2000;1:355–68.
40. Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. Stat Med. 2009;28:642–58.
41. Westgate PM. Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach. Biometrical J. 2014;56:461–76.
42. Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. The Analysis of Longitudinal Data, 2nd ed. New York: Oxford University Press; 2002.
43. Newey WK, Smith RJ. Higher order properties of GMM and generalized empirical likelihood estimators. Econometrica. 2004;72:219–55.
44. Westgate PM. A bias-corrected covariance estimator for improved inference when using an unstructured correlation with quadratic inference functions. Stat Probab Lett. 2013;83:1553–8.
45. SAS Institute Inc. SAS/STAT 9.3 Users Guide. SAS Institute Inc., Cary, NC; 2011.
46. Chen IC, Westgate PM. A novel approach to selecting classification types for time-dependent covariates for the marginal analysis of longitudinal data. Stat Methods Med Res. 2018;28:3176–86.
47. Chen LS, Prentice RL, Wang P. A penalized EM algorithm incorporating missing data mechanism for gaussian parameter estimation. Biometrics. 2014;70:312–22.

ACKNOWLEDGEMENTS

We would like to thank the people from the Division of Field Studies and Engineering at CDC's National Institute for Occupational Safety and Health who assisted in the study. The findings and conclusions in this manuscript are those of the authors and do not necessarily represent the official position of the National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention.

AUTHOR CONTRIBUTIONS

ICC was responsible for designing statistical methods, conducting a simulation study, analyzing two real-world datasets, interpreting simulation and application results, producing tables and figures, drafting the initial manuscript, revising the manuscript, and approving the final version of manuscript. SJB contributed to interpretations of simulation and application results, revised manuscript, provided feedback, and approved the final version. CFE contributed to data curation and extraction, revised manuscript, provided feedback, and approved the final version. Additionally, Whitney F. Tanner and Yu-Cheng Chen reviewed the paper and provided helpful feedback.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41370-024-00640-7.

Correspondence and requests for materials should be addressed to I-Chen Chen.