Received: 23 October 2020 | Revised: 10 July 2022 | Accepted: 11 July 2022

DOI: 10.1002/sim.9536

RESEARCH ARTICLE

Statistics in Medicine WILEY

Semiparametric analysis of a generalized linear model with multiple covariates subject to detection limits

Ling-Wan Chen[1] | **Jason P. Fine**[2] | **Eric Bair**[3] | **Victor S. Ritter**[1] | **Thomas F. McElrath**[4] | **David E. Cantonwine**[4] | **John D. Meeker**[5] | **Kelly K. Ferguson**[6] | **Shanshan Zhao**[1]

[1]Biostatistics & Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

[2]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

[3]Sciome LLC, Durham, North Carolina,

[4]Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

[5]Department of Environmental Health Sciences, University of Michigan School of Public Health, Ann Arbor, Michigan, USA

[6]Epidemiology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, USA

Correspondence
Shanshan Zhao, Biostatistics & Computational Biology Branch, National Institute of Environmental Health Sciences, 111 TW Alexander Dr, Research Triangle Park, NC 27709, USA.
Email: shanshan.zhao@nih.gov

Funding information
National Institute of Environmental Health Sciences, Grant/Award Number: ZIA ES103307

Studies on the health effects of environmental mixtures face the challenge of limit of detection (LOD) in multiple correlated exposure measurements. Conventional approaches to deal with covariates subject to LOD, including complete-case analysis, substitution methods, and parametric modeling of covariate distribution, are feasible but may result in efficiency loss or bias. With a single covariate subject to LOD, a flexible semiparametric accelerated failure time (AFT) model to accommodate censored measurements has been proposed. We generalize this approach by considering a multivariate AFT model for the multiple correlated covariates subject to LOD and a generalized linear model for the outcome. A two-stage procedure based on semiparametric pseudo-likelihood is proposed for estimating the effects of these covariates on health outcome. Consistency and asymptotic normality of the estimators are derived for an arbitrary fixed dimension of covariates. Simulations studies demonstrate good large sample performance of the proposed methods vs conventional methods in realistic scenarios. We illustrate the practical utility of the proposed method with the LIFECODES birth cohort data, where we compare our approach to existing approaches in an analysis of multiple urinary trace metals in association with oxidative stress in pregnant women.

KEYWORDS
accelerated failure time model, limit of detection, multiple exposures, nonparametric survival estimation, pseudolikelihood, Z estimation theory

1 | INTRODUCTION

In environmental studies, it is important to understand the impact of environmental mixtures on human health, via exposures to food, air, water, consumer products, and others. A key challenge in statistical analyses is that exposure concentrations below limit of detection (LOD) in biological samples are not detectable (ie, left-censored). Thus, recovering

the true effects of environmental mixtures, where multiple correlated exposures are subject to LOD, is of interest. For example, in the LIFECODES cohort of women in the Boston area who planned to deliver at the Brigham and Women's Hospital between 2006 and 2008,[1-3] researchers are interested in the relationship between 17 urinary trace metals and 8-isoprostane, an important oxidative stress marker for preterm birth, where trace metals and 8-isoprostane were measured at women's third trimester of pregnancy. However, only three metals were fully measured, while the remaining 14 metals had 0.4% to 90.2% values below LOD among women with full term birth (Supplementary Table S8).

In the presence of LOD in a single covariate, a complete-case analysis, which excludes subjects with covariate value below LOD, provides an unbiased estimator when the underlying outcome model is correctly specified, but may result in loss of efficiency.[4-6] An alternative approach is to use a substitution method, which replaces the values below LOD with an arbitrary value, such as LOD, LOD/2, or LOD/$\sqrt{2}$. Such an approach is commonly used due to its simplicity but may result in large biases.[5,7] Richardson and Ciampi[8] recommended replacing the value below LOD with the conditional expectation. This approach requires an assumption on the left tail of the covariate distribution, which is unverifiable from the observed data. Maximum likelihood approaches have also been proposed, under parametric assumptions on the censored covariate, for different types of outcome models, such as generalized linear model, Cox regression model or frailty model, and for AUC comparison.[4,5,9-12] Of course, these estimates can yield large biases when the parametric assumption is misspecified. Recently, there has been work on semiparametric approaches to relax the distributional assumption for the censored covariate. For instance, Kong and Nan[13] proposed a semiparametric accelerated failure time (AFT) model for the censored covariate with a generalized linear outcome model. Atem et al[14] proposed a semiparametric imputation approach based on a Cox model to impute the censored covariate under a linear outcome model. They further extended the method to accommodate survival outcomes.[15] Ding et al [16] proposed a semiparametric two-step importance sampling imputation for the censored covariate based on a semiparametric AFT model with a generalized linear outcome model.

In practice, there may be multiple correlated covariates subject to LOD, as discussed before in the LIFECODES study, which could benefit from careful accommodation of the correlation structure between exposures. Many studies use imputation methods for each individual covariate, which ignore the dependency between covariates. Maximum likelihood approaches have also been proposed, given a parametric form of the joint distribution for the multiple censored covariates. For example, May et al[17] used a Monte-Carlo EM algorithm to obtain the estimates with a generalized linear outcome model. Wu et al[18] and Chen et al[19] considered a Bayesian approach for a generalized linear outcome model and a Cox outcome model. In addition, multiple imputation based on a distributional assumption for the censored covariates has been explored by various authors.[9,20-22] However, these methods all require parametric assumptions for the joint distribution of the multiple censored covariates, which can be difficult to specify in practice. An extension of the semiparametric approach to multiple censored covariates based on maximum likelihood is unclear.

In this work, we adapt the semiparametric pseudo-likelihood technique in Kong and Nan[13] to an arbitrary number of covariates subject to LOD with a generalized linear outcome model. A two-stage procedure is proposed to recover the coefficients in the outcome model. In the first stage, we fit a semiparametric multivariate AFT model for the censored covariates and estimate the joint distribution of the error terms nonparametrically. We estimate the parameters of interest in the second stage with the nuisance parameters estimated from the first stage plugged into the likelihood. We describe the model and methods in Section 2, and establish the asymptotic properties in Section 3. Extensive simulations are presented in Section 4 to evaluate the finite-sample performance of the proposed methods. We use the LIFECODES birth cohort data to illustrate our method in Section 5, and conclude with remarks in Section 6.

2 | METHODS

2.1 | Likelihood framework with covariates subject to LOD

Consider a single response variable Y, q fully observed covariates $\mathbf{X} = (1, X_1, \ldots, X_q)^T$, and p covariates subject to LOD, $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$, with corresponding known LOD values $(\text{LOD}_1, \ldots, \text{LOD}_p)^T$. Here we assume all LODs are lower limits of detection but our method can be extended to upper limits or both for a single component. If $Z_j < \text{LOD}_j$, a left-censoring of Z_j is observed. By applying a monotone decreasing transformation $h_j^{-1}(\cdot)$, we can rewrite the left-censored covariate Z_j as a right-censored covariate T_j, where $Z_j = h_j(T_j)$, $\text{LOD}_j = h_j(C_j)$ and C_j is the corresponding censoring value of T_j. Thus, we observe $V_j = \min(T_j, C_j)$ and $\Delta_j = I(T_j \le C_j)$, $j = 1, \ldots, p$. We further denote $\mathbf{T} = (T_1, \ldots, T_p)^T$, $\mathbf{C} = (C_1, \ldots, C_p)^T$, $\mathbf{V} = (V_1, \ldots, V_p)^T$, $\mathbf{\Delta} = (\Delta_1, \ldots, \Delta_p)^T$, and $\mathbf{Z} = \{h_1(T_1), \ldots, h_p(T_p)\}^T = h^*(\mathbf{T})$.

Assume that Y comes from an exponential family with density

$$f_{\omega,\phi}(Y) = \exp\left\{\frac{Y\omega - b(\omega)}{a(\phi)} + c(Y,\phi)\right\}, \tag{1}$$

where ω is the natural parameter, ϕ is the dispersion parameter, and $a(\cdot)$ and $b(\cdot)$ satisfy $E(Y) = \mu = \partial b(\omega)/\partial\omega$ and $\text{var}(Y) = a(\phi)\partial^2 b(\omega)/\partial\omega\partial\omega^T$. Under a generalized linear model with a canonical link g, where $E(Y) = \mu = g^{-1}(\theta^T D)$ and $D = (\mathbf{X}^T, \mathbf{Z}^T)^T = \{\mathbf{X}^T, h^*(\mathbf{T})^T\}^T$, one can define the density of Y given $\mathbf{X}$ and $\mathbf{T}$ as $f_{\theta,\phi}(Y|\mathbf{X},\mathbf{T})$, substituting $\omega = \theta^T D$ into Equation (1). Here $\theta = (\beta^T, \gamma^T)^T$ is the regression parameter of interest, where $\beta_{(q+1)\times 1}$ and $\gamma_{p\times 1}$, correspond to $\mathbf{X}$ and $h^*(\mathbf{T})$, respectively. In this work, we focus on the canonical link g but note that any link function satisfying the regularity conditions in Section 3 can be accommodated by our method.

For the ith subject, we denote $\mathbf{T}_i$ as the vector of the transformed censored covariates, $\mathbf{T}_{-ji}$ as the vector where the jth element is removed from $\mathbf{T}_i$, and $\mathbf{T}_{-(j,k)i}$ as the vector where the jth and kth elements are removed from $\mathbf{T}_i$, $j < k$, and so on for higher dimensions. Thus, the likelihood for the observed data $(Y_i, \mathbf{X}_i, \mathbf{V}_i, \mathbf{\Delta}_i)$ can be written as

$$\begin{aligned}
&L_i(\theta,\phi; Y_i, \mathbf{X}_i, \mathbf{V}_i, \mathbf{\Delta}_i) = f_{\theta,\phi}(Y_i, \mathbf{X}_i, \mathbf{V}_i, \mathbf{\Delta}_i) \propto f(Y_i|\mathbf{V}_i, \mathbf{\Delta}_i, \mathbf{X}_i) f(\mathbf{V}_i, \mathbf{\Delta}_i|\mathbf{X}_i)\\
&= \{f_{\theta,\phi}(Y_i|\mathbf{T}_i, \mathbf{X}_i) f_{\mathbf{T}}(\mathbf{T}_i|\mathbf{X}_i)\}^{\prod_{j=1}^p \Delta_{ji}}\\
&\quad\times \prod_{j=1}^p \left\{\int_{C_j}^{\infty} f_{\theta,\phi}(Y_i|t_j, \mathbf{T}_{-ji}, \mathbf{X}_i) f_{\mathbf{T}}(\mathbf{t}|\mathbf{X}_i, \mathbf{t}_{-j} = \mathbf{T}_{-ji}) dt_j\right\}^{(1-\Delta_{ji})\prod_{l\neq j}\Delta_{li}}\\
&\quad\times \prod_{j=1}^p \prod_{k>j}^p \left\{\int_{C_j}^{\infty}\int_{C_k}^{\infty} f_{\theta,\phi}(Y_i|t_k, t_j, \mathbf{T}_{-(j,k)i}, \mathbf{X}_i) f_{\mathbf{T}}(\mathbf{t}|\mathbf{X}_i, \mathbf{t}_{-(j,k)} = \mathbf{T}_{-(j,k)i}) dt_k dt_j\right\}^{(1-\Delta_{ji})(1-\Delta_{ki})\prod_{l\notin\{j,k\}}\Delta_{li}}\\
&\quad\times\cdots\times \left\{\int_{C_1}^{\infty}\cdots\int_{C_p}^{\infty} f_{\theta,\phi}(Y_i|\mathbf{t}, \mathbf{X}_i) f_{\mathbf{T}}(\mathbf{t}|\mathbf{X}_i) d\mathbf{t}\right\}^{\prod_{j=1}^p (1-\Delta_{ji})},
\end{aligned} \tag{2}$$

where $f_{\mathbf{T}}(\mathbf{t}|\mathbf{X})$ is an unknown conditional joint distribution of $\mathbf{T}$ given $\mathbf{X}$. The likelihood in Equation (2) is a product over 2^p possible realizations of $\mathbf{\Delta}$, which will quickly get very large as p increases, and involves parameters θ, ϕ, and $f_{\mathbf{T}}$ in a complicated nonlinear form, which creates computational challenges. We notice that the parameters of interest θ are only involved in the first part of each term in the product $f_{\theta,\phi}$. Since there is no data to inform about the tail of the distribution $f_{\mathbf{T}}(\mathbf{t}|\mathbf{X})$, a flexible multivariate model for $\mathbf{T}$ with minimal assumptions on the tails is desirable.

2.2 | Semiparametric AFT model

We further assume a multivariate semiparametric AFT model for $\mathbf{T} = (T_1, \ldots, T_p)^T$,

$$T_j = h_j^{-1}(Z_j) = \alpha_j^T \mathbf{X}_j + \xi_j, \quad j = 1, \ldots, p, \tag{3}$$

where $\mathbf{X}_j$ is a subset of the fully observed covariates $\mathbf{X}$, α_j is the corresponding coefficient in the AFT model for T_j, $j = 1, \ldots, p$, and $\xi = (\xi_1, \ldots, \xi_p)^T$ follows an unknown joint distribution η, which is independent of $\mathbf{X}$. For simplicity, we assume a prespecified monotone decreasing function h_j, such as $h_j(u) = -u$ and $h_j(u) = \exp(-u)$, so that the linear relationship between T_j and $\mathbf{X}_j$ is valid. For brevity of notation, we set $\mathbf{X}_j = \mathbf{X}, j = 1, \ldots, p$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p]$. We chose to use this model specification for a couple reasons. First AFT model is widely used for modeling censored variables, without assumptions about proportional hazards and allows a direct linear relationship with other covariates. Existing semiparametric estimators of AFT model without assumptions about the error distribution also makes this model robust. Second, we allow fully observed variables $\mathbf{X}$ to contribute to the modeling of $\mathbf{T}$, which is desirable in practice. For example, in the LIFECODES data, we expect some baseline characteristics, such as maternal age, race, and BMI, to be related to some urinary metal concentrations. Third, this model allows us to introduce correlation between covariates in $\mathbf{Z}$, by sharing $\mathbf{X}$ and allowing correlation in the error distribution η. We believe through jointly estimating these AFT models, we use information more efficiently especially when some of the covariates are highly correlated. In a scenario where one of the covariates suffers from high proportions of values below LOD, this model allows better estimate when there is another

covariate in either $\mathbf{Z}$ or $\mathbf{X}$ that is highly correlated with it, comparing to when modeled individually with conventional approaches.

Under this model specification, the log-likelihood for $(Y_i, \mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\Delta}_i)$ can be further expressed as

$$\begin{aligned}\log L_i(\theta,\phi,\boldsymbol{\alpha},\eta) = &\prod_{j=1}^{p}\Delta_{ji}\{c_i(\boldsymbol{\alpha},\eta)+\log f_{\theta,\phi}(Y_i|\mathbf{T}_i,\mathbf{X}_i)\} \\ &+\sum_{j=1}^{p}(1-\Delta_{ji})\left(\prod_{l\neq j}\Delta_{li}\right)\log\left\{\int_{C_j-\alpha_j^T\mathbf{X}_i}^{\tau_j} f_{\theta,\phi}(Y_i|s_j+\alpha_j^T\mathbf{X}_i,\mathbf{T}_{-ji},\mathbf{X}_i)\eta(ds_j,ds_l=T_{li}-\alpha_l^T\mathbf{X}_i,\forall l\neq j)\right\} \\ &+\sum_{j=1}^{p}\sum_{k>j}^{p}(1-\Delta_{ji})(1-\Delta_{ki})\left(\prod_{l\notin\{j,k\}}\Delta_{li}\right) \\ &\times\log\left\{\int_{C_j-\alpha_j^T\mathbf{X}_i}^{\tau_j}\int_{C_k-\alpha_k^T\mathbf{X}_i}^{\tau_k} f_{\theta,\phi}(Y_i|s_j+\alpha_j^T\mathbf{X}_i,s_k+\alpha_k^T\mathbf{X}_i,\mathbf{T}_{-(j,k)i},\mathbf{X})\eta(ds_j,ds_k,ds_l=T_{li}-\alpha_l^T\mathbf{X}_i,\forall l\neq\{j,k\}\right\} \\ &+\cdots \\ &+\left\{\prod_{j=1}^{p}(1-\Delta_{ji})\right\}\log\left\{\int_{C_1-\alpha_1^T\mathbf{X}_i}^{\tau_1}\cdots\int_{C_p-\alpha_p^T\mathbf{X}_i}^{\tau_p} f_{\theta,\phi}(Y_i|s_1+\alpha_1^T\mathbf{X}_i,\ \ldots\ ,s_p+\alpha_p^T\mathbf{X}_i,\mathbf{X}_i)\eta(ds_1,\ \ldots\ ,ds_p)\right\},\end{aligned} \tag{4}$$

where $c_i(\boldsymbol{\alpha},\eta)$ only involves $(\boldsymbol{\alpha},\eta)$ and is constant in θ, and τ_j is a truncation value for the jth residual as defined in Condition C5 in Section 3, $j=1,\ \ldots\ ,p$. In practice, all residuals are finite with bounded $\alpha_j^T\mathbf{X}$. While in theory τ_j should be deterministic, we find that taking τ_j to be an arbitrary value larger than the empirical residuals for each term performs well in the simulations in Section 4 and the data analysis in Section 5.

2.3 | Two-stage pseudo-likelihood estimation

The full log-likelihood in Equation (4) involves complicated integration over 2^p possible realizations of $\boldsymbol{\Delta}$. We recall that the parameter of interest θ is only involved in $f_{\theta,\phi}$, the first terms of each summation in Equation (4). We can thus treat $(\phi,\boldsymbol{\alpha},\eta)$ as nuisance parameters, and estimate θ and $(\phi,\boldsymbol{\alpha},\eta)$ separately. To reduce complexity, we propose a two-stage procedure by first estimating the nuisance parameters $(\phi,\boldsymbol{\alpha},\eta)$, and then estimating θ from the pseudo-likelihood with $(\phi,\boldsymbol{\alpha},\eta)$ replaced by their estimates from the previous stage. The details of the procedure are as follows.

In Stage 1, the nuisance parameters $(\phi,\boldsymbol{\alpha},\eta)$ are estimated, with various approaches possible. The dispersion parameter ϕ in the generalized linear model may be estimated using the complete cases only with any valid method of estimation $(\hat{\phi})$. The regression parameter $\boldsymbol{\alpha}$ in the AFT models can be estimated individually either by rank-based methods[23-25] or by least-squares approaches,[26,27] with R packages **rankreg** and **aftgee**, respectively $(\hat{\boldsymbol{\alpha}})$. The joint distribution of the AFT model residuals, η, may be estimated by applying a nonparametric multivariate Kaplan-Meier (K-M) estimator to the estimated residuals $\hat{\boldsymbol{\xi}}=\mathbf{T}-\hat{\boldsymbol{\alpha}}^T\mathbf{X}$, as introduced in Prentice and Zhao[28] $(\hat{\eta}_{\hat{\boldsymbol{\alpha}}})$. This nonparametric estimate is based on decomposing a higher dimensional joint survivor function into lower dimensional survivor functions and a cross ratio. For example, when $p=2$, the bivariate survivor function $S(\xi_1,\xi_2)$ can be expressed as

$$S(\xi_1,\xi_2)=S(\xi_1,0)S(0,\xi_2)\prod_{0}^{\xi_1}\prod_{0}^{\xi_2}S(s_1,s_2)S(s_1^-,s_2^-)/\{S(s_1^-,s_2)S(s_1,s_2^-)\}.$$

We plug in K-M estimators for $S(\xi_1,0)$ and $S(0,\xi_2)$ and empirical estimates for the cross ratio, and calculate $\hat{\eta}(\xi_1,\xi_2)=1-\hat{S}(\xi_1,0)-\hat{S}(0,\xi_2)+\hat{S}(\xi_1,\xi_2)$. Note that this estimator simplifies to the K-M estimator with $p=1$ and the Dabrowska estimator with $p=2$.[29,30]

In Stage 2, we estimate θ from the log-likelihood with $(\hat{\phi},\hat{\boldsymbol{\alpha}},\hat{\eta}_{\hat{\boldsymbol{\alpha}}})$ plugged in. This gives the log-pseudo-likelihood of the observed data $\{Y_i,\mathbf{X}_i,\mathbf{V}_i,\boldsymbol{\Delta}_i\}_{i=1}^n$, defined by

$$\log \mathrm{PL}(\theta)=\sum_{i=1}^{n}\log L_i(\theta,\hat{\phi},\hat{\boldsymbol{\alpha}},\hat{\eta}_{\hat{\boldsymbol{\alpha}}}), \tag{5}$$

where $\log L_i(\cdot)$ is defined in Equation (4). The maximum pseudo-likelihood estimator $\hat{\theta}$ is obtained by solving $\partial \log \mathrm{PL}(\theta)/\partial\theta = 0$ via the Newton-Raphson algorithm where we set the initial value of θ from the complete-case analysis.

2.4 | Consideration with high-dimensional data

Theoretically, the proposed method works for an arbitrary p-dimensions of covariates subject to LOD. However, in practice, the computational burden may be prohibitive as p increases. Specifically, there are two major hurdles: the estimation of the p-dimensional joint distribution of AFT model error terms (η) and the optimization of $\log \mathrm{PL}(\theta)$ which contains $2^p - 1$ integrals. Computational time can be shown to be of order n^p. Below we suggest some approaches to simplify the computation.

First, we carefully preselect the covariates to be included in $\mathbf{Z}$ to reduce p. In practice, there may be a large number of covariates subject to LOD at different levels, as in the LIFECODES data. Although it is desirable to process all these variables with our proposed algorithm, it is most important to include covariates with 5% to 70% measurements below LOD in $\mathbf{Z}$ when computational power is limited. For covariates with less than 5% measurements below LOD, a substitution method with an appropriate replacement value is likely to introduce limited bias. For covariates with more than 70% measurements below LOD, we recommend modeling them as dichotomized variables. Although this will lead to changes in the interpretation of regression parameters, it is more reasonable than assuming an overall linear relationship when most of the values are unobserved.

In a lower dimensional setting, calculating the nonparametric multivariate K-M estimation of the joint distribution η is fast but the computational time increases rapidly as p increases. To reduce computational burden, we propose a naive estimator of η by making an independence assumption where $\hat{\eta}(\xi_1, \ldots, \xi_p) = \prod_{j=1}^{p} \hat{\eta}_j(\xi_j)$, and $\hat{\eta}_j(\xi_j)$ is an estimate of ξ_j distribution such as a K-M estimate. This marginal approximation approach may result in a biased estimate of η but in practice, with moderate correlation between ξ_j's, the computational time is much shorter while the bias is reasonably small, as suggested by simulation studies in Section 4. To illustrate the gains in computing speed, we present an example with $p = 3$ and show the computational time using different estimators of η in Figure 1A. We highlight that calculating the trivariate K-M estimator for η takes 30 minutes with $n = 400$ while calculating the marginal approximation takes 1 minute. The difference in computational time can be much more significant with larger p. Thus, the marginal approximation approach may be preferred in situations where the multivariate K-M estimator is computationally infeasible.

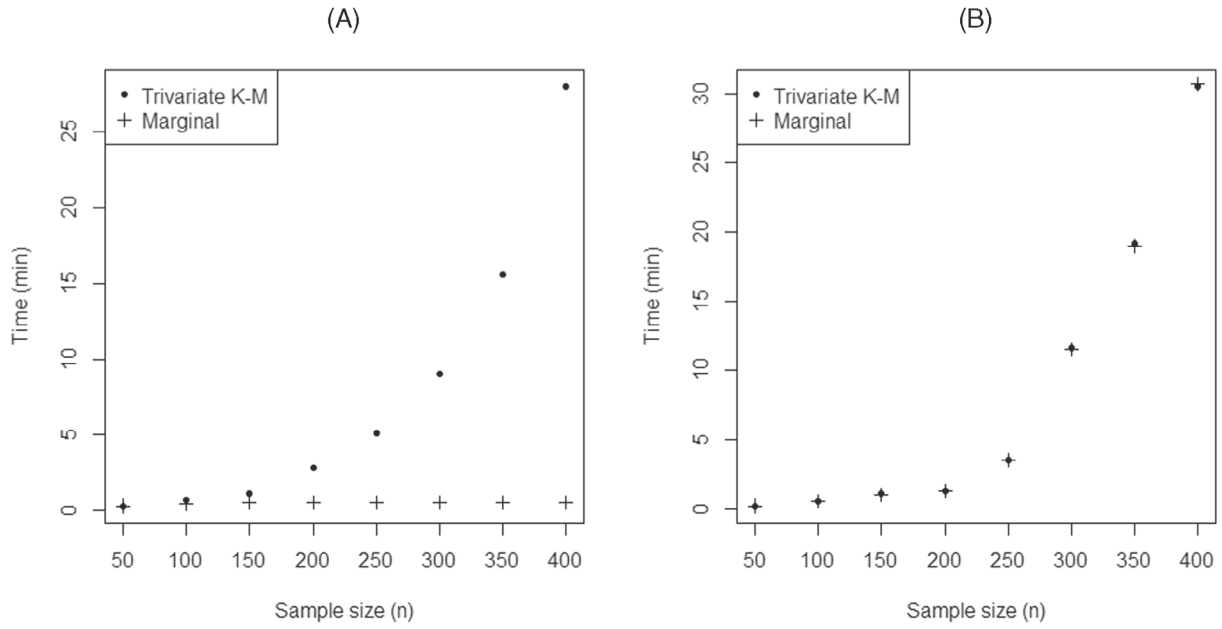

FIGURE 1 (A) Computational time for estimating η using the trivariate K-M and the marginal approximation at various sample sizes when $p = 3$. (B) Computational time for maximizing of $\log \mathrm{PL}(\theta)$ using the trivariate K-M and the marginal approximation at various sample sizes when $p = 3$

Another computational challenge is in maximizing $\log \mathrm{PL}(\theta)$ over θ because of the $2^p - 1$ multiple integrals in the expression. Due to using nonparametric estimates of η, these integrals are calculated empirically rather than analytically, at each iteration of the Newton-Raphson procedure to estimate θ. We noticed that the computational time increase dramatically with sample size n. In the example above with $p = 3$, the required time to estimate θ is less than 1 minute with $n = 100$ but increases to 30 minutes with $n = 400$, regardless of the estimator of η (Figure 1B). The computing time may be acceptable for obtaining point estimates of the coefficients but may be impractical for bootstrap variance estimation based on a large number of resampled datasets from the original data. Thus, we suggest using a reduced sample size when performing the bootstrap variance estimation as described in Section 3. The variance estimator can be adjusted by the ratio of sample sizes in the original data and the bootstrap sample. Asymptotically, this approach yields an unbiased variance estimator and achieves substantial reductions in computing time.

We further propose to reduce computational time via the Monte Carlo (MC) integration when evaluating the $2^p - 1$ multiple integrals. The MC integration is a numerical integration procedure that approximates a definite integral by evaluating the integrand at a set of points that are selected randomly. Since we are assuming that the components of η are independent, the MC integration may be performed by sampling uniformly over the possible values of η. For the simulation examples presented here, we randomly selected 100 000 points in η whenever there were over 1 000 000 possible values of η (the integrand was evaluating at all possible values of η when there were less than 1 000 000 possible values). In the previous example with $p = 3$ and $n = 400$, by applying the MC integration to the estimation, the computational time to estimate θ reduces significantly from 30 minutes to 20 seconds. Note that the MC integration was a multi-threaded process (2 cores were available). These simulations were performed using R 3.6.1 on a 2.9 GHz PC with 2 cores and 16 GB of RAM. With our experience, computational time is reasonable when we apply the marginal approximation approach with the MC integration for the estimation, and using small sample size for the bootstrap variance estimation.

3 | ASYMPTOTIC PROPERTIES

In this section, we establish the consistency and asymptotic normality of the pseudo-likelihood estimator $\hat{\theta}$ which is the solution to $\partial \log \mathrm{PL}(\theta)/\partial\theta = 0$. The asymptotic properties for the univariate case were shown in Kong and Nan.[13] In our work, the extension to arbitrary p-dimensional censored covariates involves nontrivial modifications of earlier proofs. This occurs, in part, because of the complicated nature of $\mathrm{PL}(\theta)$ and in part, because of the nonparametric estimation of the joint distribution of the errors in the AFT models. The derivatives of $\log \mathrm{PL}(\theta)$ cannot be written analytically with respect to η and care is needed to properly account for "noise" introduced by the estimation of η.

Let $\mathcal{Y}$ and $\mathcal{X}$ be the sample spaces of the response variable Y and the covariate $\mathbf{X}$, respectively. Denote $(\Theta, \mathcal{A}, \mathcal{H})$ as the parameter spaces of $(\theta, \boldsymbol{\alpha}, \eta)$. We redefine the design matrix $D = \{X_1, \ldots, X_q, h_1(T_1), \ldots, h_p(T_p)\}^T$ in the regression model as $D_{(\mathbf{X},\mathbf{T})}$, and let $\dot{h}_j(t_j) = dh_j(t_j)/dt$ and $\ddot{h}_j(t_j) = d^2h_j(t_j)/dt_j^2, j = 1, \ldots, p$. Denote the marginal distribution of ξ_j for the jth AFT model associated with α_{j0} as $\eta_{j,\alpha_{j0}}(s_j)$, and its first and second derivatives as $\dot{\eta}_j(s_j) = d\eta_{j,\alpha_{j0}}(s_j)/ds_j$ and $\ddot{\eta}_j(s_j) = d^2\eta_{j,\alpha_{j0}}(s_j)/ds_j^2, j = 1, \ldots, p$. Let the derivative of the log-likelihood with respect to θ of n observations $\{Y_i, \mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\Delta}_i\}_{i=1}^n$ be

$$\Psi_n(\theta, \phi, \boldsymbol{\alpha}, \eta) = \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i, \mathbf{X}_i, \mathbf{V}_i, \boldsymbol{\Delta}_i; \theta, \phi, \boldsymbol{\alpha}, \eta), \tag{6}$$

where $\psi(\cdot)$ is a random map for a single observation and is the derivative of (4) with respect to θ. Replacing $(\phi, \boldsymbol{\alpha}, \eta)$ with $(\hat{\phi}, \hat{\boldsymbol{\alpha}}, \hat{\eta})$ in (6), we have the pseudo-likelihood estimating equation $\Psi_n(\theta, \hat{\phi}, \hat{\boldsymbol{\alpha}}, \hat{\eta}) = 0$ where the solution is our proposed estimator $\hat{\theta}$. Denote $\Psi(\theta, \phi, \boldsymbol{\alpha}, \eta)$ as a deterministic function, defined as

$$\Psi(\theta, \phi, \boldsymbol{\alpha}, \eta) = E\{\psi(Y, \mathbf{X}, \mathbf{V}, \boldsymbol{\Delta}; \theta, \phi, \boldsymbol{\alpha}, \eta)\}.$$

In addition, for a function f of a random variable W which follows a distribution P, we define

$$Pf = \int f(w)dP(w), \quad \mathbb{P}_n f = n^{-1}\sum_{i=1}^{n} f(W_i), \quad \mathbb{G}_n f = n^{-1/2}(\mathbb{P}_n - P)f,$$

and p^* as an outer probability. The regularity conditions are listed below:

C1. $\mathbf{X}$ and $\mathbf{Z}$ are uniformly bounded.

C2. $\Psi(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0)$ has a unique solution in θ, θ_0.

C3. For any constant $U < \infty$, $\sup_{t_j \in [C_j, U]} |h_j(t_j)| \le c_0 < \infty$, $\sup_{t_j \in [C_j, U]} |\dot{h}_j(t_j)| \le c_1 < \infty$, and $\sup_{t_j \in [C_j, U]} |\ddot{h}_j(t_j)| \le c_2 < \infty$, where c_0, c_1 and c_2 are constants, $j = 1, \ldots, p$.

C4. (i) $\dot{\eta}_j(s_j)$ and $\ddot{\eta}_j(s_j)$ are bounded, and $\int (d \log \dot{\eta}_j(s_j)/ds_j)^2 \dot{\eta}_j(s_j) ds_j < \infty$, $j = 1, \ldots, p$; (ii) $\eta_{0,\boldsymbol{\alpha}_0}$ is continuously differentiable with bounded partial derivatives up to pth order.

C5. If $\hat{\boldsymbol{\alpha}}$ is a $n^{1/2}$-consistent estimator of $\boldsymbol{\alpha}_0$, $||\hat{\eta}_{\hat{\alpha}}(s_1, \ldots, s_p) - \eta_0(s_1, \ldots, s_p)|| \to 0$ in outer probability p^* and

$$\sup_{s_j \in [C_j - c_{3j}, \tau_j], j = 1, \ldots, p} |\hat{\eta}_{\hat{\alpha}}(s_1, \ldots, s_p) - \eta_0(s_1, \ldots, s_p)| = O_{p^*}(n^{-1/2}),$$

where $\sup_{\boldsymbol{\alpha} \in \mathcal{A}, \mathbf{x} \in \mathcal{X}} |\alpha_j^T \mathbf{x}| = c_{3j} < \infty$, $j = 1, \ldots, p$, and there exists a function $m_1(\alpha_0, \eta_0, \mathbf{X}, \mathbf{V}, \boldsymbol{\Delta})$ such that $\sqrt{n}(\hat{\eta}_{\hat{\alpha}} - \eta_0) = \mathbb{G}_n m_1(\alpha_0, \eta_0, \mathbf{X}, \mathbf{V}, \boldsymbol{\Delta}) + o_p(1)$.

C6. $a(\phi)$ is a monotone function with bounded derivatives $\dot{a}(\cdot)$ and $\ddot{a}(\cdot)$, satisfying that $|1/a(\phi)| \le c_4 < \infty$ for a constant c_4.

C7. $\dot{b}(\cdot)$ is a bounded monotone function and $\ddot{b}(\cdot)$ is a bounded Lipschitz function.

C8. Let $\mathcal{K} = \{y \in \mathcal{Y}, \theta \in \Theta, |1/a(\phi)| < c_4, \mathbf{x} \in \mathcal{X}, t_j \in [C_j, U], j = 1, \ldots, p\}$. There exist constants $k_j < \infty, j = 1, \ldots, 5$, such that (i) $\sup_{\mathcal{K}} |f_{\theta,\phi}(y|\mathbf{t}, \mathbf{x})[y - \dot{b}\{\theta^T D_{(\mathbf{x},\mathbf{t})}\}]| \le k_1$; (ii) $\sup_{\mathcal{K}} |\partial f_{\theta,\phi}(y|\mathbf{t}, \mathbf{x})/\partial \phi| \le k_2$; (iii) $\sup_{\mathcal{K}} |\left\{\partial f_{\theta,\phi}(y|\mathbf{t}, \mathbf{x})/\partial \phi\right\} [y - \dot{b}\{\theta^T D_{(\mathbf{x},\mathbf{t})}\}]| \le k_3$; (iv) $\sup_{\mathcal{K}} |\partial \left(f_{\theta,\phi}(y|t, x)[y - \dot{b}\{\theta^T D_{(\mathbf{x},\mathbf{t})}\}]\right)/\partial \mathbf{t}| \le k_4$; (v) $\sup_{\mathcal{K}} |\partial \left(f_{\theta,\phi}(y|t, x)[y - \dot{b}\{\theta^T D_{(\mathbf{x},\mathbf{t})}\}]\right)/\partial \theta| \le k_5$.

C9. (i) There exists constant truncation values $\tau_j < \infty$ such that $P(V_j - \alpha_j^T \mathbf{X} > \tau_j, j = 1, \ldots, p | \mathbf{X} = \mathbf{x}) \ge c_5 > 0$ for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_p] \in \mathcal{A}$; (ii) There exist constants $a_j > 0$ (for $j = 1, \ldots, p$), $a_{jk} > 0$ (for $j > k = 1, \ldots, p$), $\ldots, a_{1\cdots p} > 0$ and $c_6 > 0$ such that

$$\begin{aligned}
&\int_{C_j - \alpha_j^T \mathbf{X}}^{\tau_j} f_{\theta,\phi}(Y | s_j + \alpha_j^T \mathbf{X}, \mathbf{T}_{-j}, \mathbf{X}) \eta(ds_j, ds_l = T_l - \alpha_l^T \mathbf{X}, \forall l \ne j) \ge a_j, \\
&\int_{C_j - \alpha_j^T \mathbf{X}}^{\tau_j} \int_{C_k - \alpha_k^T \mathbf{X}}^{\tau_k} f_{\theta,\phi}(Y | s_j + \alpha_j^T \mathbf{X}, s_k + \alpha_k^T \mathbf{X}, \mathbf{T}_{-(j,k)}, \mathbf{X}) \\
&\quad \times \eta(ds_j, ds_k, ds_l = T_l - \alpha_l^T \mathbf{X}, \forall l \ne \{j, k\}) \ge a_{jk}, \\
&\quad \vdots \\
&\int_{C_1 - \alpha_1^T \mathbf{X}}^{\tau_1} \cdots \int_{C_p - \alpha_p^T \mathbf{X}}^{\tau_p} f_{\theta,\phi}(Y | s_j + \alpha_j^T \mathbf{X}, j = 1, \ldots, p, \mathbf{X}) \eta(ds_1, \ldots, ds_p) \ge a_{1\cdots p}
\end{aligned}$$

with probability 1 for any $\theta \in \Theta$, and $|\phi - \phi_0| + |\boldsymbol{\alpha} - \boldsymbol{\alpha}_0| + ||\eta - \eta_0|| < c_6$.

Condition C1 asserts the boundedness of covariates, which is often met in practice. Condition C2 is an identifiability condition, which ensures the consistency of the proposed estimator $\hat{\theta}$. We show that $\dot{\Psi}(\theta_0, \phi_0, \boldsymbol{\alpha}_0, \eta_0) = \partial \Psi(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0)/\partial \theta|_{\theta = \theta_0}$ is negative definite under Condition C5 in the proof of Theorem 1, which implies that $\dot{\Psi}(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0)$ is a continuous matrix of θ and also negative definite in a neighborhood of θ_0. This guarantees that θ_0 is the unique solution of $\Psi(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0) = 0$ in a neighborhood of θ_0. Since the initial value in the algorithm is obtained from the complete-case analysis, which is known to be $n^{1/2}$-consistent, the solution from the two-stage method should also be in the same neighborhood and consistent. Condition C3 holds for many commonly used transformation functions. Condition C4 is the usual assumption for multivariate AFT models. Condition C5 asserts that the estimator $\hat{\eta}_{\hat{\alpha}}$ is $n^{1/2}$-consistent and has a limiting normal distribution. The proofs are tedious for arbitrary $p > 1$ with generic $\hat{\boldsymbol{\alpha}}$ and $\hat{\eta}_{\hat{\alpha}}$. We state high level conditions which can be checked on a case by case basis. We note that condition C6 has been proved for $p = 1$ and should also hold for $p > 1$ with $n^{1/2}$-consistent estimators for $\boldsymbol{\alpha}_0$ and η_0. Conditions C6 to C8 assert the boundedness of various functions in the outcome model, which automatically hold for commonly used generalized linear models. Condition C9 is for technical convenience, which can be obtained by truncating the response variable Y such that $|Y| \le M < \infty$ for a large constant M, and then truncating the residuals in the AFT models with some constants $\tau_j' < \tau_j$, $j = 1, \ldots, p$. In the simulation section, satisfactory results are achieved without implementing such truncation steps.

Under the above regularity conditions, we establish the following theorem.

Theorem 1. *Suppose models (1) and (3) hold. Under the regularity conditions, the two-stage pseudo-likelihood estimator* $\hat{\theta}$*, satisfying* $\Psi(\hat{\theta}, \hat{\phi}, \hat{\boldsymbol{\alpha}}, \hat{\eta}_{\hat{\boldsymbol{\alpha}}}) = 0$*, converges in outer probability* p^* *to the true value,* θ_0*, and* $\sqrt{n}(\hat{\theta} - \theta_0)$ *converges weakly to a Gaussian distribution with mean zero and covariance matrix* Ω*, given in the supplementary material.*

To prove Theorem 1, we use the well-established Z-estimation theory in Nan and Wellner[31] and generalize the proof from $p = 1$[13] to an arbitrary p, with details in the supplementary material. Under the regularity conditions, it can be shown that $\Psi(\theta, \hat{\phi}, \hat{\boldsymbol{\alpha}}, \hat{\eta}_{\hat{\boldsymbol{\alpha}}})$ converges uniformly to $\Psi(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0)$, where $(\phi_0, \boldsymbol{\alpha}_0, \eta_0)$ are the true values of $(\phi, \boldsymbol{\alpha}, \eta)$, as $n \to \infty$. Thus, if θ_0 is the unique solution to $\Psi_n(\theta, \phi_0, \boldsymbol{\alpha}_0, \eta_0)$ in Θ, then $\hat{\theta}$ is consistent for θ_0. Via Lemma 2 in the supplementary material, we have the asymptotic linear representation for $n^{1/2}(\hat{\theta} - \theta_0)$ which provides the asymptotic normality. The variance matrix Ω is extremely complicated and cannot be derived analytically. We employ bootstrapping for variance estimation and related inferences, which performs well in Section 4.

4 | SIMULATION STUDIES

Extensive simulation studies were conducted to evaluate the finite-sample performance of the proposed methods. We first present simulations with $p = 2$ and 10, followed by a simulation study motivated by the LIFECODES study. We focus on continuous outcomes with linear outcome models here due to the length limitation of this paper. Simulation results with binary outcomes were similar and can be found in the supplementary material. For simplicity, we restricted to a uniform monotone decreasing transformation function $h_j(\cdot) = h(\cdot)$ for $j = 1, \ldots, p$.

We started with $p = q = 2$. The two fully observed covariates $(X_1, X_2)^T$ were generated from Ber(0.5) and $N(1, 1)$, respectively. The two covariates subject to LOD, $\mathbf{Z} = (Z_1, Z_2)^T$, were generated from $Z_1 = h(T_1)$ and $Z_2 = h(T_2)$, where

$$T_j = h^{-1}(Z_j) = \alpha_j^T \mathbf{X} + \xi_j, \quad \text{for } j = 1, 2,$$

and $\alpha_1 = (-0.25, -0.5, -0.25)^T$, $\alpha_2 = (-0.25, -0.25, -0.5)^T$, and $(\xi_1, \xi_2)^T$ followed a bivariate distribution η. The outcome Y was generated by $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma_1 Z_1 + \gamma_2 Z_2 + \epsilon$, where $\beta_0 = \beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 1$ and $\epsilon \sim N(0, 1)$. To evaluate the proposed method under different scenarios, we considered two transformation functions $h(\cdot)$ and two joint distributions η: $h(t) = -t$ (ie, $T_j = -Z_j$) or $h(t) = \exp(-t)$ (ie, $T_j = -\log(Z_j)$), and $\eta = \text{MVN}\{(0, 0)^T, \Sigma_1\}$ (multivariate normal) or $\eta = 0.5\text{MVN}\{(0, 0)^T, \Sigma_1\} + 0.5\text{MVN}\{(0, 0)^T, \Sigma_2\}$ (a mixture of multivariate normals), where

$$\Sigma_1 = 1/4^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \quad \text{and} \quad \Sigma_2 = 1/8^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

and $\rho = 0.5$. Furthermore, LOD_j was chosen to have 25% or 50% marginal censoring rate for $j = 1, 2$ with overall censoring rate around 40% or 70%, respectively. We generated samples with size 200 or 400, and repeated 1000 times.

For each simulated dataset, we implemented eight methods: analysis using the full data, complete-case analysis, substitution methods with three replacement values (LOD, LOD/$\sqrt{2}$, LOD/2) and three versions of our proposed two-stage approach. We used either rank-based method with Gehan weight[24,25] or by least-squares approaches or least-square method[26] to fit the semiparametric AFT models, and then estimated η with bivariate K-M estimator. We also employed the two-stage approach with marginal approximation by naively estimating η as a product of marginal distributions. The estimates of $\boldsymbol{\alpha}$ using either rank-based or least square based methods were similar. Thus, we only presented the results for the two-stage approach with marginal approximation using rank-based method for $\boldsymbol{\alpha}$ in this paper.

For each method, we computed the average bias, empirical standard error, mean of estimated standard deviations, and coverage rate of 95% confidence intervals (CI) for the regression parameter of interest, $\theta = (\beta^T, \gamma^T)^T$. For the proposed methods, we performed 200 bootstrap replicates with sample size equal to the original sample size to estimate the standard deviation. For the full data analysis, the complete-case approach and three substitution methods, the estimated standard deviations were obtained from the regression model. The simulation results for the scenario where $T_j = -\log(Z_j)$ and $\eta = \text{MVN}\{(0, 0)^T, \Sigma_1\}$ were given in Table 1. Results for the other scenarios are similar (Supplementary Tables S1-S3).

Substitution methods tended to yield biased estimates for regression parameters of both partially observed $\mathbf{Z}$ and fully observed $\mathbf{X}$, and biases increased as censoring rates increased. Complete-case analysis and two-stage approaches with bivariate K-M estimator both gave small biases, while the two-stage approach with marginal approximation yielded slightly larger biases as expected. In addition, the empirical standard errors with the proposed approaches were smaller

TABLE 1 Simulation results for linear outcome model where $p = q = 2$, $T = -\log(Z)$, and the multivariate normal as the error term distribution

			Bias					Empirical SE				
%<LOD	n	Method	β_0	β_1	β_2	γ_1	γ_2	β_0	β_1	β_2	γ_1	γ_2
25%	200	Full data	0.006	0.006	0.005	−0.007	0.002	0.181	0.189	0.130	0.126	0.089
		Complete cases	−0.003	0.014	0.009	−0.009	0.002	0.288	0.228	0.186	0.143	0.099
		LOD	−0.433	0.243	0.330	0.023	−0.088	0.210	0.188	0.125	0.133	0.089
		LOD/$\sqrt{2}$	0.160	0.060	0.108	−0.052	−0.046	0.182	0.192	0.134	0.121	0.089
		LOD/2	0.627	0.018	0.011	−0.153	−0.036	0.173	0.198	0.143	0.115	0.091
		2-Stage (rank)	0.014	0.000	−0.006	−0.005	0.004	0.191	0.192	0.135	0.128	0.090
		2-Stage (LS)	0.013	0.000	−0.006	−0.004	0.004	0.191	0.192	0.134	0.128	0.090
		2-Stage (marg)	−0.009	−0.008	−0.003	0.005	0.003	0.191	0.192	0.134	0.128	0.089
	400	Full data	−0.003	−0.001	0.001	0.004	−0.002	0.130	0.139	0.090	0.092	0.061
		Complete cases	0.004	−0.003	−0.002	0.004	−0.002	0.197	0.169	0.128	0.102	0.069
		LOD	−0.440	0.241	0.326	0.031	−0.091	0.148	0.139	0.088	0.097	0.061
		LOD/$\sqrt{2}$	0.150	0.055	0.103	−0.043	−0.048	0.128	0.140	0.092	0.089	0.060
		LOD/2	0.617	0.012	0.004	−0.144	−0.036	0.120	0.143	0.097	0.085	0.061
		2-Stage (rank)	0.003	−0.005	−0.008	0.007	−0.001	0.134	0.142	0.093	0.093	0.062
		2-Stage (LS)	0.003	−0.005	−0.008	0.007	−0.001	0.133	0.142	0.093	0.093	0.062
		2-Stage (marg)	−0.018	−0.014	−0.006	0.016	−0.002	0.134	0.141	0.093	0.093	0.061
50%	200	Full data	0.006	0.006	0.005	−0.007	0.002	0.181	0.189	0.130	0.126	0.089
		Complete cases	0.022	0.015	0.006	−0.016	0.003	0.494	0.316	0.261	0.179	0.123
		LOD	−1.426	0.627	0.716	0.066	−0.142	0.286	0.186	0.112	0.156	0.093
		LOD/$\sqrt{2}$	0.000	0.342	0.450	−0.091	−0.151	0.201	0.194	0.123	0.125	0.086
		LOD/2	0.942	0.271	0.362	−0.235	−0.194	0.175	0.204	0.135	0.110	0.087
		2-Stage (rank)	−0.019	−0.016	−0.012	0.015	0.003	0.227	0.210	0.145	0.141	0.095
		2-Stage (LS)	−0.019	−0.016	−0.013	0.015	0.004	0.226	0.209	0.144	0.141	0.096
		2-Stage (marg)	−0.067	−0.050	−0.029	0.040	0.008	0.225	0.210	0.144	0.139	0.094
	400	Full data	−0.003	−0.001	0.001	0.004	−0.002	0.130	0.139	0.090	0.092	0.061
		Complete cases	0.008	−0.003	0.000	0.003	−0.003	0.342	0.229	0.177	0.126	0.081
		LOD	−1.424	0.628	0.714	0.071	−0.147	0.203	0.139	0.080	0.112	0.065
		LOD/$\sqrt{2}$	−0.005	0.343	0.445	−0.087	−0.152	0.139	0.143	0.089	0.089	0.061
		LOD/2	0.936	0.275	0.357	−0.233	−0.191	0.121	0.149	0.097	0.079	0.063
		2-Stage (rank)	−0.017	−0.014	−0.011	0.019	−0.002	0.159	0.156	0.099	0.101	0.064
		2-Stage (LS)	−0.017	−0.014	−0.011	0.020	−0.002	0.159	0.156	0.098	0.101	0.064
		2-Stage (marg)	−0.064	−0.048	−0.027	0.045	0.002	0.159	0.154	0.098	0.100	0.064

(Continues)

TABLE 1 (Continued)

%<LOD	n	Method	Mean estimated SD					Coverage rate of 95% CIs				
			β_0	β_1	β_2	γ_1	γ_2	β_0	β_1	β_2	γ_1	γ_2
25%	200	Full data	0.181	0.190	0.128	0.129	0.087	0.951	0.958	0.957	0.965	0.950
		Complete cases	0.280	0.232	0.183	0.144	0.099	0.947	0.959	0.946	0.951	0.950
		LOD	0.209	0.188	0.119	0.137	0.088	0.457	0.730	0.226	0.954	0.824
		LOD/$\sqrt{2}$	0.177	0.194	0.126	0.126	0.086	0.860	0.946	0.840	0.938	0.915
		LOD/2	0.162	0.200	0.134	0.118	0.085	0.046	0.960	0.933	0.757	0.918
		2-Stage (rank)	0.189	0.196	0.133	0.132	0.090	0.946	0.957	0.945	0.962	0.947
		2-Stage (LS)	0.189	0.196	0.133	0.132	0.090	0.944	0.957	0.947	0.962	0.947
		2-Stage (marg)	0.189	0.195	0.132	0.131	0.089	0.942	0.957	0.951	0.959	0.947
	400	Full data	0.127	0.134	0.089	0.090	0.060	0.945	0.937	0.944	0.945	0.949
		Complete cases	0.196	0.163	0.128	0.100	0.068	0.947	0.940	0.942	0.947	0.946
		LOD	0.147	0.133	0.084	0.096	0.060	0.155	0.541	0.031	0.937	0.674
		LOD/$\sqrt{2}$	0.124	0.136	0.088	0.088	0.059	0.762	0.924	0.773	0.918	0.871
		LOD/2	0.114	0.141	0.094	0.083	0.059	0.000	0.949	0.946	0.578	0.912
		2-Stage (rank)	0.132	0.137	0.093	0.092	0.061	0.952	0.934	0.951	0.950	0.952
		2-Stage (LS)	0.132	0.137	0.093	0.092	0.061	0.953	0.935	0.948	0.955	0.951
		2-Stage (marg)	0.132	0.137	0.092	0.091	0.061	0.944	0.936	0.950	0.942	0.949
50%	200	Full data	0.181	0.190	0.128	0.129	0.087	0.951	0.958	0.957	0.965	0.950
		Complete cases	0.476	0.319	0.253	0.174	0.116	0.935	0.948	0.939	0.940	0.945
		LOD	0.287	0.187	0.113	0.159	0.095	0.000	0.074	0.000	0.934	0.666
		LOD/$\sqrt{2}$	0.194	0.193	0.119	0.128	0.084	0.941	0.584	0.047	0.902	0.567
		LOD/2	0.159	0.201	0.127	0.111	0.079	0.000	0.712	0.200	0.438	0.320
		2-Stage (rank)	0.231	0.216	0.145	0.146	0.096	0.956	0.968	0.949	0.968	0.960
		2-Stage (LS)	0.229	0.216	0.144	0.145	0.095	0.951	0.963	0.946	0.964	0.961
		2-Stage (marg)	0.231	0.214	0.144	0.144	0.094	0.951	0.960	0.945	0.958	0.955
	400	Full data	0.127	0.134	0.089	0.090	0.060	0.945	0.937	0.944	0.945	0.949
		Complete cases	0.331	0.223	0.176	0.120	0.079	0.941	0.930	0.946	0.941	0.939
		LOD	0.201	0.132	0.079	0.110	0.065	0.000	0.004	0.000	0.897	0.366
		LOD/$\sqrt{2}$	0.136	0.136	0.084	0.089	0.058	0.941	0.290	0.001	0.832	0.272
		LOD/2	0.112	0.142	0.089	0.078	0.055	0.000	0.501	0.027	0.153	0.087
		2-Stage (rank)	0.159	0.151	0.101	0.100	0.065	0.937	0.931	0.955	0.938	0.952
		2-Stage (LS)	0.157	0.150	0.100	0.100	0.065	0.938	0.933	0.955	0.942	0.954
		2-Stage (marg)	0.158	0.149	0.100	0.098	0.064	0.927	0.929	0.945	0.917	0.942

Note: Eight methods were implemented, including analysis with full data, complete-case approach, substitution methods with three replacement values: LOD, LOD/$\sqrt{2}$, and LOD/2, two-stage rank-based approach with multivariate KM estimator for η (2-stage (rank)), two-stage least-square approach with multivariate KM estimator for η (2-stage (LS)), and two-stage approach with marginal approximation (2-stage (marg)). Estimated SD* was from bootstrap for three version of the proposed method and from regression for full data analysis, complete-case approach and substitution methods.

than those for complete-case approach and only slightly larger than using the full dataset, which implied efficient use of data. Mean estimated standard deviations were close to the empirical standard errors with all approaches. The coverage rates could be very different from the nominal level with substitution methods, while at the nominal level for all three versions of our proposed approach. We explored the proposed method for smaller sample sizes ($n = 50$ and 100) and found similar patterns as with $n = 200$ and 400 when comparing across all the methods. All three versions of our proposed method had some slightly larger biases with smaller sample size but biases decreased as sample size increased, and coverage probabilities always remained close to 95% (Supplementary Table S4).

We further evaluated the influence of using the marginal approximation compared to using bivariate K-M estimator for η under different strengths of correlation. We fixed $T_j = -\log(Z_j)$, $\eta = \text{MVN}\{(0,0)^T, \Sigma_1\}$, marginal censoring rate at 50%, sample size 400, while varying ρ as 0.25, 0.5 or 0.75, where the corresponding correlation between Z_1 and Z_2 was 0.66, 0.73, or 0.80. The results were given in Table 2. The average bias using the two-stage marginal approach tended to be larger when ρ increased but still performed reasonably well even with $\rho = 0.75$ and these biases were much lower than using substitution methods. The coverage rates were close but slightly below the nominal level as ρ increased. This implied that the independence assumption may affect the results when ρ was large and the censoring rates were high, but all three versions of our proposed approach still performed much better than substitution methods.

When the dimension of $\mathbf{Z}$ is greater than 3, the computational burden becomes more severe. As discussed in Section 2.4, we recommend using the two-stage approach with marginal approximation and MC integration, and a smaller sample size in bootstrap variance estimation. Here, we presented a simulation with $p = 10$ to show the performance. Considering two fully observed covariates $\mathbf{X} = (1, X_1, X_2)^T$ where $X_1 \sim \text{Ber}(0.5)$ and $X_2 \sim N(1,1)$, we generated ten left-censored variables $\mathbf{Z} = (Z_1, \ldots, Z_{10})^T$ with $Z_j = h(T_j) = -T_j$ for $j = 1, \ldots, 10$, by

$$\mathbf{T} = \boldsymbol{\alpha}^T \mathbf{X} + \xi,$$

where $\mathbf{T} = (T_1, \ldots, T_{10})^T$,

$$\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_{10}] = \begin{bmatrix} -0.2 & -0.35 & -0.3 & -0.25 & -0.4 & -0.25 & -0.3 & -0.25 & -0.35 & -0.25 \\ -0.5 & -0.25 & -0.5 & -0.25 & -0.5 & -0.25 & -0.25 & -0.25 & -0.25 & -0.25 \\ -0.25 & -0.5 & -0.25 & -0.25 & -0.25 & -0.5 & -0.25 & -0.5 & -0.5 & -0.25 \end{bmatrix},$$

and $(\xi_1, \ldots, \xi_{10})^T \sim \eta = \text{MVN}\{(0, \ldots, 0)^T, \Sigma\}$ with

$$\Sigma = 1/2^2 \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_3 \end{pmatrix}.$$

Here $\mathbf{R}_1$ and $\mathbf{R}_3$ were 3×3 correlation matrices with all off-diagonal entries as 0.25 and 0.75, respectively, and $\mathbf{R}_2$ was a 4×4 correlation matrix with all off-diagonal entries as 0.5. This mimicked a practical situation that there was natural grouping of exposures, where exposures were correlated within groups but independent between groups. We let $Y = \beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \epsilon$, where all elements in β and γ were set as 1 and $\epsilon \sim N(0,1)$. The marginal censoring rate was set as 20% for all Z_j with an overall censoring rate around 64%. We generated data with sample size 400 and repeated the simulation 1000 times.

For each simulated dataset, we implemented six methods: analysis with the full data, complete-case approach, substitution methods with three replacement values (LOD, LOD/$\sqrt{2}$, LOD/2), and our proposed two-stage approach with marginal approximation. For our proposed two-stage approach with marginal approximation, the MC integration was applied to the estimation, and the standard deviations were estimated using 200 bootstrap replicates each with sample size 100, and then adjusted by a factor of 2. The results are shown in Table 3. Compared to substitution approaches, the two-stage approach performed well across regression coefficients for each covariate but had some remaining bias for the intercept, which was likely due to the independence assumption. Further increasing the sample size helped to reduce this bias (results not shown here). The two-stage marginal method was also more efficient than the complete-cases analysis with smaller standard errors. The adjusted bootstrap standard deviations were

TABLE 2 Simulation results for linear outcome model where $p = q = 2$, $T = -\log(Z)$, the multivariate normal as the error term distribution, $n = 400$, and marginal censoring rate at 50%

		Bias					Empirical SE				
ρ	**Method**	β_0	β_1	β_2	γ_1	γ_2	β_0	β_1	β_2	γ_1	γ_2
0.25	Full data	−0.005	0.001	−0.001	0.002	0.000	0.122	0.135	0.089	0.082	0.055
	Complete cases	−0.017	0.005	−0.004	0.005	0.001	0.333	0.236	0.186	0.111	0.075
	LOD	−1.331	0.671	0.728	0.005	−0.134	0.190	0.129	0.078	0.099	0.057
	LOD/$\sqrt{2}$	0.032	0.382	0.461	−0.125	−0.147	0.137	0.133	0.088	0.081	0.055
	LOD/2	0.953	0.305	0.377	−0.254	−0.196	0.120	0.140	0.097	0.073	0.058
	2-Stage (rank)	−0.027	−0.010	−0.011	0.014	0.002	0.155	0.148	0.098	0.092	0.058
	2-Stage (LS)	−0.027	−0.010	−0.011	0.014	0.002	0.155	0.148	0.099	0.093	0.058
	2-Stage (marg)	−0.052	−0.030	−0.022	0.028	0.006	0.156	0.148	0.098	0.092	0.057
0.5	Full data	−0.003	−0.001	0.001	0.004	−0.002	0.130	0.139	0.090	0.092	0.061
	Complete cases	0.008	−0.003	0.000	0.003	−0.003	0.342	0.229	0.177	0.126	0.081
	LOD	−1.424	0.628	0.714	0.071	−0.147	0.203	0.139	0.080	0.112	0.065
	LOD/$\sqrt{2}$	−0.005	0.343	0.445	−0.087	−0.152	0.139	0.143	0.089	0.089	0.061
	LOD/2	0.936	0.275	0.357	−0.233	−0.191	0.121	0.149	0.097	0.079	0.063
	2-Stage (rank)	−0.017	−0.014	−0.011	0.019	−0.002	0.159	0.156	0.099	0.101	0.064
	2-Stage (LS)	−0.017	−0.014	−0.011	0.020	−0.002	0.159	0.156	0.098	0.101	0.064
	2-Stage (marg)	−0.064	−0.048	−0.027	0.045	0.002	0.159	0.154	0.098	0.100	0.064
0.75	Full data	−0.009	−0.001	0.000	0.004	0.000	0.131	0.140	0.086	0.116	0.074
	Complete cases	−0.002	−0.004	0.000	0.003	−0.001	0.341	0.224	0.177	0.151	0.098
	LOD	−1.562	0.565	0.710	0.180	−0.183	0.206	0.131	0.076	0.132	0.075
	LOD/$\sqrt{2}$	−0.059	0.294	0.436	−0.033	−0.165	0.142	0.139	0.083	0.107	0.069
	LOD/2	0.911	0.244	0.338	−0.213	−0.185	0.122	0.146	0.093	0.093	0.070
	2-Stage (rank)	−0.015	−0.021	−0.011	0.024	−0.004	0.167	0.158	0.097	0.127	0.078
	2-Stage (LS)	−0.015	−0.021	−0.011	0.024	−0.004	0.167	0.158	0.097	0.127	0.078
	2-Stage (marg)	−0.084	−0.071	−0.027	0.068	−0.007	0.169	0.155	0.098	0.124	0.076

(Continues)

TABLE 2 (Continued)

		Mean estimated SD*					Coverage rate of 95% CIs				
ρ	**Method**	β_0	β_1	β_2	γ_1	γ_2	β_0	β_1	β_2	γ_1	γ_2
0.25	Full data	0.130	0.134	0.092	0.082	0.055	0.957	0.950	0.951	0.956	0.954
	Complete cases	0.339	0.234	0.185	0.113	0.073	0.951	0.949	0.947	0.955	0.944
	LOD	0.202	0.130	0.079	0.101	0.059	0.000	0.000	0.000	0.949	0.363
	LOD/$\sqrt{2}$	0.138	0.135	0.085	0.082	0.054	0.946	0.187	0.002	0.669	0.201
	LOD/2	0.113	0.141	0.090	0.072	0.051	0.000	0.407	0.020	0.062	0.035
	2-Stage (rank)	0.158	0.150	0.103	0.092	0.059	0.945	0.942	0.958	0.935	0.954
	2-Stage (LS)	0.158	0.149	0.103	0.092	0.059	0.946	0.940	0.958	0.937	0.954
	2-Stage (marg)	0.157	0.148	0.102	0.090	0.058	0.940	0.938	0.955	0.918	0.952
0.5	Full data	0.127	0.134	0.089	0.090	0.060	0.945	0.937	0.944	0.945	0.949
	Complete cases	0.331	0.223	0.176	0.120	0.079	0.941	0.930	0.946	0.941	0.939
	LOD	0.201	0.132	0.079	0.110	0.065	0.000	0.004	0.000	0.897	0.366
	LOD/$\sqrt{2}$	0.136	0.136	0.084	0.089	0.058	0.941	0.290	0.001	0.832	0.272
	LOD/2	0.112	0.142	0.089	0.078	0.055	0.000	0.501	0.027	0.153	0.087
	2-Stage (rank)	0.159	0.151	0.101	0.100	0.065	0.937	0.931	0.955	0.938	0.952
	2-Stage (LS)	0.157	0.150	0.100	0.100	0.065	0.938	0.933	0.955	0.942	0.954
	2-Stage (marg)	0.158	0.149	0.100	0.098	0.064	0.927	0.929	0.945	0.917	0.942
0.75	Full data	0.127	0.138	0.089	0.108	0.072	0.946	0.941	0.963	0.930	0.946
	Complete cases	0.335	0.220	0.177	0.145	0.096	0.951	0.944	0.951	0.942	0.949
	LOD	0.202	0.135	0.079	0.129	0.076	0.000	0.006	0.000	0.700	0.293
	LOD/$\sqrt{2}$	0.135	0.140	0.084	0.104	0.068	0.911	0.446	0.003	0.938	0.319
	LOD/2	0.112	0.145	0.089	0.089	0.063	0.000	0.604	0.037	0.332	0.187
	2-Stage (rank)	0.162	0.157	0.101	0.121	0.079	0.950	0.941	0.952	0.934	0.936
	2-Stage (LS)	0.161	0.156	0.100	0.121	0.079	0.947	0.938	0.950	0.936	0.930
	2-Stage (marg)	0.163	0.154	0.100	0.118	0.076	0.912	0.929	0.948	0.906	0.933

Note: Eight methods were implemented, including analysis with full data, complete-case approach, substitution methods with three replacement values: LOD, LOD/$\sqrt{2}$, and LOD/2, two-stage rank-based approach with multivariate KM estimator for η (2-stage (rank)), two-stage least-square approach with multivariate KM estimator for η (2-stage (LS)), and two-stage approach with marginal approximation (2-stage (marg)). Estimated SD* was from bootstrap for three version of the proposed method and from regression for full data analysis, complete-case approach and substitution methods.

TABLE 3 Simulation results for linear outcome model where $p = 10$, $q = 2$, $T = -Z$, the multivariate normal as the error term distribution, and $n = 400$

	Method	β_0	β_1	β_2	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	γ_9	γ_{10}
Bias	Full data	0.001	0.006	0.001	0.001	0.001	−0.005	−0.007	0.003	0.007	−0.002	−0.011	0.007	−0.003
	Complete cases	−0.009	−0.001	0.001	0.007	0.006	−0.009	−0.007	0.010	0.006	−0.002	−0.009	0.002	−0.001
	LOD	−2.023	0.399	0.656	0.111	−0.071	0.105	0.195	0.154	−0.121	0.199	−0.170	−0.146	0.509
	LOD/$\sqrt{2}$	−1.654	0.314	0.531	0.093	−0.053	0.085	0.153	0.141	−0.096	0.161	−0.146	−0.077	0.393
	LOD/2	−1.373	0.266	0.459	0.081	−0.051	0.068	0.130	0.117	−0.083	0.137	−0.133	−0.053	0.327
	2-Stage (marg)	−0.233	−0.094	−0.098	0.042	0.027	0.035	0.050	0.051	0.040	0.054	0.004	0.028	0.065
Empirical SE	Full data	0.114	0.128	0.109	0.108	0.108	0.108	0.128	0.127	0.127	0.121	0.167	0.163	0.167
	Complete cases	0.397	0.222	0.187	0.203	0.193	0.203	0.244	0.236	0.231	0.231	0.312	0.295	0.307
	LOD	0.209	0.173	0.165	0.152	0.153	0.158	0.189	0.186	0.180	0.172	0.234	0.229	0.247
	LOD/$\sqrt{2}$	0.192	0.166	0.159	0.144	0.142	0.147	0.179	0.173	0.170	0.162	0.221	0.214	0.235
	LOD/2	0.181	0.162	0.155	0.139	0.136	0.140	0.172	0.164	0.164	0.156	0.213	0.204	0.226
	2-Stage (marg)	0.161	0.148	0.126	0.127	0.124	0.124	0.153	0.151	0.147	0.141	0.190	0.189	0.198
Mean estimated SD*	Full data	0.112	0.133	0.109	0.107	0.107	0.107	0.129	0.129	0.129	0.128	0.170	0.170	0.170
	Complete cases	0.380	0.228	0.187	0.205	0.197	0.204	0.239	0.234	0.229	0.239	0.298	0.298	0.308
	LOD	0.201	0.173	0.131	0.167	0.164	0.167	0.199	0.197	0.193	0.199	0.255	0.255	0.254
	LOD/$\sqrt{2}$	0.180	0.167	0.128	0.157	0.152	0.155	0.188	0.181	0.181	0.186	0.239	0.236	0.241
	LOD/2	0.167	0.163	0.126	0.151	0.144	0.147	0.181	0.171	0.174	0.178	0.228	0.224	0.232
	2-Stage (marg)	0.179	0.175	0.147	0.146	0.145	0.147	0.175	0.172	0.171	0.174	0.231	0.231	0.230
Coverage rate of 95% CIs	Full data	0.942	0.956	0.962	0.945	0.954	0.947	0.954	0.956	0.950	0.964	0.956	0.965	0.956
	Complete cases	0.938	0.958	0.951	0.940	0.954	0.946	0.945	0.936	0.945	0.961	0.943	0.951	0.964
	LOD	0.000	0.357	0.006	0.925	0.949	0.917	0.855	0.895	0.927	0.856	0.915	0.940	0.498
	LOD/$\sqrt{2}$	0.000	0.536	0.027	0.933	0.960	0.921	0.892	0.895	0.942	0.892	0.917	0.958	0.654
	LOD/2	0.000	0.637	0.077	0.934	0.953	0.935	0.908	0.906	0.935	0.912	0.919	0.959	0.720
	2-Stage (marg)	0.766	0.952	0.924	0.961	0.976	0.975	0.970	0.968	0.973	0.975	0.987	0.976	0.962

Note: Six methods were implemented, including analysis with full data, complete-case approach, substitution methods with three replacement values: LOD, LOD/$\sqrt{2}$, and LOD/2, and two-stage approach with marginal approximation (2-stage (marg)). Estimated SD* was from bootstrap with sample size 100 for the two-stage marginal approach and from regression for full data analysis, complete-case approach and substitution methods.

slightly larger than the empirical standard errors. One way to improve the estimate of standard deviation was to use original sample size 400 in the bootstrapping. The computing time for each bootstrap replicate with sample size 100 and 400 were 15 and 90 seconds, respectively. Considering the massive reduction in computational time with sample size 100, the variance estimation results were acceptable, with reasonable coverage probabilities for 95% CI. We also considered a scenario for $Z_j = h(T_j) = \exp(-T_j)$, $j = 1, \ldots, 10$ and found the results to be similar (Supplementary Table S5).

Another solution we suggested for reducing the computational time under high dimensional covariates subject to LOD is to carefully choose the covariates to be included in $\mathbf{Z}$, while treating the other covariates subject to LOD either as binary or using substitution values. We conducted a simulation based on the LIFECODES data to investigate the performance of the proposed method with this preselection. In this simulation, we bootstrapped the fully observed demographic variables: baseline maternal age, race, education, insurance, prepregnancy BMI, and gestational age and specific gravity at the third trimester visit, with sample size $n = 252$, and generated the 17 metal values through AFT models within each bootstrapped sample plus an error term from a multivariate normal distribution with variance range from 0.05 to 1.00), and correlation range from −0.07 to 0.70, as estimated from the LIFECODES dataset. The outcome Y was simulated with a linear regression on both the demographic variables and the metals. The regression coefficients for both the AFT models and the outcomes were from data analysis results in Section 5. As mentioned before, of the 17 metals, only 3 were fully observed, and 3 had less than 5% of values below LOD, which were substituted by LOD/$\sqrt{2}$. Four other metals had more than 70% of values below LOD, and they were dichotomized to indicator variables of whether above LOD or not (0: below LOD; 1: above LOD). The LOD value for each metal was set to have the similar percentage of values below LOD for the metal in the real data, by generating the corresponding metals with a massive sample size and calculating the percentiles. We applied five methods, including complete-case approach, substitution methods with three replacement values (LOD, LOD/$\sqrt{2}$, LOD/2), and our proposed two-stage approach with marginal approximation, to analyze 1000 simulated data sets, and results were given in Supplementary Table S6. We noticed the results were consistent with the previous simulations, under this more practical simulation setting. Our proposed method was subject to limited biases for all the metals, no matter if they were included in $\mathbf{Z}$ or not, and the efficiency was improved as compared to the complete-case analysis. In this setting, the substitution methods performed reasonable for most covariates, but the performance was less stable than our proposed method, with some large biases and low coverage probabilities.

5 | DATA ANALYSIS

We illustrate our proposed method with data from the LIFECODES birth cohort[1,3,32] of women delivered at the Brigham and Women's Hospital in Boston, MA, during 2006 to 2008.We focus on a subset of 252 women who delivered full term and had urinary trace metal measurements at their third trimester, to explore the linear relationship between 17 urinary trace metals (arsenic (As), barium (Ba), beryllium (Be), cadmium (Cd), copper (Cu), chromium (Cr), mercury (Hg), manganese (Mn), molybdenum (Mo), nickel (Ni), lead (Pb), selenium (Se), tin (Sn), thallium (Tl), uranium (U), tungsten (W), and zinc (Zn)) and a urinary oxidative stress biomarker, 8-isoprostane. Among these metals, only 3 (As, Mo, Zn) were detected in all samples, 3 (Se, Ba, Mn) had less than 5% measurements below LOD, 7 (Sn, Hg, Cu, Ni, Tl, Pb, Cd) had 5% to 70% values below LOD, and 4 (W, Cr, U, Be) had more than 70% values below LOD (Supplementary Table S8). For the four metals with heavy censoring, we dichotomized them into binary variables: 1 if observed; 0 otherwise. Thus, the overall censoring rate of all metals was 65.3%. We further log transformed the concentration of the 13 continuous metals and 8-isoprostane for normality. Pairwise correlations between the 13 log-transformed continuous metal concentrations were between −0.13 and 0.79 based on complete data (Supplementary Figure S1), and correlation between the four binary-type metals were between 0.10 and 0.41. In a previous analysis, Kim et al[32] found five metals (Se, Mn, Cu, Tl, Be) to be associated with 8-isoprostane while adjusting for demographic covariates and replacing values below LOD with LOD/$\sqrt{2}$ in the 13 continuous metals concentrations and 8-isoprostane, with a nested case-control cohort of 92 women who delivered preterm and 269 women who delivered full term. Here, we present a reanalysis of this dataset but restricting to women delivered full term with detected value in 8-isoprostane and fully observed demographics variables, while accounting for the LOD issue in the metal measurements.

Here we considered the 7 metals with 5-70% values below LOD as $\mathbf{Z}$ ($p = 7$). We further replaced the values below LOD in Se, Ba, and Mn by their corresponding LOD/$\sqrt{2}$ and included them as fully observed variables $\mathbf{X}$, together with the 3

fully observed metals, 4 binary metals, and baseline demographic variables. Specifically, demographic variables included: baseline maternal age, race, education, insurance, prepregnancy BMI, and gestational age and specific gravity at the third trimester visit (Supplementary Table S9).

Results with complete-case analysis, substitution method with LOD/$\sqrt{2}$ for all continuous metals, and our proposed two-stage approach with marginal approximation, are summarized in Table 4. We observe that 8-isoprostane increases as Mn and Cu increase, and as Tl decreases with the substitution method ($\hat{\theta}$: 0.121, 0.536, −0.180, respectively) and our proposed two-stage approach performed similarly while Tl is marginally significant ($\hat{\theta}$: 0.116, 0.579, −0.110, respectively). This generally agrees with Kim et al.[32] Pairwise correlations between the residual terms in the 7 AFT models were between −0.045 and 0.372 (Supplementary Figure S2) suggesting the two-stage marginal approximation approach is reasonable. Instead, the complete case analysis identified Zn, Ba, Cu, and Tl to be significantly associated with 8-isoprostane. Although complete-case analysis is unbiased in theory, the results could be unreliable in practice due to small sample size ($n = 83$ in this analysis). The estimated standard deviations from the two-stage marginal approximation approach are much smaller than those from the complete-case analysis which suggests efficiency gain. In this analysis, point estimates for metals with our proposed method are relatively similar to those from the substitution method, which suggests the choice of LOD/$\sqrt{2}$ is not bad. However, in other applications, it may not be the case and could be challenging to decide "the most appropriate" for the replacement value due to lack of observations below LOD.

We did another analysis by further including Se, Ba, and Mn in $\mathbf{Z}$ ($p = 10$), to understand if we gain additional information. Results are similar to those in the previous analysis (Table 4): Mn, Cu, Tl are significantly associated with 8-isoprostane ($\hat{\theta}$: 0.118, 0.628, −0.144, respectively). This comparison shows that the substitution of the three metals (Se, Ba, Mn) has a minimal impact on the results since only 10 women were affected.

TABLE 4 Results of the estimated coefficients and 95% confidence intervals for the 17 metals in 8-isoprostane analysis

		Estimate (95% confidence interval)			
Metal	**<LOD%**	**Complete cases ($n = 92$)**	**LOD/$\sqrt{2}$**	**2-Stage (marg)-p7**	**2-Stage (marg)-p10**
As	0	−0.081 (−0.228, 0.067)	−0.039 (−0.123, 0.045)	−0.003 (−0.109, 0.103)	0.002 (−0.101, 0.106)
Mo	0	−0.110 (−0.406, 0.185)	0.046 (−0.108, 0.199)	0.058 (−0.126, 0.242)	0.043 (−0.117, 0.203)
Zn	0	−0.296 (−0.576, −0.015)*	−0.052 (−0.193, 0.088)	0.005 (−0.154, 0.164)	0.033 (−0.110, 0.177)
Se	0.4	0.266 (−0.226, 0.757)	0.113 (−0.169, 0.396)	0.305 (−0.026, 0.635)	0.173 (−0.087, 0.434)
Ba	1.6	−0.291 (−0.521, −0.061)*	−0.017 (−0.122, 0.089)	0.039 (−0.066, 0.143)	0.035 (−0.129, 0.199)
Mn	2.0	0.036 (−0.124, 0.196)	0.121 (0.021, 0.221)*	0.116 (0.003, 0.229)*	0.118 (0.006, 0.231)*
Sn	5.6	0.074 (−0.074, 0.222)	0.050 (−0.030, 0.130)	0.004 (−0.090, 0.098)	0.007 (−0.084, 0.099)
Cu	6.7	0.544 (0.038, 1.049)*	0.536 (0.279, 0.794)*	0.579 (0.337, 0.821)*	0.628 (0.314, 0.941)*
Hg	8.3	−0.125 (−0.306, 0.057)	0.073 (−0.015, 0.160)	0.076 (−0.046, 0.198)	0.106 (−0.021, 0.234)
Ni	13.9	0.364 (−0.030, 0.759)	−0.041 (−0.197, 0.116)	0.026 (−0.162, 0.214)	0.056 (−0.153, 0.266)
Tl	15.5	−0.250 (−0.541, 0.040)	−0.180 (−0.273, −0.088)*	−0.110 (−0.230, 0.011)	−0.144 (−0.275, −0.012*)
Pb	25.8	0.274 (0.011, 0.537)	0.028 (−0.081, 0.138)	0.052 (−0.087, 0.191)	0.040 (−0.087, 0.166)
Cd	58.3	0.131 (−0.081, 0.343)	0.014 (−0.110, 0.137)	−0.029 (−0.160, 0.101)	−0.034 (−0.160, 0.092)
W	79.4	−0.032 (−0.347, 0.282)	−0.104 (−0.310, 0.102)	−0.108 (−0.376, 0.160)	−0.088 (−0.356, 0.179)
Cr	88.1	−0.201 (−0.534, 0.132)	−0.135 (−0.397, 0.127)	−0.041 (−0.339, 0.257)	−0.140 (−0.434, 0.154)
U	89.3	0.143 (−0.276, 0.562)	0.171 (−0.138, 0.480)	0.128 (−0.207, 0.463)	0.153 (−0.228, 0.535)
Be	90.1	0.066 (−0.359, 0.490)	−0.207 (−0.510, 0.096)	−0.198 (−0.579, 0.183)	−0.214 (−0.540, 0.112)

Note: Implemented methods including complete-case approach, substitution method with LOD/$\sqrt{2}$ as the replacement value and two-stage approach with marginal approximation for $p = 7$ or 10 (2-stage (marg)-p7, 2-stage (marg)-p10). Estimated standard deviations were from bootstrap for the proposed method and from regression for complete-case approach and substitution method.

*Result for this metal is significant.

6 | DISCUSSION

We proposed a two-stage semiparametric approach to carefully address the issue of multiple covariates subject to LOD in a generalized linear outcome model. Substitution method, although convenient to use in practice, could result in large biases if the value is not a good representation of the left-tail of the distribution. Our proposed method, instead, provides unbiased results by estimating the joint distribution of the censored covariates semiparametrically. While our proposed approach is computationally challenging when the number of covariates subject to LOD is large, we suggested several solutions which ease the computational burden for practical usage. In addition, we recommended using our proposed method to the covariates with censoring rates between 5% and 70%, while applying the substitution method to the covariates with very low censoring rates and dichotomize covariates with heavy censoring rates. Although the magnitude and interpretation for the corresponding coefficients may be biased, the directions of their effects keep the same. Our simulation studies with large p suggested reasonable computational time and appropriate performance when applying these solutions.

Consistent estimation of the coefficients in the semiparametric AFT model can be obtained using either a rank-based or least-square approach. The least-squares approach requires longer computing time. For example, when $p = 5$, $q = 2$, and $n = 400$, the rank-based method requires 3 seconds in the estimation of AFT parameters but the least square approach needs 1 minute (Supplementary Table S10). Thus, we recommend conducting the rank-based method if p is considerably large and the AFT model involves many fully observed covariates. Furthermore, for the estimation of the joint distribution of residuals in the AFT models, the marginal approach which assumes independence is convenient to compute for high dimensional p but may provide biased results in the intercept term when strong correlations between covariates exist. A possible way to improve the marginal approach is to consider the pairwise correlation in the joint distribution which allows some dependence and use the composite pairwise likelihood approach for the joint distribution.

A number of methodologic extensions are of interest. Our approach is readily applicable when there are higher orders of covariates in the outcome model given the AFT models hold. Such approach can also be applied to explore the interaction between these covariates. In addition, variable selection for high-dimensional covariates subject to LOD is important in model building and is under investigation. The use of non-generalized linear models, such as survival regression and quantile regression, would be practically useful. Adaptation of the proposed two-stage methods to such settings is a topic of future research. Furthermore, the proposed methods are not directly applicable to case-control and case-cohort designs, where covariates are sampled conditionally on outcome status. Such designs are commonly employed when covariates are expensive to measure, which is important for further exploration. Finally, the topic of appropriately handling LOD is of particular interest in environmental mixture analysis, due to the potentially high correlation between components in the mixture. Our approach makes full use of the dependency between these environmental exposures to improve accuracy and efficiency, while do not make strict assumptions about their joint distribution. Although motivated by environmental mixture analysis, the proposed methods are generally applicable to any biomarker studies that may have multiple correlated biomarkers subject to LOD.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZIA ES103307). We would like to thank Drs. Clarice Weinberg and Greg Dinse at the National Institution of Environmental Health Sciences, NIH for their helpful suggestions, and the scientific computing support provided by Sciome LLC.

DATA AVAILABILITY STATEMENT

R code for the proposed methods with a sample example can be found at https://github.com/lingwanchen/Semi-mLOD.

ORCID

Ling-Wan Chen https://orcid.org/0000-0003-1343-8469

REFERENCES

1. Ferguson KK, McElrath TF, Meeker JD. Environmental phthalate exposure and preterm birth. *JAMA Pediatr*. 2014;168(1):61-67.
2. Ferguson KK, McElrath TF, Chen Y-H, Loch-Caruso R, Mukherjee B, Meeker JD. Repeated measures of urinary oxidative stress biomarkers during pregnancy and preterm birth. *Am J Obstet Gynecol*. 2015;212(2):208-e1.

3. Kim SS, Meeker JD, Carroll R, et al. Urinary trace metals individually and in mixtures in association with preterm birth. *Environ Int*. 2018;121:582-590.
4. D'Angelo G, Weissfeld L, GenIMS Investigators. An index approach for the Cox model with left censored covariates. *Stat Med*. 2008;27(22):4502-4514.
5. Nie L, Chu H, Liu C, Cole SR, Vexler A, Schisterman EF. Linear regression with an independent variable subject to a detection limit. *Epidemiology*. 2010;21(Suppl 4):S17.
6. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. Hoboken: Wiley; 2019.
7. Helsel DR. *Nondetects and Data Analysis. Statistics for Censored Environmental Data*. Hoboken: Wiley-Interscience; 2005.
8. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol*. 2003;157(4):355-363.
9. Arunajadai SG, Rauh VA. Handling covariates subject to limits of detection in regression. *Environ Ecol Stat*. 2012;19(3):369-391.
10. Cole SR, Chu H, Nie L, Schisterman EF. Estimating the odds ratio when exposure has a limit of detection. *Int J Epidemiol*. 2009;38(6):1674-1680.
11. Sattar A, Sinha SK, Wang X-F, Li Y. Frailty models for pneumonia to death with a left-censored covariate. *Stat Med*. 2015;34(14):2266-2280.
12. Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. *Biometrics*. 2008;64(3):895-903.
13. Kong S, Nan B. Semiparametric approach to regression with a covariate subject to a detection limit. *Biometrika*. 2016;103(1):161-174.
14. Atem FD, Qian J, Maye JE, Johnson KA, Betensky RA. Linear regression with a randomly censored covariate: application to an Alzheimer's study. *J Royal Stat Soc Ser C (Appl Stat)*. 2017;66(2):313-328.
15. Atem FD, Matsouaka RA, Zimmern VE. Cox regression model with randomly censored covariates. *Biom J*. 2019;61(4):1020-1032.
16. Ding Y, Kong S, Kang S, Chen W. A semiparametric imputation approach for regression with censored covariate with application to an AMD progression study. *Stat Med*. 2018;37(23):3293-3308.
17. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Stat Med*. 2011;30(20):2551-2561.
18. Wu H, Chen Q, Ware LB, Koyama T. A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: an application to acute lung injury. *J Appl Stat*. 2012;39(8):1733-1747.
19. Chen Q, Wu H, Ware LB, Koyama T. A Bayesian approach for the Cox proportional hazards model with covariates subject to detection limit. *Int J Stat Med Res*. 2014;3(1):32.
20. Lee M, Kong L, Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. *Stat Med*. 2012;31(17):1838-1848.
21. Bernhardt PW, Wang HJ, Zhang D. Flexible modeling of survival data with covariates subject to detection limits via multiple imputation. *Comput Stat Data Anal*. 2014;69:81-91.
22. Bartlett JW, Seaman SR, White IR, Carpenter JR, Alzheimer's Disease Neuroimaging Initiative*. Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Stat Methods Med Res*. 2015;24(4):462-487.
23. Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank tests. *Biometrika*. 1990;77(4):845-851.
24. Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. *Biometrika*. 2003;90(2):341-353.
25. Jin Z, Lin DY, Ying Z. Rank regression analysis of multivariate failure time data based on marginal linear models. *Scand J Stat*. 2006;33(1):1-23.
26. Jin Z, Lin DY, Ying Z. On least-squares regression with censored data. *Biometrika*. 2006;93(1):147-161.
27. Chiou SH, Kang S, Kim J, Yan J. Marginal semiparametric multivariate accelerated failure time model with generalized estimating equations. *Lifetime Data Anal*. 2014;20(4):599-618.
28. Prentice RL, Zhao S. Nonparametric estimation of the multivariate survivor function: the multivariate Kaplan–Meier estimator. *Lifetime Data Anal*. 2018;24(1):3-27.
29. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.
30. Dabrowska DM. Kaplan-Meier estimate on the plane. *Ann Stat*. 1988;16(4):1475-1489.
31. Nan B, Wellner JA. A general semiparametric Z-estimation approach for case-cohort studies. *Stat Sin*. 2013;23(3):1155.
32. Kim SS, Meeker JD, Keil AP, et al. Exposure to 17 trace metals in pregnancy and associations with urinary oxidative stress biomarkers. *Environ Res*. 2019;179:108854.

SUPPORTING INFORMATION

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

How to cite this article: Chen L-W, Fine JP, Bair E, et al. Semiparametric analysis of a generalized linear model with multiple covariates subject to detection limits. *Statistics in Medicine*. 2022;41(24):4791-4808. doi: 10.1002/sim.9536