Annals of Work Exposures and Health, 2024, **68**, 846–858
https://doi.org/10.1093/annweh/wxae068
Advance access publication 14 August 2024
Original Article

BOHS British Occupational Hygiene Society
The Chartered Society for Worker Health Protection
OXFORD

Quantile regression for longitudinal data with values below the limit of detection and time-dependent covariates—application to modeling carbon nanotube and nanofiber exposures

I-Chen Chen*, Stephen J. Bertke, Matthew M. Dahm

Division of Field Studies and Engineering, National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, 1090 Tusculum Avenue, MS-R14, Cincinnati, OH 45226, USA

*Corresponding author: Email: okv0@cdc.gov

Abstract

Background: In studies of occupational health, longitudinal environmental exposure, and biomonitoring data are often subject to right skewing and left censoring, in which measurements fall below the limit of detection (LOD). To address right-skewed data, it is common practice to log-transform the data and model the geometric mean, assuming a log-normal distribution. However, if the transformed data do not follow a known distribution, modeling the mean of exposure may result in bias and reduce efficiency. In addition, when examining longitudinal data, it is possible that certain covariates may vary over time.

Objective: To develop predictive quantile regression models to resolve the issues of left censoring and time-dependent covariates and to quantitatively evaluate if previous and current covariates can predict current and/or future exposure levels.

Methods: To address these gaps, we suggested incorporating different substitution approaches into quantile regression and utilizing a method for selecting a working type of time dependency for covariates.

Results: In a simulation study, we demonstrated that, under different types of time-dependent covariates, the approach of multiple random value imputation outperformed the other approaches. We also applied our methods to a carbon nanotube and nanofiber exposure study. The dependent variables are the left-censored mass of elemental carbon at both the respirable and inhalable aerosol size fractions. In this study, we identified some potential time-dependent covariates with respect to worker-level determinants and job tasks.

Conclusion: Time dependency for covariates is rarely accounted for when analyzing longitudinal environmental exposure and biomonitoring data with values less than the LOD through predictive modeling. Mistreating the time-dependency as time-independency will lead to an efficiency loss of regression parameter estimation. Therefore, we addressed time-varying covariates in longitudinal exposure and biomonitoring data with left-censored measurements and illustrated an entire conditional distribution through different quantiles.

Key words: Biomonitoring; environmental exposure; left censoring; limit of detection; quantile regression; time-dependent covariate.

What's Important About This Paper?

Longitudinal environmental exposure and biomonitoring data are subject to right skewing and left censoring. To address the potential bias when these data do not follow a known distribution, marginal quantile regression can be implemented using different substitution approaches. This study also discusses an approach to handle time-dependent covariates. Incorrectly treating the time-dependency as time-independency will result in an efficiency loss of regression parameter estimation.

Received: October 26, 2023. Accepted: July 18, 2024.

Introduction

Statistical models used to identify workplace determinants of exposure have been increasingly used in occupational exposure studies (Jin et al. 2011; Dahm et al. 2019). The dependent variable in these models is environmental exposure or biomonitoring data collected on workers over time, i.e. the concentration level of a metabolite or analyte in urine, blood, or other biological matrices, or a personal environmental air or dermal sample. The independent variables measured over time could consist of many different job-related characteristics, such as work tasks, direct or indirect exposure, or quantity of materials handled. The longitudinal exposure data are often skewed to the right and assumed to follow a log-normal distribution (Leidel et al. 1977). However, if the logged data do not follow a known distribution, modeling the conditional mean of the log-normal exposure outcome might not be ideal because the estimated mean and standard deviation will be subject to extreme values. Also, the use of geometric mean, i.e. the exponentiated mean of the logged data, can be questionable when the log-transformed data are asymmetric (Helsel 2006). Therefore, quantile regression for correlated outcomes is recommended as an alternative providing a complete illustration of a dependent variable's entire conditional distribution. This model has advantages compared to mean regression in that it makes no parametric assumptions on the error distribution and is robust to potential outliers (Tang and Leng 2011; Fu *et al.* 2015).

Quantile regression modeling can be complicated when workers are sampled over time and in the presence of left censoring. Left-censored exposure data occurs when laboratory instrument procedure has a limit of detection (LOD) below which, no observed value is given. Analysis of nondetected or left-censored exposure data has been increasingly discussed in occupational health (Jin et al. 2011). Most recently, the marginal quantile model (Chen et al. 2021) incorporating a substitution method of LOD/2 with an estimating equations approach (Fu et al. 2015) has been shown to perform well through simulations for left-censored exposure data that are not log-normally distributed and are heavily right-skewed with low correlated repeated measures. This model is also desirable for log-normal outcome data with large censoring and high correlation. However, the substitution method of LOD/2 or LOD/$\sqrt{2}$ lacks a unique replaced value, potentially leading to bias in regression parameter estimation, particularly in scenarios with a high percentage of censoring (Hornung and Reed 1990). Despite this limitation, the substitution method (Hornung and Reed 1990; Burstyn and Teschke 1999) remains widely utilized by industrial hygienists for assigning values to measurements falling below the LOD in occupational exposure analyses. Moreover, this substitution approach is not advisable unless censored values are less than 10% of the data set (Lubin et al. 2004). Instead, the use of a multiple imputation technique might be appropriate and can result in unbiased estimates and nominal confidence intervals (Lubin et al. 2004). Recently, the β-substitution method was shown to have superior results, in terms of bias, relative to the LOD/2 or LOD/ $\sqrt{2}$ substitution method (Ganser and Hewett 2010). It also performed as well or better than the maximum likelihood estimation and Kaplan–Meier methods under high censoring (<50%) and small sample sizes ($n < 20$) (Huynh et al. 2014).

Methods used for longitudinal data in the presence of time-dependent covariates have been developed for the marginal modeling of the mean (Pepe and Anderson 1994; Lai and Small 2007; Chen and Westgate 2017) and the marginal quantile regression model (Chen and Westgate 2021), in which the median and different quantiles are modeled. The estimating equations, and thus estimates, can be unbiased when employing an independence working correlation structure in the presence of time-dependent covariates (Pepe and Anderson 1994). Furthermore, the regression parameter estimation can be inefficient because not all valid moment conditions are adequately utilized (Fitzmaurice 1995; Wang and Carey 2003). Therefore, several approaches have been proposed to improve estimation efficiency by examining the valid moments corresponding to 4 existing types of time-dependent covariates (Lai and Small 2007; Zhou et al. 2014; Chen and Westgate 2017). Although these approaches require analysts to presume the type of time dependency, this presumption is often unknown in practice. As a result, methods are further developed via an empirical likelihood of weighing moment conditions that are not assured to have consistent estimation (Leung et al. 2013), a hypothesis testing of a covariate type (Lai and Small 2007), and a correlation testing of each moment (Lalonde et al. 2014). However, these testing techniques might lead to high type II error rates because too many moment conditions are deemed valid. To avoid biased regression parameter estimation, a criterion accounting for both the bias and efficiency of regression parameter estimation with respect to time-dependent covariates and minimizing mean squared error (MSE) has been proposed (Chen and Westgate 2019, 2021). This criterion also provides consistent estimation. To the best of our knowledge, in the literature on environmental exposure and biomonitoring data, time-dependent covariates have rarely been analyzed, they are generally treated as time-independent instead. Specifically, we are interested in if the determinants may change over time and cause feed-back effects from the exposure outcome or dependent variable.

In this manuscript, we propose an extension to the approach of marginal quantile regression for longitudinal exposure data featuring nondetects (Chen et al. 2021). This extension involves the incorporation of multiple fill-in methods to handle measurements less than the LOD. Specifically, these methods assign a single value using LOD/2, LOD/$\sqrt{2}$, or β-substitution method (Ganser and Hewett 2010). Additionally, a series of values, including a multiple random value imputation technique (Lubin et al. 2004) and a method imputing values for censored data based on the linear equation of a quantile-quantile (QQ) plot (Pleil 2016a, 2016b). With these approaches for longitudinal data containing time-varying covariates, we employ the MSE criteria to select a working classification type of time-dependency (Chen and Westgate 2019, 2021). We conduct a simulation study to compare the estimation performances of the marginal quantile regression model with different fill-in methods, relative to estimating equations with an independence working structure, under different types of time-dependency and a range of LOD proportions, and to assess the MSE's utility. Lastly, we demonstrate the proposed methods by applying them to a dataset of environmental workplace exposures to carbon nanotube and nanofibers (CNT/F). This longitudinal dataset, with left censoring, features the dependent variables of elemental carbon (EC) mass at both the respirable and inhalable aerosol size fractions and the independent variables, or potential time-dependent covariates, of worker-level determinant and job task information that have been measured over time.

Methods

Quantile regression model with substitution approaches

For ease of illustration, consider a longitudinal study where there are n independent subjects each of which is measured at T distinct time points. For example, explore an association of urinary samples measuring biomonitoring outcomes collected at T multiple time points from n independent workers and environmental air samples that are measured over time. In general, the number of time points is allowed to vary across subjects. Additionally, the ordering of time points occurring within subjects matters and within-subject correlation usually decreases over time. The lack of prior knowledge of this correlation would prevent one from drawing valid statistical inference, i.e. the higher the correlation incorrectly specified, the greater the biased standard error estimates of regression parameters produced (Diggle et al. 2002).

Marginal quantile regression shares similar theoretical properties with those firstly proposed by Koenker and Bassett (1978) for independent outcomes, but it additionally models within-subject correlations. Detailed description of the marginal quantile regression can be found in Supplementary Appendix I. In the regression model, let dependent variable, $\boldsymbol{y}_i = [y_{i1}, \ldots, y_{iT}]^T$, denotes the observed outcome vector for the ith subject without any assumptions of underlying distribution. To fill in an observed value not detected at the jth time for the ith subject, we propose to assign y_{ij} an imputed value using substitution approaches if $y_{ij} < \text{LOD}_{ij}$. The aims of the manuscript are to handle different degrees of left-censored outcome data in the presence of time-dependent covariates. We note that while substitution approaches are not ideal in conventional analyses that calculate arithmetic means, their use remains relevant in quantile modeling, where quantiles such as medians are being estimated. Specifically, regardless of distributional assumptions on the nondetected outcomes, changing the substitution value can significantly impact the estimated mean, but it will have a minimal effect, and often no effect, on the calculation of the median.

Single and multiple value imputation techniques are generally adopted for environmental exposure and biomonitoring samples that have concentration levels less than the LOD. Single value imputation assigning a value ranging from 0 to the LOD is the most common approach for displaying summary statistics. The use of LOD/$\sqrt{2}$ as the assigned value is demonstrated to provide a better estimation of the arithmetic mean and standard deviation than the use of LOD/2 for less-skewed data, while the LOD/2 should be considered when the data are highly skewed or have geometric standard deviation approximately 3 or greater (Hornung and Reed 1990; Burstyn and Teschke 1999). Another single value imputation is β-substitution approach (Ganser and Hewett 2010). The β-substitution method derives the calculation of a β factor to adjust each nondetected value based on the uncensored data. Through a simulation study, parameter estimates resulting from the β-substitution method have smaller biases and improved root MSEs relative to the LOD/2 and LOD/$\sqrt{2}$ substitution methods. See Appendix A of Ganser and Hewett (2010) for β-substitution method algorithm and Appendix B's example demonstration.

Multiple random value imputation provides a preferable alternative for the left-censored exposure data (Thi´ebaut and Jacqmin-Gadda 2004; Pleil 2016a, 2016b). Along with the use of a bootstrap procedure, i.e. randomly sampling with replacement, a maximum likelihood estimation is utilized to estimate parameters for a parametric distribution based on the uncensored data and censoring proportion. Therefore, an underlying distribution with the estimated parameters can be used for imputing values below the LOD

(Lubin et al. 2004). With this multiple value imputation, the imputed values can also be generated by regressing covariate(s) with environmental exposure and biomonitoring outcomes. Recently, multiple order value imputation was introduced to delineate the natural logarithm of the uncensored exposure concentration levels versus the *Z*-scores by fitting a linear equation presented in a QQ-plot (Pleil, 2016a, 2016b). The best fit equation is then used to project imputed values for the censored data onto the space spanned by the calculated *Z*-scores.

Types of time-dependent covariates and selection method

Four existing types of time-dependent covariates have been presented in the longitudinal data analysis for marginal models (Lai and Small 2007; Zhou et al. 2014) and implemented in quantile regression modeling (Chen and Westgate 2021) by modifying the inverse of a working correlation structure, $\boldsymbol{R}_i^{\gamma\,-1}$, to utilize every valid moment condition based on the specific type of time-dependent covariate. At a given quantile level of γ, the kth covariate is classified as a type I time-dependent covariate if the estimating equation's moment condition, $\mathrm{E}\left[X_{kis}v_i^{sj}\left(\gamma - I\left[y_{ij} \le X_{kij}\beta_k^{\gamma}\right]\right)\right] = 0$, holds for all s and j, where s and $j = 1,\ldots,T$, a type II if the moment is for $s > j$, a type III if the moment does not hold for some $s > j$, and a type IV representing the opposite of a type II if the moment for $s \le j$. Invalid moment conditions and efficiency loss for parameter estimation would occur if β_k^{γ} corresponds to a time-dependent covariate and the moment does not hold for some s and j. The interpretations of these 4 types of time-dependency are as follows.

- A common type I covariate is time itself or age.
- A type II covariate occurs when previous and current time points covariate values (independent variable) predict current and future time points exposure outcomes (dependent variable).
- A type III covariate occurs when previous and current time points covariate values predict current and future time points exposure outcomes and that current time points exposure outcomes predict future time points covariate values, creating a feedback cycle.
- A type IV covariate occurs when future time points covariate values are affected by current time points exposure outcomes.

Using the modified quantile regression modeling method requires that industrial hygienists or data analysts realize the type of time-dependent covariate, although this is often unknown in most scenarios. In addition, if covariate values and exposure outcomes are found to be associated, determining the specific type of time-dependency would be beneficial for interpreting the direction of impact between covariates and outcomes. Therefore, we account for an empirical MSE minimization criterion that results in the least variable estimation possible to select a working covariate type (Chen and Westgate 2019, 2021). The criterion is given by

$$\widehat{MSE}\left(\widehat{\boldsymbol{\beta}}_s^{\gamma}\right) = \widehat{Cov}\left(\widehat{\boldsymbol{\beta}}_s^{\gamma}\right) + \left(\widehat{\boldsymbol{\beta}}_s^{\gamma} - \widehat{\boldsymbol{\beta}}_{III}^{\gamma}\right)\left(\widehat{\boldsymbol{\beta}}_s^{\gamma} - \widehat{\boldsymbol{\beta}}_{III}^{\gamma}\right)^T,$$

where $\widehat{\boldsymbol{\beta}}_s^{\gamma}$ is the vector of regression parameter estimates in which the time-dependency is assumed to be type s, s = I, II, III, or IV, and $\widehat{Cov}\left(\widehat{\boldsymbol{\beta}}_s^{\gamma}\right)$ is an estimated covariance matrix of $\widehat{\boldsymbol{\beta}}_s^{\gamma}$. In this study, we did not take type IV into account, as its results can be comparable to those corresponding to a type II due to similar definitions and its cause-and-effect relationship rarely occur in practice. Only one up-to-date study targeting all valid moments to reach efficient parameter estimation has been proposed for marginal quantile regression in the presence of time-dependent covariates (Chen and Westgate 2021). However, there is no study discussing time-varying covariates in longitudinal exposure data with nondetects. Therefore, we further propose to extend the approaches used to improve estimation efficiency and to choose an unknown type of time-dependency for a marginal quantile model that substitution approaches are incorporated into.

Simulation study and results

In the simulation study, we compare the performances of our proposed quantile regression models to the quantile model incorporating an independence working correlation structure. We also considered 2 levels of proportions below the LOD, i.e. 20% and 40%, below the LOD, 3 quantile levels, i.e. 25th, 50th or median, and 75th, and 3 scenarios or types of time-dependent covariates (description of models can be found in Supplementary Appendix II). Under each regression model, we accommodated left-censored outcome data using 4 substitution approaches, i.e. β-substitution, LOD/2, and multiple random and order value imputations. The β-substitution method was also used as a referent model to compare with the other 3 substitution methods. Three modeling cases follow multivariate normal distribution, multivariate Student's *t*-distribution with 3 df, and multivariate log-normal distribution, respectively. The simulation process was demonstrated using a true first-order autoregressive (AR-1) working correlation structure and a correlation parameter $\alpha = 0.7$ (high correlation among time points from the same subject), as AR-1 may be favorable over other structures such as banded and exchangeable (EXCH), in a longitudinal study (Diggle et al. 2002). In any given setting, the number of subjects (n) is 200

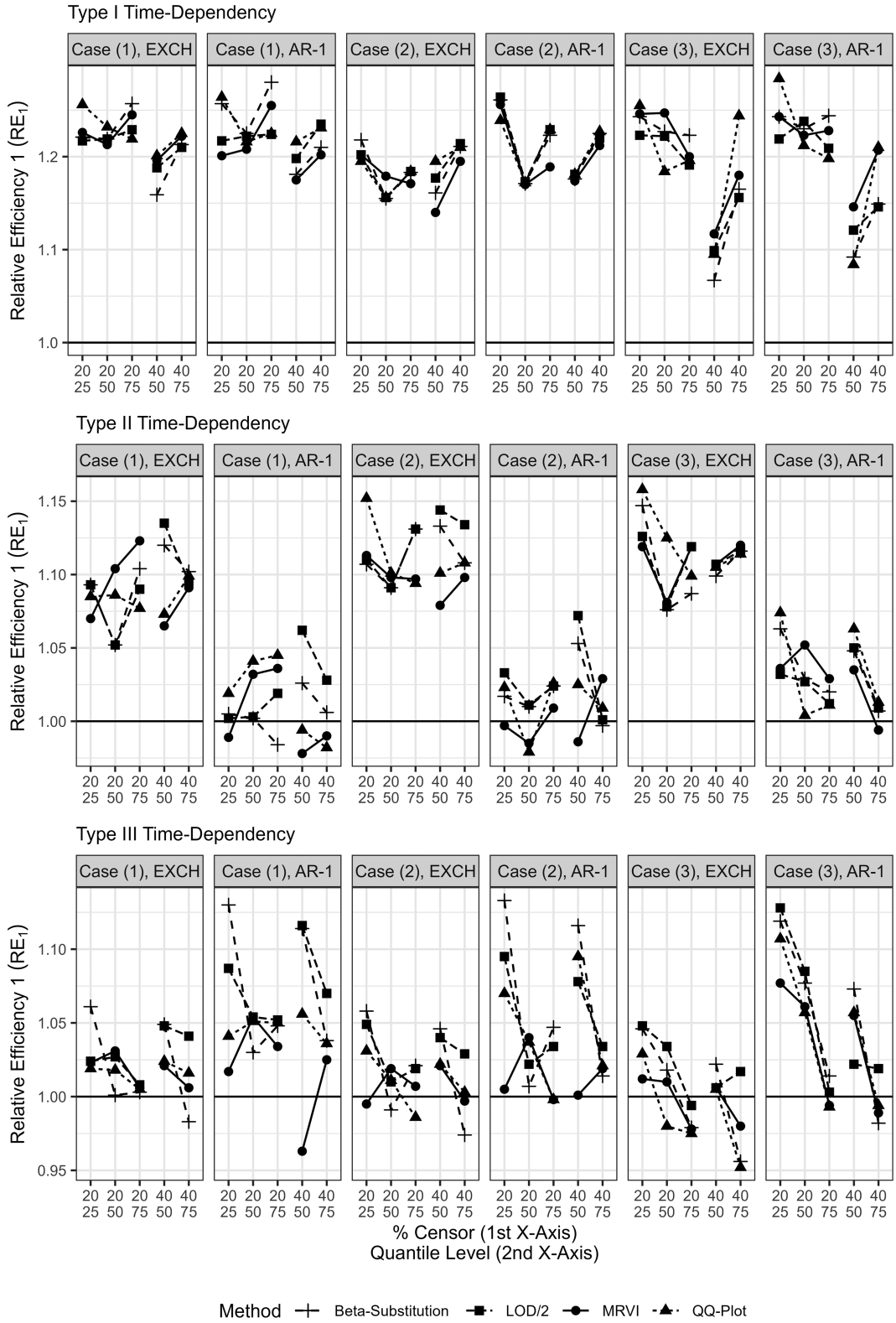

Fig. 1. Simulation results for the models with 4 substitution methods, β-substitution (beta-substitution), LOD/2 substitution, multiple random value imputation (MRVI), and multiple ordered value imputation (QQ-plot), under different underlying distributions (cases (1) to

and each subject has five repeated measurements (T) in a time manner. Each setting was conducted through 1000 simulations. The simulations were carried out using R version 4.2.3 (R Core Team 2023).

To examine differences in performances of regression parameter estimation, we calculated the MSEs allowing for the consideration of both the empirical bias and variability that that may arise from the estimated regression parameters. In Figs. 1 and 2 we displayed ratios of empirical MSEs, which we referred to as relative efficiencies (REs), to explore the performances of working correlation structures and substitution approaches, accordingly. For any given RE_1 and RE_2 in the figures, the denominator is the MSE resulting from the use of our quantile models. The numerator, as well as the referent value, of RE_1 is the MSE from the use of a quantile model without within-subject correlations, while the numerator of RE_2 is the MSE from the use of β-substitution method. The greater the RE, which is larger than one, i.e. the horizontal line in the figure, the smaller the MSE of the proposed quantile model, thus demonstrating its superiority over the referent model. In Supplementary Figure S1, we presented the number of times a working type of covariate was chosen out of 1000 simulations.

When comparing to the marginal model incorporating an independence working correlation structure, i.e. absence of within-subject correlation, results corresponding to either a true covariate type I, II, or III time dependency showed that, overall, the proposed methods using different substitution approaches to assign nondetected values are more efficient for any given censoring proportions and quantile levels (Fig. 1). Relative to an independence quantile model, the RE_1 results ranged from 1.07 to 1.28, 0.97 to 1.16, 0.95 to 1.13, respectively, over 3 types and resulted in greater efficiencies under consideration of the true type I or II. Additionally, the RE_2 results demonstrated that the multiple random value imputation outperformed the other approaches. In contrast, the β-substitution and multiple order value imputation exhibited suboptimal performance, as evidenced by their REs being less than one across settings with higher censoring proportion (40%) (Fig. 2). The substitution approaches chose most often the desired time-dependent type of covariate and the results with respect to selection frequencies were comparable regardless of the given approach (Supplementary Fig. S1). Note that empirical biases of regression parameter estimation were often negligible and MSEs of estimates were similar to the findings in the previous literature, and therefore the results were not shown. Also, the smaller the number of subjects (n), the greater the empirical bias and MSE that correspond to theoretical properties.

Carbon nanotube and nanofiber exposure study

We apply our methods to a study of worker-level determinants and standardized job tasks to predict future carbon nanotube and nanofiber (CNT/F) workplace exposures. The study was conducted in 15 unique facilities processing and/or manufacturing CNT/F across the United States using a longitudinal database collected by the National Institute for Occupational Safety and Health between 2012 and 2016 (Dahm et al. 2018, 2019). Previous cross-sectional studies were designed to examine the association of health effects and exposure to CNT/F, but lack the ability to infer causality (Liou et al. 2015). Through the selection of time-dependent type of covariate for latency periods of workplace factors and exposure determinants in a longitudinal study, the objective of our study was to determine if exposures to CNT/F from a previous period may influence future determinants or covariate values, which in turn may have an impact on future exposure outcomes.

A total of 120 participants from 15 facilities participated in the study (number of independent subjects or n is 120). Longitudinal measurements were collected over 2 full work-shifts, i.e. 2 samples, (number of repeated measurements or T) from 117 participants and 3 or 4 work-shifts from 3 participants, resulting in a total of 245 measurements. The exposure metrics to CNT/F were the mass of elemental carbon (EC) at both the respirable and inhalable aerosol size fractions, which are in the unit of $\mu g/m^3$. All values for EC exposures were background-corrected because of the interference from anthropogenic sources of EC. Detailed correction methods can be found in Dahm et al. (2018, 2019). Numbers of exposure measurements below the LOD were 44 and 16 for respirable and inhalable exposures, respectively. Also, measurements with negative background-corrected concentration levels were found in 70 respirable samples and in 76 inhalable samples. Adding up the measurements under these 2 scenarios, i.e. 114 and 92, the percentages of EC left-censored respirable and inhalable exposure levels needing to be assigned a value using substitution approaches were 46.5% and 37.6%, correspondingly. We considered 3

(3)) with incorrectly specified exchangeable (EXCH) or correctly specified AR-1 working correlation structure, comparing with the use of an independence working correlation structure. Relative efficiency (RE_1) compares the empirical mean squared error (MSE) from the use of an independence correlation to the empirical MSE from the use of a dependence correlation. The greater the RE_1, which is larger than one, i.e. the horizontal line, the smaller the MSE of the proposed quantile model, thus demonstrating its superiority over the referent model.

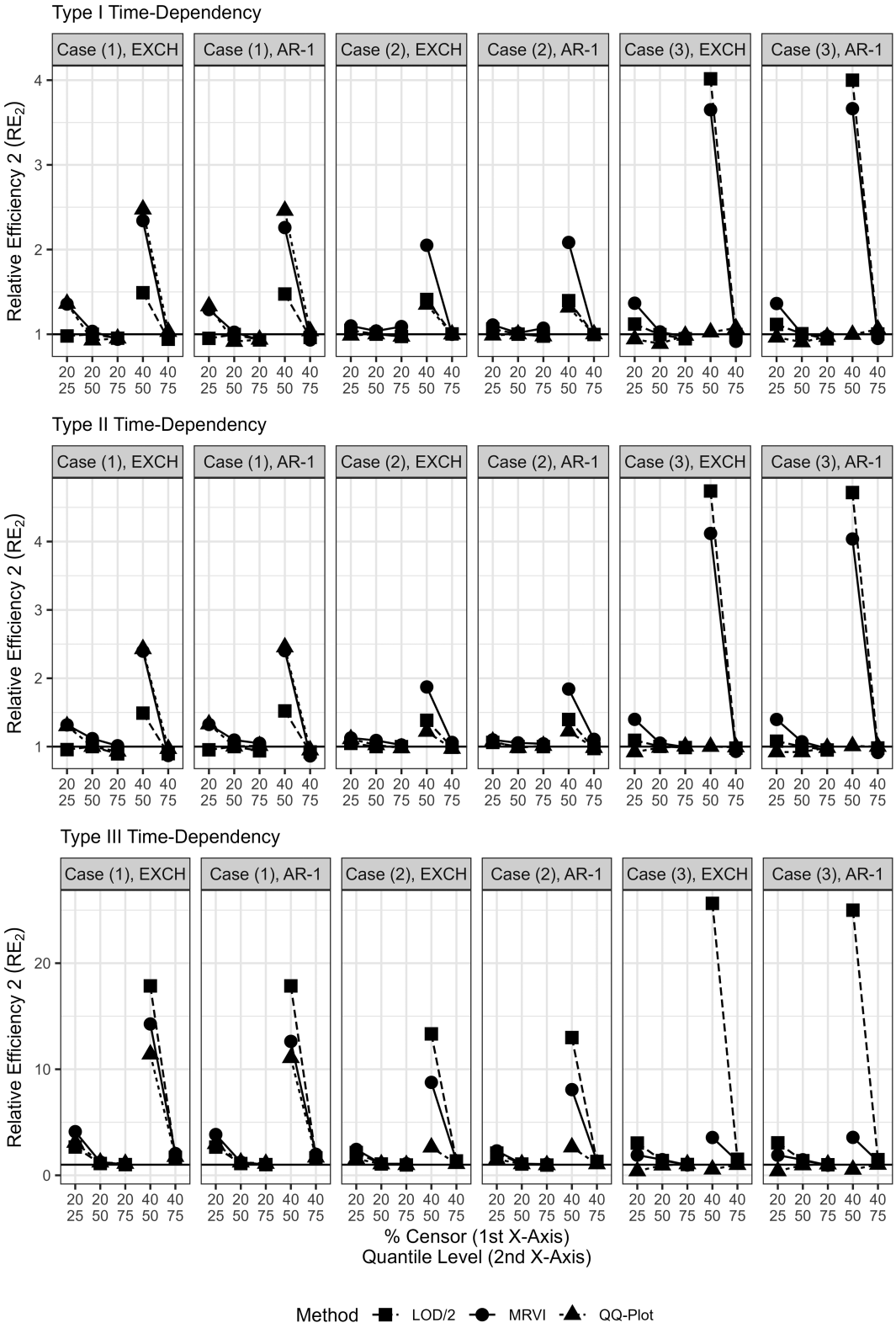

Fig. 2. Simulation results of relative efficiency (RE_2) that compares the empirical mean squared error (MSE) from the use of β-substitution (beta-substitution) to the empirical MSE from the use of the other substitution methods, including LOD/2 substitution,

Table 1. Frequency, percent, and descriptive statistics, including mean, SD, median, interquartile range (IQR), geometric mean (GM), and geometric standard deviation, for worker-level determinant, job tasks, and exposure outcomes.

Covariate					Mean (SD)
Worker-level determinant	Minutes of direct exposure to CNT/F				144.4 (156.6)
Job tasks	Minutes spent in office/desk work				206.8 (206.3)
	Minutes spent powder handling and postprocessing				29.2 (83.8)
Outcome[a]	**Type**	**% Censoring**	**Mean (SD)**	**Median (IQR)**	**GM (GSD)**
Respirable elemental carbon (EC) (μg/m^3)	Original	46.5	0.87 (5.55)	0.05 (0.02–0.38)	0.07 (8.93)
	Log-transformed		−2.70	−3.00	
Inhalable elemental carbon (EC) (μg/m^3)	Original	37.6	3.71 (18.2)	0.30 (0.02–1.22)	0.15 (25.2)
	Log-transformed		−1.93	−1.20	

[a]Outcome dataset was from Dahm et al. (2019) with the use of LOD/2 substitution approach.

determinants that were associated with decreased or increased exposure potentials based on the previous findings (Dahm et al. 2019), including time spent in a job with direct exposure to CNT/F, time spent performing office/desk work, and time spent performing powder handling/postprocessing of CNT/F powders, all measured in minutes.

Both distributions of EC respirable and inhalable exposure metrics for CNT/F were skewed to the right as evidenced by the arithmetic means being greater than the medians. Moreover, the GMs were not close to the medians, especially the respirable EC exposure, indicating the data might not be log-normally distributed (Table 1). The distribution of log-transformed respirable EC outcomes was asymmetric and left-skewed with greater median compared to mean. A QQ-plot examination also confirmed its skewness. Another motivation to apply a quantile regression model for these data is that the outcomes in both exposure metrics had potential outliers that could affect estimates of the means if a traditional mean regression is conducted (Supplementary Fig. S2).

We adopted a published model (Dahm et al. 2019) and followed its data manipulation by log-transforming the outcome data of the 2 EC exposures for ease of comparisons, but employed marginal quantile regression with an AR-1 working correlation structure at 3 quantile levels, γ = 0.5, 0.75, and 0.95. Results of 95th quantile were presented because potential outliers or influential points were found in the datasets (Supplementary Fig. S2). The chosen quantiles are based on a guideline (Chen et al. 2021) that any sample quantiles above the 47th and 38th quantiles can be reported when 46.5% and 37.6% of the EC respirable and inhalable exposure data were censored, respectively. Tables 2 and 3 provide the selected type of time dependency and the estimates of the regression parameter, SE, and 95% CI. The factor in Tables 2 and 3 represent the exponent of the regression parameter estimate, e.g. when 1-min change in direct exposure to CNT/F, the exponent of the estimate for the 50th quantile is interpreted as a change in μg/m^3 of the median of the exposure outcome. Note that the results of multiple random value imputation were included because it outperformed over the other substitution approaches in the simulation study. Also, we did not focus on type I because there was no time variable in the considered covariate list.

The estimated correlation parameters used to construct the AR-1 correlation structure ranged from 0.15 to 0.34, 0.23 to 0.31, and 0.48 to 0.54 for the EC respirable exposure data at the 50th, 75th, and 95th quantile levels, correspondingly, and from 0.62 to 0.65, 0.41 to 0.45, and 0.68 to 0.76 for 50th, 75th, and 95th quantiles of the EC inhalable exposure, indicating small to moderate within-subject correlations among longitudinal measurements collected from the participants. The 2 substitution approaches for each exposure outcome at the 50th, 75th, and 95th quantile levels yielded same directions and similar magnitudes for regression parameter estimates (Tables 2 and 3), which were also consistent with literature findings examining associations of exposure determinants and CNT/F (Kuijpers et al. 2016; Dahm *et al.* 2018, 2019). In addition to

multiple random value imputation (MRVI), and multiple ordered value imputation (QQ-plot), under different underlying distributions (cases (1) to (3)) with incorrectly specified exchangeable (EXCH) or correctly specified AR-1 working correlation structure. The greater the RE_2, which is larger than one, i.e. the horizontal line, the smaller the MSE of the proposed quantile model, thus demonstrating its superiority over the referent model.

Table 2. Type of time dependency, estimates of parameter and SE, 95% CI, and factor for covariate of interest from elemental carbon (EC) datasets.

Outcome	Covariate	Quantile[a]	Type	Estimate	SE	95% CI	Factor[b]
Logarithmic respirable EC	Direct exposure to CNT/F	50th		0.002 [c]	0.001	0.001 to 0.004	1.002
		75th	3	0.002	0.001	−0.001 to 0.005	1.002
		95th	3	0.006 [c]	0.002	0.003 to 0.009	1.006
	Office/desk work	50th		−0.002 [c]	0.001	−0.003 to −0.001	0.998
		75th	3	−0.003 [c]	0.001	−0.004 to −0.002	0.997
		95th	3	−0.003 [c]	0.001	−0.005 to −0.001	0.997
	Powder handling and postprocessing	50th		0.002	0.002	−0.001 to 0.005	1.002
		75th		0.008	0.004	−0.000 to 0.016	1.008
		95th		0.013 [c]	0.002	0.009 to 0.016	1.013

[a]Any quantile levels greater than or equal to 47th percentile can be calculated (censoring proportion is 46.5% for EC respirable exposure metric).
[b]Exponent of the estimate.
[c]*P*-value < 0.05.

Table 3. Type of time dependency, estimates of parameter and SE, 95% CI, and factor for covariate of interest from elemental carbon (EC) datasets.

Outcome	Covariate	Quantile[a]	Type	Estimate	SE	95% CI	Factor[b]
Logarithmic Inhalable EC	Direct exposure to CNT/F	50th	3	0.004[c]	0.001	0.003 to 0.005	1.004
		75th	2	0.002	0.001	−0.000 to 0.004	1.002
		95th	2	0.010[c]	0.003	0.005 to 0.015	1.010
	Office/desk work	50th	3	−0.004[c]	0.001	−0.006 to −0.003	0.996
		75th		−0.003[c]	0.001	−0.005 to −0.002	0.997
		95th	2	−0.005[c]	0.001	−0.007 to −0.004	0.995
	Powder handling and postprocessing	50th	2	0.005	0.003	−0.001 to 0.011	1.005
		75th	3	0.009[c]	0.004	0.001 to 0.017	1.009
		95th		0.009[c]	0.004	0.001 to 0.016	1.009

[a]Any quantile levels greater than or equal to 38th percentile can be calculated (censoring proportion is 37.6% for EC inhalable exposure metric).
[b]Exponent of the estimate.
[c]*P*-value < 0.05.

providing estimates of regression parameters for different quantiles (Supplementary Figs. S3 and S4), which is not possible through traditional regression approaches, we further quantified the associations of covariates and future exposure outcomes. Previous and current direct exposures to CNT/F might positively affect current and future respirable EC levels at the 75th and 95th quantiles, and inhalable EC levels at the median. Also, the current respirable and inhalable EC levels could positively predict the future duration of direct CNT/F exposure, creating a type III time-dependency. Additionally, previous and current direct exposures to CNT/F significantly and positively predicted current inhalable EC levels at the 75th and 95th quantile levels, suggesting a type II time dependency. Exposures from the participants who performed office/desk work created a feedback cycle with a negative impact to and from respirable EC levels at the 75th and 95th quantiles and median inhalable EC levels, and removed the feedback cycle to predict inhalable EC levels significantly and negatively at the 95th quantile. The feedback cycle with a negative impact indicates that the less exposed a participant is currently, the more time the participant will spend in the office in the future. However, the covariate could be treated as a type II time dependency because the difference between types II and III was negligible. Although no time-dependency was investigated between job task covariate of powder handling and postprocessing and respirable EC, we found that, as the quantile level increased, the participants who performed this job had increased magnitude of exposure. This pattern was also found between

powder handling and postprocessing and inhalable EC from the median to the 75th quantile. Moreover, type II and III time-dependencies were explored for the powder handling and postprocessing job task that was used to predict inhalable EC levels at the 50th and 95th quantiles, correspondingly. Note that based on the prospective from an industrial hygienist with experience in CNT/F exposure assessments, the cycle from current exposure outcome to future covariate value in a type III time-dependency can be ruled out because the times spent performing the corresponding job tasks were controllable to partially controllable, and therefore, statistical significance does not imply practical significance. The corresponding R code and functions for implementing the proposed methods were provided in Supplementary Appendices III and IV.

Discussion

Mean regression analyses have gained widespread attention in environmental exposure and biomonitoring studies for analyzing longitudinal right-skewed exposure outcomes with nondetects. However, in certain real-world datasets, the use of mean regression models can be influenced by skewness and the presence of potential outliers, which could impact the mean more than the median. In such scenarios, employing quantile regression without requiring a specified error distribution for modeling the conditional quantiles of the response variable is recommended. Marginal quantile regression targeting the 50th quantile or median also presents itself as an alternative for analyzing log-normal exposure outcomes, particularly for high censoring proportion and substantial within-subject correlation (Chen et al. 2021). This arises from the equivalence between modeling the conditional mean of log-normal exposure data and modeling the conditional median of log-transformed data. To address the gap of nondetects, we proposed incorporating available fill-in or substitution methods for utilizing nondetects below LOD. Furthermore, in a longitudinal study, the values of covariates may vary over time. Incorrectly treating the time-dependency as time-independency will lead to an efficiency loss of regression parameter estimation. To improve parameter estimation, we proposed utilizing a criterion that results in the least variable estimation possible to select a type of time-dependent covariate. Through a simulation study, we demonstrated that our methods outperformed previous approaches considering an independence working correlation structure in the absence of within-subject correlation for time-dependent covariates and had consistent results with the existing literature using the selection criterion.

We limited our consideration to an available but less parsimonious AR-1 working correlation structure for marginal quantile regression models because it is often preferred over other structures in longitudinal studies (Diggle *et al.* 2002). The use of an AR-1 correlation structure in marginal models is an additional advantage, as it may not be accommodated in existing random effects models used for left-censored repeated measures data because the covariance matrix of the random error term at the subject level cannot be determined (Jin et al. 2011). In future studies, a Gaussian pseudolikelihood selection technique (Fu et al. 2015) or a general stationary autocorrelation structure (Lu and Fan 2015), rather than a parametric likelihood, may be accounted for to avoid the need to specify any correlation structures.

In terms of choices of substitution approaches, direct truncation is not recommended for the left-censored values, as it has the potential to distort measures of central tendency and dispersion, ultimately leading to decreased accuracy and precision. The replacement of values below the LOD with zeros is also not a recommended approach, particularly when dealing with data distributions that are log-transformed, as the logarithmic zero is undefined. Note that the quantile models also allow for multiple LODs to be present in exposure data with longitudinal measures, as the censoring proportion can be calculated accordingly.

The simulation study presented in this manuscript utilized marginal quantile regression models with balanced repeated measurements over time, and univariable results were reported. However, the proposed methods can be extended to accommodate subjects with varying numbers of time points, as demonstrated in the application example. It should be noted that while there is the option to choose multiple types of time-dependent covariates at any given quantile level, we assumed the presence of only one time-dependent covariate of unknown type for simplicity. We also explored the use of marginal quantile models with multiple types of time-dependent covariates, but the results were comparable to those presented and are therefore not included. In addition, statistical significance does not necessarily mean practical significance. We suggested that researchers should consult with industrial hygienists to pursue a practical perspective before interpreting specific types of time-dependent covariate or causality, i.e. overrule a statistical perspective or rule out a feedback cycle.

Our study has several limitations that should be acknowledged. First, the simulations were performed using parametric distributions, and thus, it may be necessary to evaluate other departures from normality and log-normality, such as data that are skewed to the left or follow a chi-squared or inverse gamma distribution. The simulation datasets also lacked gold standard referent estimation for every quantile

level, which could potentially limit the conclusiveness of comparisons. We recommend that industrial hygienists or data analysts utilize graphical and testing examinations to ensure that distributional assumptions of mean regression are met. If the sampled data significantly deviates from its underlying distribution, no method may produce unbiased estimates. In such situations, quantile regression may be considered as a safe approach, albeit with a potential loss of efficiency, such as wide confidence intervals and large *P*-values. Second, the substitution approaches are commonly used to calculate summary statistics for real-world exposure data with left censoring. Nevertheless, regardless of the specific approach selected, all imputed values are estimates rather than real data, subjecting to the LOD of laboratory instruments and containing unobserved errors. Moreover, multiple value and order imputation techniques require a common parametric distribution with parameters estimated from the uncensored data to impute values for observations below the LOD. These techniques may produce biased or unstable statistics when data are asymmetric after log-transformation. LOD/2 and β-substitution methods are less complex to implement and calculate, although they pose difficulties for performing standard normality tests. LOD/2 approach assumes that all values below the LOD can be represented as a single point mass at half the LOD. The assumption is an oversimplification and fails to accurately reflect the actual variability and distribution of nondetects. Also, all substitution approaches rely on distribution assumptions, so the proposed marginal quantile model technically no longer avoids any distributional assumptions. Lastly, assuming the independence of all imputed values for measurements below the LOD may lead to positively biased SE estimates, which translates into incorrect estimates of the sampling variability.

Future work could consider maximum likelihood-based covariance matrix estimators that enable the presence of the left-censored values without imputing them (Pesonen et al. 2015), although the imputation technique would still be constrained by the underlying distributional assumptions. Also, Bayesian multiple imputation methods have been shown to improve their performances with the use of more informative priors (Huynh et al. 2016) and have weaker distributional assumptions. The distribution-based Bayesian methods for left-censored bivariate and multivariate longitudinal data (Chen et al. 2011, 2013) are worthy to apply for exposure data with time-dependent covariates. However, the use of Bayesian methods highly depends on the prior information. Parameter estimation efficiency can further be enhanced using composite marginal quantile regression in which multiple quantiles share common characteristics in the presence of time-varying covariates (Yang et al. 2017). A simultaneous approach to selecting a working correlation structure and determining the type of time-dependency in covariates could also be developed. It should be noted that our methods demonstrated superior powers in the presence of type I or II time-dependency, but coverage probabilities were sub-nominal in some settings. Further research is necessary to ensure proper coverage probabilities in all scenarios, although coverage probability is not the primary focus of inference in this manuscript.

Conclusions

Quantile regression modeling offers an alternative perspective on the conditional distribution of a longitudinal outcome above the LOD, enables a more comprehensive examination of variable of interest by evaluating various quantiles, and is robust to potential outliers. The distributional assumptions are not always be specified in practice. Additionally, in a longitudinal study, the values of covariate may vary over time. Falling to account for time-dependency of the covariate might result in inefficient regression parameter estimation. Therefore, we applied a criterion to select different defined types of time-dependent covariates. After replacing nondetectable values using different substitution approaches, the simulation study suggested that multiple random value imputation is appropriate for longitudinal data with left censoring.

Acknowledgements

We would like to thank the people from the Division of Field Studies and Engineering and Health Effects Laboratory Division at CDC's National Institute for Occupational Safety and Health who assisted in the data collection for this study.

Funding

This work was not supported by any funding.

Conflict of interest

The authors declare no conflict of interest.

Disclaimer

The findings and conclusions in this manuscript are those of the author(s) and do not necessarily represent the official position of the National Institute for

Occupational Safety and Health, Centers for Disease Control and Prevention.

Data availability

The simulation and application data can be shared on request to the corresponding author.

Supplementary material

Supplementary material is available at *Annals of Work Exposures and Health* online.

References

Burstyn I, Teschke K. Studying the determinants of exposure: a review of methods. Am Ind Hyg Assoc J. 1999:60(1):57–72. https://doi.org/10.1080/00028899908984423

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

Chen H, Quandt SA, Grzywacz JG, Arcury TA. A Bayesian multiple imputation method for handling longitudinal pesticide data with values below the limit of detection. Environmetrics. 2013:24(2):132–142. https://doi.org/10.1002/env.2193

Chen IC, Bertke SJ, Curwin BD. Quantile regression for exposure data with repeated measures in the presence of non-detects. J Expo Sci Environ Epidemiol. 2021:31(6):1057–1066. https://doi.org/10.1038/s41370-021-00345-1

Chen IC, Westgate PM. Improved methods for the marginal analysis of longitudinal data in the presence of time-dependent covariates. Stat Med. 2017:36(16):2533–2546. https://doi.org/10.1002/sim.7307

Chen IC, Westgate PM. A novel approach to selecting classification types for time-dependent covariates in the marginal analysis of longitudinal data. Stat Methods Med Res. 2019:28(10-11):3176–3186. https://doi.org/10.1177/0962280218799529

Chen IC, Westgate PM. Marginal quantile regression for longitudinal data analysis in the presence of time-dependent covariates. The Int J Biostat. 2021:17(2):267–282. https://doi.org/10.1515/ijb-2020-0010

Dahm MM, Bertke SJ, Schubauer-Berigan MK. Predicting occupational exposures to carbon nanotubes and nanofibers based on workplace determinants modeling. Ann. Work Expo. Health. 2019:63(2):158–172. https://doi.org/10.1093/annweh/wxy102

Dahm MM, Schubauer-Berigan MK, Evans DE, Birch ME, Bertke SJ, Beard JD, Erdely A, Fernback JE, Mercer RR, Grinshpun SA. Exposure assessments for a cross-sectional epidemiologic study of US carbon nanotube and nanofiber workers. Int J Hyg Environ Health. 2018:221(3):429–440. https://doi.org/10.1016/j.ijheh.2018.01.006

Diggle PJ, Heagerty PJ, Liang KY, Zeger SL. 2002. The analysis of longitudinal data. 2nd ed. New York (NY): Oxford University Press; 2002.

Fitzmaurice GMA. A caveat concerning independence estimating equations with multiple multivariate binary data. Biometrics. 1995:51(1):309–317. https://doi.org/10.2307/2533336

Fu L, Wang YG, Zhu M. A Gaussian pseudolikelihood approach for quantile regression with repeated measurements. Comput Stat Data Anal. 2015:84:41–53. https://doi.org/10.1016/j.csda.2014.11.002

Ganser GH, Hewett P. An accurate substitution method for analyzing censored data. J Occup Environ Hyg. 2010:7(4):233–244. https://doi.org/10.1080/15459621003609713

Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere. 2006:65(11):2434–2439. https://doi.org/10.1016/j.chemosphere.2006.04.051

Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg. 1990:5(1):46–51. https://doi.org/10.1080/1047322x.1990.10389587

Huynh T, Quick H, Ramachandran G, Banerjee S, Stenzel M, Sandler DP, Engel LS, Kwok RK, Blair A, Stewart PA. A comparison of the β-substitution method and a Bayesian method for analyzing left-censored data. Ann Occup Hyg. 2016:60(1):56–73. https://doi.org/10.1093/annhyg/mev049

Huynh T, Ramachandran G, Banerjee S, Monteiro J, Stenzel M, Sandler DP, Engel LS, Kwok RK, Blair A, Stewart PA. Comparison of methods for analyzing left-censored occupational exposure data. Ann Occup Hyg. 2014:58(9):1126–1142. https://doi.org/10.1093/annhyg/meu067

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(1):97–112. https://doi.org/10.1093/annhyg/meq061

Koenker R, Bassett G. Regression quantiles. Econometrica. 1978:46(1):33–50. https://doi.org/10.2307/1913643

Kuijpers E, Bekker C, Fransman W, Brouwer D, Tromp P, Vlaanderen J, Godderis L, Hoet P, Lan Q, Silverman D et al. Occupational exposure to multi-walled carbon nanotubes during commercial production synthesis and handling. Ann Occup Hyg. 2016:60(3):305–317. https://doi.org/10.1093/annhyg/mev082

Lai TL, Small DS. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. J Roy Stat Soc Ser B: Stat Methodol 2007:69(1):79–99. https://doi.org/10.1111/j.1467-9868.2007.00578.x

Lalonde TL, Wilson JR, Yin J. GMM logistic regression models for longitudinal data with time-dependent covariates and extended classifications. Stat Med. 2014:33(27):4756–4769. https://doi.org/10.1002/sim.6273

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.

Leung DHY, Small DS, Qin J, Zhu M. Shrinkage empirical likelihood estimator in longitudinal analysis with time-dependent covariates–application to modeling the health of Filipino children. Biometrics. 2013:69(3):624–632. https://doi.org/10.1111/biom.12039

Liou SH, Tsai CSJ, Pelclova D, Schubauer-Berigan MK, Schulte PA. Assessing the first wave of epidemiological studies of nanomaterial workers. J Nanopart Res. 2015:17(10):413. https://doi.org/10.1007/s11051-015-3219-7

Lu X, Fan Z. Weighted quantile regression for longitudinal data. Comput Stat. 2015:30(2):569–592. https://doi.org/10.1007/s00180-014-0550-x

Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004:112(17):1691–1696. https://doi.org/10.1289/ehp.7199

Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Commun Stat–Simul Comput. 1994:23(4):939–951. https://doi.org/10.1080/03610919408813210

Pesonen M, Pesonen H, Nevalainen J. Covariance matrix estimation for left-censored data. Comput Stat Data Anal 2015:92:13–25. https://doi.org/10.1016/j.csda.2015.06.005

Pleil JD. QQ-plots for assessing distributions of biomarker measurements and generating defensible summary statistics. J Breath Res. 2016a:10(3):035001. https://doi.org/10.1088/1752-7155/10/3/035001

Pleil JD. Imputing defensible values for left-censored 'below level of quantitation' (LoQ) biomarker measurements. J Breath Res. 2016b:10(4):045001. https://doi.org/10.1088/1752-7155/10/4/045001

R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2023. URL https://www.R-project.org/

Tang CY, Leng C. Empirical likelihood and quantile regression in longitudinal data analysis. Biometrika. 2011:98(4):1001–1006. https://doi.org/10.1093/biomet/asr050

Thiébaut R, Jacqmin-Gadda H. Mixed models for longitudinal left-censored repeated measures. Comput Methods Programs Biomed. 2004:74(3):255–260. https://doi.org/10.1016/j.cmpb.2003.08.004

Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika. 2003:90(1):29–41. https://doi.org/10.1093/biomet/90.1.29

Yang CC, Chen YH, Chang HY. Composite marginal quantile regression analysis for longitudinal adolescent body mass index data. Stat Med. 2017:36(21):3380–3397. https://doi.org/10.1002/sim.7355

Zhou Y, Lefante J, Rice J, Chen S. Using modified approaches on marginal regression analysis of longitudinal data with time-dependent covariates. Stat Med. 2014:33(19):3354–3364. https://doi.org/10.1002/sim.6171