Journal of Exposure Science & Environmental Epidemiology (2021) 31:1047–1056
https://doi.org/10.1038/s41370-021-00331-7

ARTICLE

Simultaneous modeling of detection rate and exposure concentration using semi-continuous models to identify exposure determinants when left-censored data may be a true zero

Melissa C. Friesen[1] · Hyoyoung Choo-Wosoba[2] · Philippe Sarazin[3,4] · Jooyeon Hwang[5] · Pamela Dopart[1] · Daniel E. Russ[6] · Nicole C. Deziel[7] · Jérôme Lavoué[4] · Paul S. Albert[2] · Bin Zhu[2]

Received: 5 November 2020 / Revised: 7 April 2021 / Accepted: 15 April 2021 / Published online: 18 May 2021

Abstract

Background Most methods for treating left-censored data assume the analyte is present but not quantified. Biased estimates may result if the analyte is absent such that the unobserved data represents a mixed exposure distribution with an unknown proportion clustered at zero.

Objective We used semi-continuous models to identify time and industry trends in 52,457 OSHA inspection lead sample results.

Method The first component of the semi-continuous model predicted the probability of detecting concentrations ≥ 0.007 mg/m^3 (highest estimated detection limit, 62% of measurements). The second component predicted the median concentration of measurements ≥ 0.007 mg/m^3. Both components included a random-effect for industry and fixed-effects for year, industry group, analytical method, and other variables. We used the two components together to predict median industry- and time-specific lead concentrations.

Results The probabilities of detectable concentrations and the median detected concentrations decreased with year; both were also lower for measurements analyzed for multiple (vs. one) metals and for those analyzed by inductively-coupled plasma (vs. atomic absorption spectroscopy). The covariance was 0.30 (standard error = 0.06), confirming the two components were correlated.

Significance We identified determinants of exposure in data with over 60% left-censored, while accounting for correlated relationships and without assuming a distribution for the censored data.

Keywords statistical modeling · left-censored data · occupational lead exposure

Introduction

Environmental exposure monitoring data commonly includes measurements that cannot be quantified by the laboratory for the target analyte. Referred to as

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41370-021-00331-7.

✉ Melissa C. Friesen
friesenmc@mail.nih.gov

[1] National Cancer Institute, Division of Cancer Epidemiology & Genetics, Occupational and Environmental Epidemiology Branch, Rockville, MD, USA

[2] National Cancer Institute, Division of Cancer Epidemiology & Genetics, Biostatistics Branch, Rockville, MD, USA

[3] Institut de recherche Robert-Sauvé en santé et en sécurité du travail, Chemical and Biological Hazards Prevention, Montréal, QC, Canada

[4] Department of Occupational and Environmental Health, Université de Montréal, Montréal, QC, Canada

[5] Department of Occupational and Environmental Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA

[6] Office of Intramural Research, Center for Information Technology, National Institutes of Health, Bethesda, MD, USA

[7] Yale School of Public Health, Yale University, New Haven, CT, USA

"undetected," "below limit of detection (or quantification)," or "left-censored" samples, unquantified measurements can represent a substantial portion of the measurements. Methods to handle these measurements to minimize bias have continued to evolve, with best practices employing imputation or substitution methods that are based on the observed distribution of detected samples [1–10]. These methods generally assume that the target analyte in the left-censored data is present but not quantifiable by the laboratory method and make unverifiable assumptions about the distribution of the marker below the detection limit. This is a reasonable assumption in many analyses of well-defined exposure scenarios focused on a single analyte. However, many analytes, such as individual metals and solvents, are now commonly measured as part of an analyte panel. On these panels, it is plausible that one or more of the measured analytes may be truly absent (true zero), rather than present but not detected, and were measured because of concern regarding a different analyte on the panel. This results in an unobserved bimodal distribution that includes a cluster at or near zero and a log-normal distribution of measured and undetected, but present concentrations. Moreover, we often do not have information available to a priori determine whether the analyte in the left-censored sample is "absent" or "present but not detected" [11, 12]. A mixture model to address this issue has been applied to biomarker data [13]; however, this model does not account for correlated measurements in longitudinally collected data. As a result, statistical approaches are needed to account for a dependent repeated measure with a bimodal distribution that includes an unobserved cluster at or near zero to avoid biased estimates.

Our motivation was to evaluate predictors of exposure of environmental air measurements of lead in U.S. workplaces to support reconstruction of historical lead exposures for epidemiologic studies. The lead data were obtained from the online Chemical Exposure Hazard Database (CEHD) of inspection measurements reported by the U.S. Occupational Safety and Health (OSHA) Salt Lake City Laboratory. The lead measurements were collected over more than 20 years (1984–2009), from workplaces representing 575 industry groups, and were analyzed by a single laboratory using multiple analytical methods that included targeted analysis of lead and lead as part of a metal panel. Previous analyses of these data found that 58% of the measurements were unquantified by the laboratory and reported as 0 [11, 12]. Moreover, the unquantified proportion varied by whether the CEHD measurement was also recorded in the OSHA Integrated Management Information System (IMIS) database, with 74% and 50% of the CEHD measurements unquantified for the subsets that were recorded and not recorded in IMIS, respectively, suggesting selective reporting. However, both subsets had very similar cumulative distribution functions for the quantified measurements. As a result, we attributed the varying proportions below the LOD, in part, to measurements analyzed for lead as part of a metal panel analysis but where lead was not the analyte of interest during the inspection (i.e., true zeros), and to potential selective underreporting of measurements below the LOD in the IMIS database. Using back-of-the-envelope calculations, we estimate an approximately five-fold relative difference in the overall geometric mean between treating all measurements less than the LOD as present vs. absent (not shown). Given these concerns, traditional methods that assumed a single exposure distribution and that the analyte was present, but unquantified, would expect to result in biased estimates.

Semi-continuous models have been designed to handle analysis of a data set that has both repeated measures and an outcome (dependent) variable with a bimodal distribution with a cluster at zero that may either be observed or unobserved [14]. The approach splits the analysis to simultaneously model both the occurrence and intensity of the outcome variable while accounting for correlation in repeated measures. In a semi-continuous model, the first component, the occurrence model, is a generalized linear mixed model (logistic link) that evaluates predictors of the probability of a measurement being non-zero or being above a threshold value (e.g., a detection limit). The second component, the intensity model, is a linear mixed model that evaluates predictors of the detected analyte concentrations. Both components can incorporate a single random effect. The semi-continuous model is flexible in that (i) it makes no explicit assumption about the distribution below the threshold value (e.g., no assumptions are made about the proportion of "true zeros" or results that are "present but not quantified") and (ii) it allows for the same or different fixed effect predictors in the two model components and correlated random effects between the two components. The approach simultaneously evaluates the two models while accounting for their lack of independence. Modeling the components independently can induce bias when the random effects are correlated but mis-specified as independent [15]. This model structure have been used in an increasing variety of applications, including evaluations of food and activity diaries [16, 17], driving behaviors [18], biomarkers [19, 20], and treatment effectiveness [21]. We are unaware of any prior application of semi-continuous models to environmental or occupational monitoring data.

In the present analysis, we demonstrate the methodology of applying a semi-continuous mixed-effects model, using as a case study the OSHA CEHD lead measurements (as the dependent variable) to identify predictors of workplace lead exposures over more than two decades. We then used the models to predict industry- and time-specific lead concentrations to support future exposure assessment efforts.

Resolving this statistical challenge is a crucial first step in using the CEHD data for exposure assessment and other purposes. Similar analytical approaches can be applied to other environmental measurement data with mixed distributions and large proportions of non-detects.

Methods

Data source and data treatment

The lead measurements for the years 1984 to 2009 were obtained from the Chemical Exposure Health Data (CEHD) online data base that reports the laboratory analytical results from OSHA's central laboratory in Salt Lake City [11, 12]. The time period was chosen to be consistent with our prior publication; additional measurement years are now available online. The variables available in CEHD are described by Lavoue et al. [11] and include inspection number, four-digit Standard Industry Classification (SIC4) code from the 1972 (pre-1987) or 1987 version (1987 onwards), sampling date, lead concentration in mg/m^3, sample type (e.g., personal, area, bulk), instrument type, sample duration, field number, and other laboratory variables. The lead concentrations for measurements below the laboratory limit of quantification were reported in the database as "0". We restricted our analyses to personal samples and excluded all analytical blanks.

We undertook several cleaning and data treatment steps. Initial cleaning of this data was undertaken to exclude erroneous records, as described in Lavoue et al. [11]. In addition, because the measurements with the same sampling number would correspond to partial work shift measurements, we used the procedure described in Sarazin et al. [22] to aggregate 73,115 lead analysis records to obtain a single total sampling time and time weighted average concentration for each unique sampling number, resulting in a data set of 52,467 sample results. We recoded 1972 SIC4 assignments to the 1987 classification system. From SIC4, we also derived a broad industry group from the first digit (SIC1). We recoded all instrument types into two broad categories: atomic absorption spectroscopy (AAS) and inductively coupled plasma (ICP). We categorized total sampling duration into four categories: 0- < 60, 60- < 120, 120- < 600, and ≥ 600 min. We derived a variable indicating whether the result belonged to a panel, where panel = "yes" if there was more than one chemical analyte with the same field number. The database includes measurements that have been collected by inspectors under both federal OSHA plans and state OSHA plans. We categorized state into type of OSHA plan: federal OSHA, state OSHA, or partial federal OSHA (mix of federal and state components).

Statistical analysis

We applied a semi-continuous model that incorporated an occurrence model for R_{ij} and an intensity model for Y_{ij} when $R_{ij} = 1$. In this formulation, R_{ij} denotes whether the lead result Y_{ij} for the ith subject and jth time point is at or above a threshold value c (here 0.007 mg/m^3):

$$R_{ij} = \begin{cases} 0, \ if\ Y_{ij} < c \\ 1, \ if\ Y_{ij} \geq c \end{cases},$$

R_{ij} is assumed to follow a Bernoulli distribution with probability of R_{ij}, P_{ij}. Because the detection limit varied over the measurement period and by analytical method, we set c to a single detection limit of 0.007 mg/m^3, the highest observed quantification limit, and assigned a zero to any result below 0.007 mg/m^3. Quantification limits varied from 0.002 to 0.007 mg/m^3.

We constructed the occurrence model to evaluate the probability of a sample result being at or above the threshold using a logit link function: $logit(p_{ij}) = X_{1ij}^T\beta_1 + u_{1i}$. We constructed the intensity model, using only the detected measurements, where the detected observations were assumed to be lognormally distributed: $\log(E(Y_{ij}|R_{ij} = 1)) \sim N(X_{2ij}^T\beta_2 + u_{2i}, \sigma_e^2)$. Both models incorporated a random variable, u_{1i} and u_{2i}, respectively, denoting the SIC4 code to account for clustering of observations within industries and to allow us to use the best linear unbiased predictors for each SIC4 to estimate industry-specific estimates [23]. With 575 SIC4 industries in the data, most with sparse data, incorporating SIC4 as a random-effect allowed us to borrow strength from the mean of the broad industry group (SIC1, incorporated as fixed-effect) when estimating the SIC4-specific mean. u_{1i} and u_{2i} were assumed to be normally distributed with mean vector of zeros and covariance matrix. β_1 and β_2 were fixed effects; X_{1ij}^T and X_{2ij}^T were the fixed effect design matrices of both the occurrence and intensity models, respectively, designating the model covariates. We used the same covariates for both models, $X_{1ij}^T = X_{2ij}^T$; however, the framework allows for the covariates to differ between models. These covariates were calendar year, sample duration, analytical method, SIC1, OSHA plan type, and panel sample (yes = > 1 analyte measured; 0 = only lead measured). We also evaluated calendar year interactions with analytical method, OSHA plan type, and SIC1. The model was built using *MIXCORR*, a SAS macro developed by Tooze et al. [14], which is provided in Supplementary Appendix A.

For each SIC4 code, we used the occurrence model to estimate the probability of a result being ≥ 0.007 mg/m^3 and the intensity model to estimate the mean log lead concentration for those results that were ≥0.007 mg/m^3. We then combined the two models to estimate the median lead concentration for each SIC4 code using distributional parameters. This allowed us to make no assumptions of whether the undetected sample result was a “true zero” or “"present but unquantified”. For example, if the occurrence model estimated that the probability of being above the detection limit was 70%, we then used the 20th percentile from the intensity model to obtain the median concentration. For industries with a predicted probability of detection above the threshold of <50% we could only state that the predicted industry-specific median was below that threshold. The median based on the mixed-distribution model-based estimates was obtained by:

$$med(Y_{ij}|\beta_1, \beta_2, u_{1i}, u_{2i}, \sigma_e^2, c) = \begin{cases} below\, c, & if\, p_{ij} < 0.5 \\ exp\left[\Phi_{Z_{ij}}^{-1}\left(0.5 - \frac{1}{1+exp(X_{1ij}^T\beta_1+u_{1i})}\right)\right], & otherwise \end{cases}$$

where $\Phi_{Z_{ij}}^{-1}$ is the inverse of the standard normal cdf as a quantile function, and $Z_{ij} = \frac{Y_{ij}-X_{2ij}^T\beta_2-u_{2i}}{\sigma_e}$.

We also computed the median on only the values above the LOD as $exp(X_{2ij}^T\beta_2 + u_{2i})$. Estimates were obtained by replacing parameters with their estimates in the above formulas. Variances associated with these median estimates were estimated using the bootstrap to obtain 95% confidence intervals (95%CI). For these predictions, we set the following variables as the reference categories: year = 1984, analytical method = AAS, sample duration 120–600 min, OSHA plan = federal, and number of analytes on a panel = 1. These predictions were performed in R and the code is provided in supplemental Appendix B. The code can be modified to obtain other predicted percentiles, such as the 75th or 95th percentile.

Results

Measurement descriptives

The lead data set comprised 52,467 sample results, with 32,601 (62%) samples reported with concentrations < 0.007 mg/m^3. The samples below the 0.007 mg/m^3 threshold were more frequently analyzed using ICP (76%) than AAS (34%). Of these, 889 ICP and 1015 AAS samples had a non-zero lead concentration in CEHD (3.6% of all samples) but were treated as “zero” in the statistical models.

The data set included samples from 575 SIC4 codes, with a median of 14 (interquartile range, IQR: 3–61) and maximum of 2485 samples per SIC4 (Supplemental Table 1). The median proportion of samples ≥ 0.007 mg/m^3 was 11% (IQR 0–35%), with 203 (35%) SIC4 codes with no measurements ≥ 0.007 mg/m^3. The proportion of samples analyzed using ICP varied by SIC4, with a median of 75% analyzed using ICP (IQR 43–100%). The 20 SIC4 codes with the most samples had a minimum of 626 samples, with 8 to 93% of the samples ≥0.007 mg/m^3 and 15 to 95% analyzed by ICP (Supplemental Table 1). We observed moderate positive correlation between the number of samples and the proportion of samples ≥ 0.007 mg/m^3 (Spearman correlation coefficient, rho = 0.44) and slightly negative correlation between the proportion of samples ≥ 0.007 mg/m^3 and the proportion of samples analyzed using ICP (rho = −0.28) (not shown).

Model parameters

In the occurrence model (Table 1), we found that the probability of detectable results decreased with calendar year, samples analyzed using ICP vs. AAS, all duration categories compared to 120–600 min, and samples from states with partial federal OSHA plans vs. state only or federal only. For samples analyzed by AAS, the probability of detection increased for measurements that were analyzed for multiple metals. In addition, broad industry differences were observed, with lower probability of detected results (compared to SIC1 = 3, Manufacturing industry codes 300–399) for SIC1 = 2 (Manufacturing industry codes 200–299) and higher probability for all other SIC1 groups. We also observed that temporal trends varied by analytical method, type of OSHA plan, and broad industry categories. For example, the odds of a detectable result decreased 6% per year for SIC1 = 3 (1 – exp($\mathrm{B}_{\text{Year-1984}}$)) and 11% for SIC1 = 8 (1 – exp($\mathrm{B}_{\text{Year-1984}}$) *exp($\mathrm{B}_{\text{Year-1984, SIC1 = 8}}$)).

In the intensity model (Table 1), we observed declining time trends for samples analyzed using ICP, for measurements collected in states with partial state OSHA plans, and for samples collected in SIC1 group 5/6 (Wholesale and retail trade, finance, insurance, real estate). Increasing time trends were observed for SIC1 group 0/1 (Agriculture, forestry, fishing, mining, construction). Lead concentrations were lower when analyzed using ICP vs. AAS, for duration > 600 min vs. 120–600 min, and when analyzed by AAS for multiple metals vs. only lead. Lead concentrations were higher for durations of 0–60 and 60–120 min compared to 120–600 min, for states with partial OSHA plans vs. federal OSHA plans, and for SIC1 groups 0/1, 5/6, 7, and 9 compared to SIC1 = 3 (see Table 2 for industry category definitions). The between-industry variance (*Var(u*$_2$*)*) was much smaller than the residual variance (σ_e^2).

The covariance between the occurrence and intensity models was statistically different than 0 at 0.30 (95% confidence interval: 0.18–0.42)

Industry-specific estimates

The predicted industry-specific estimates are listed in Supplemental Table 1 for all SIC4 codes. Across the 575 SIC4

Table 1 Model parameters for the occurrence and intensity models.

Variable	N measurements in category	Occurrence Model: Logistic Regression (≥ 0.007 vs. <0.007)		Intensity Model: Linear Regression (concentration lognormally transformed)	
		Relative difference, exp(ß)	95% Confidence Interval	Relative difference, exp(ß)	95% Confidence Interval
Intercept	52,467	0.76	(0.63–0.93)	0.05	(0.04–0.06)
Year-1984	52,467	0.94	(0.93–0.95)	1.00	(0.99–1.00)
Analytical method					
ICP	35,661	0.20	(0.18–0.21)	0.81	(0.75–0.87)
AAS	16,806	1.00		1.00	
Sample duration			(1.00–1.00)		(1.00–1.00)
0–60 min	3121	0.44	(0.40–0.49)	2.90	(2.55–3.27)
60–120 min	2954	0.90	(0.81–0.99)	1.54	(1.40–1.69)
120–600 min	46,163	1.00		1.00	
>600 min	229	0.89	(0.62–1.26)	1.09	(0.82–1.46)
OSHA Plan					
State	26,000	0.99	(0.92–1.08)	1.03	(0.96–1.11)
Partial state	5775	0.85	(0.73–0.97)	1.13	(0.97–1.31)
Federal	20,692	1.00		1.00	
Number of metals analyzed, if analyzed by AAS					
1, Lead only	5863	1.00		1.00	
>1, Metal panel	46,604	1.18	(1.08–1.30)	0.72	(0.67–0.77)
Broad Industry Group (SIC1987 1st digit)					
SIC1 = 0 or 1, Agriculture, forestry, fishing, mining, construction	4993	4.60	(2.72–7.86)	2.37	(1.81–3.09)
SIC1 = 2, Manufacturing industries 200–299	2363	0.86	(0.57–1.29)	1.04	(0.79–1.37)
SIC1 = 3, Manufacturing industries 300–399	39,404	1.00		1.00	
SIC1 = 4, Transportation, communications, electric, gas, sanitary service	999	1.50	(0.82–2.75)	1.37	(0.91–2.06)
SIC1 = 5 or 6, Wholesale and retail trade, finance, insurance, real estate	2115	1.93	(1.13–3.31)	1.62	(1.15–2.28)
SIC1 = 7, Services industries 700–799	1789	2.77	(1.60–4.81)	1.74	(1.26–2.40)
SIC1 = 8, Service industries 800–899	278	1.06	(0.39–2.86)	1.14	(0.50–2.64)
SIC1 = 9, Public administration	526	1.98	(0.83–4.73)	1.79	(1.03–3.13)
Interactions with year					
(Year–1984)*Analytical Method ICP	35,661	1.03	(1.02–1.04)	0.98	(0.98–0.99)
(Year-1984)*State OSHA plan	26,000	1.01	(1.00–1.02)	1.00	(1.00–1.01)
(Year-1984)*Partial State OSHA plan	5775	1.01	(1.00–1.02)	0.98	(0.97–0.99)
(Year-1984)*SIC1 = 0 or 1	4993	0.98	(0.97–0.99)	1.02	(1.01–1.03)
(Year-1984)*SIC1 = 2	2363	0.98	(0.96–1.00)	1.01	(0.99–1.03)
(Year-1984)*SIC1 = 4	999	1.01	(0.99–1.04)	1.01	(0.99–1.04)
(Year-1984)*SIC1 = 5 or 6	2115	0.98	(0.97–1.00)	0.99	(0.97–1.00)
(Year-1984)*SIC1 = 7	1789	0.99	(0.97–1.01)	1.00	(0.98–1.01)
(Year-1984)*SIC1 = 8	278	0.95	(0.89–1.01)	0.99	(0.92–1.07)
(Year-1984)*SIC1 = 9	526	0.97	(0.93–1.00)	0.97	(0.94–1.01)
Variance components		Variance (SE)		Variance (SE)	
$Var(\mu_1)$		1.55 (0.15)			
$Var(\mu_2)$				0.237 (0.035)	
σ_e^2				2.14 (0.02)	
$Cov\ (u_1,u_2)$				0.303 (0.061)	

codes, the median predicted detection probability was 0.51 (IQR 0.37–0.71), the median predicted detected concentration was 0.041 mg/m^3 (IQR 0.033–0.54), and the median predicted lead concentration adjusted for the probability of detection was < 0.007 mg/m^3 (75th percentile, 0.015 mg/m^3). For the 575 industries, the distribution of the predicted probability of an industry-specific measurement being > 0.007 mg/m^3 is shown in Fig. 1 and the distribution of the median predicted lead concentrations is shown in Fig. 2.

Table 2 Predicted probability of detection, median concentration of detected values, and predicted median concentration for the 20 SIC4 industry codes with the most measurements.

4-digit SIC code	Industry	*N*	% ≥0.007	% ICP	% AAS	Probability concentration ≥0.007 (95% CI)	Predicted median concentration of values ≥0.007 (95% CI)	Predicted median concentration, adjusted for probability ≥0.007 (95% CI)
3691	Storage batteries	2485	93	15	85	0.972 (0.965–0.977)	0.058 (0.047–0.073)	0.052 (0.042–0.066)
3341	Copper smelting and refining, secondary-MFG	2431	73	48	52	0.919 (0.899–0.933)	0.072 (0.058–0.091)	0.054 (0.042–0.069)
3366	Copper foundries	2135	75	67	33	0.944 (0.929–0.953)	0.051 (0.041–0.064)	0.041 (0.033–0.052)
1721	Painting and paper hanging	1769	82	40	60	0.968 (0.949–0.979)	0.356 (0.247–0.540)	0.317 (0.215–0.488)
3321	Gray and ductile iron foundries	1755	34	89	11	0.787 (0.743–0.818)	0.028 (0.023–0.036)	0.012 (0.009–0.017)
5093	Scrap and waste materials	1496	61	72	28	0.927 (0.869–0.959)	0.069 (0.047–0.103)	0.053 (0.033–0.080)
3443	Fabricated plate work (boiler shops)	1264	10	92	8	0.449 (0.389–0.498)	0.033 (0.027–0.041)	<0.007 (<0.007)
3312	Blast furnaces	1106	43	76	24	0.818 (0.779–0.846)	0.031 (0.025–0.039)	0.016 (0.012–0.021)
3441	Expansion joints (structural shapes): iron and steel-MFG	1002	9	94	6	0.431 (0.371–0.479)	0.038 (0.030–0.047)	<0.007 (<0.007)
3499	Fabricated metal products	986	9	91	9	0.447 (0.387–0.495)	0.051 (0.041–0.064)	<0.007 (<0.007)
3471	Plating and polishing	940	13	91	9	0.505 (0.444–0.554)	0.033 (0.027–0.042)	0.001 (<0.007–0.004)
3714	Motor vehicle parts	870	24	74	26	0.620 (0.561–0.665)	0.026 (0.021–0.033)	0.005 (0.003–0.007)
1799	Special trade contractors	845	55	48	52	0.891 (0.835–0.926)	0.175 (0.121–0.265)	0.116 (0.075–0.186)
3339	Primary nonferrous metals, NEC	778	83	29	71	0.939 (0.924–0.950)	0.086 (0.069–0.108)	0.069 (0.055–0.087)
3325	Steel foundries, NEC	777	16	92	8	0.577 (0.516–0.624)	0.023 (0.019–0.029)	0.003 (0.001–0.005)
3365	Aluminum foundries	726	43	75	25	0.815 (0.775–0.843)	0.045 (0.036–0.056)	0.022 (0.017–0.029)
3523	Farm machinery and equipment	717	12	92	8	0.495 (0.434–0.544)	0.078 (0.063–0.098)	<0.007 (<0.007–0.007)
3731	Ship building and repairing	701	27	83	17	0.716 (0.663–0.754)	0.061 (0.049–0.077)	0.019 (0.014–0.027)
3715	Truck trailers	683	15	91	9	0.541 (0.479–0.589)	0.069 (0.056–0.087)	0.005 (<0.007–0.011)
3444	Pipe, sheet metal*MFG	626	8	95	5	0.406 (0.348–0.454)	0.028 (0.023–0.036)	<0.007 (<0.007)

For these predictions, we set the following variables to the following reference categories: year = 1984, analytical method = AAS, sample duration 120–600 min, OSHA plan = federal, number of analytes on a panel = 1.

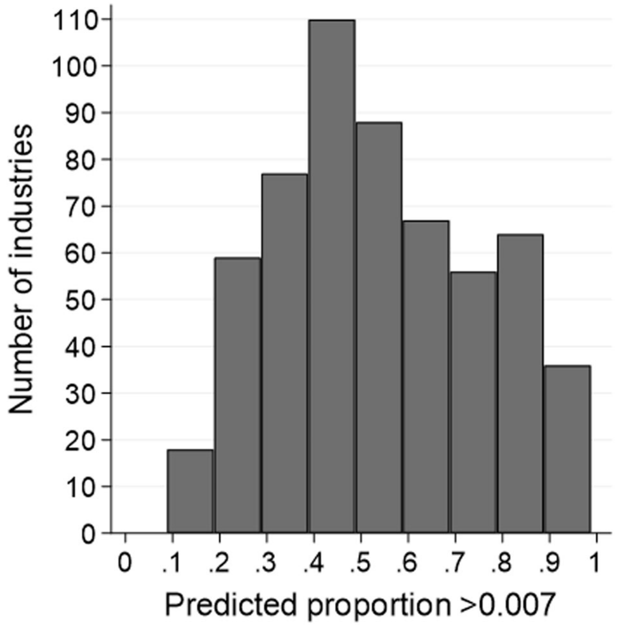

Fig. 1 Distribution of predicted probability of a measurement being ≥0.007 mg/m^3 for 575 industries. Each bar shows the number of industries for each 0.1 unit range of predicted probability.

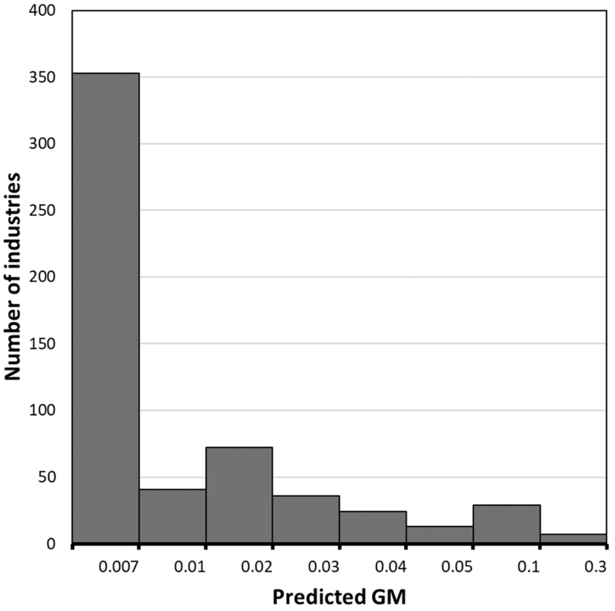

Fig. 2 Distribution of predicted GMs for all 575 industries. Each bar shows the number of industries for each range of predicted GMs.

The predicted industry-specific estimates for the 20 SIC4 codes with the most samples are listed in Table 2. The highest predicted median concentration, adjusted for probability of detection, was for SIC4 = 1721 (Painting and Paper Hanging). This SIC4 was the 4th most frequently measured SIC code with 1769 samples and had a predicted probability of detection of 0.968 (95%CI: 0.949–0.979), a predicted median detected concentration of 0.356 (95%CI: 0.247–0.540) mg/m^3, and predicted median concentration adjusted for probability of detection of 0.317 (95%CI: 0.215–0.488) mg/m^3. The highest predicted detection probability was observed for the SIC code with the most measurements, SIC4 = 3691 Storage Batteries. However, 5 of the 20 most frequently measured SIC4 codes (SIC4 codes: 3441, 3443, 3444, 3499, 3523; defined in Table 2) has predicted probabilities less than 0.50; measurements for these SIC4 codes were predominantly analyzed using ICP (>90%). As a result, although predicted median detected concentrations ranged from 0.028 to 0.51 mg/m^3, for these SIC4 codes (and all others with predicted probability of detection < 0.5) we can only state that the overall median lead concentrations was < 0.007 mg/m^3.

Figure 3 shows the moderate to high correlation between the industry-specific predicted detection probability (*x*-axis) and the predicted median detected concentration (y-axis) (rho = 0.73).

Discussion

Semi-continuous modeling of workplace lead measurements allowed us to identify important exposure determinants, such as temporal trends and industry-specific differences, in a data set with 58% of measurements reported as “0,” without requiring assumptions on the shape of the exposure distribution for the non-quantified measurements. The substantial co-variance observed between the model occurrence and intensity components supports our use of semi-continuous models in this case study and suggests that model estimates would have been biased if the occurrence and intensity components were modeled separately. Moreover, this approach also allowed us to retain generalizability of our findings to the entire group of industries monitored by OSHA and analyzed by OSHA’s Salt Lake City laboratory, in comparison to analyses that would restrict the data set to industries with a sufficient detection rate.

The same determinants were statistically significant in both the occurrence and intensity models; however, the directions of effect were not always the same. Temporal effects varied by analytical method, type of OSHA plan, and industry. Across all interactions, we observed a decreasing probability of detecting a lead result above our threshold over time. In contrast, the median magnitude of the detected concentrations increased over time. Using the two components together led to an overall decrease in median lead concentrations that is consistent with temporal patterns observed in analyses of lead measurements from OSHA’s IMIS database [24] and other sources of lead measurements [7, 8].

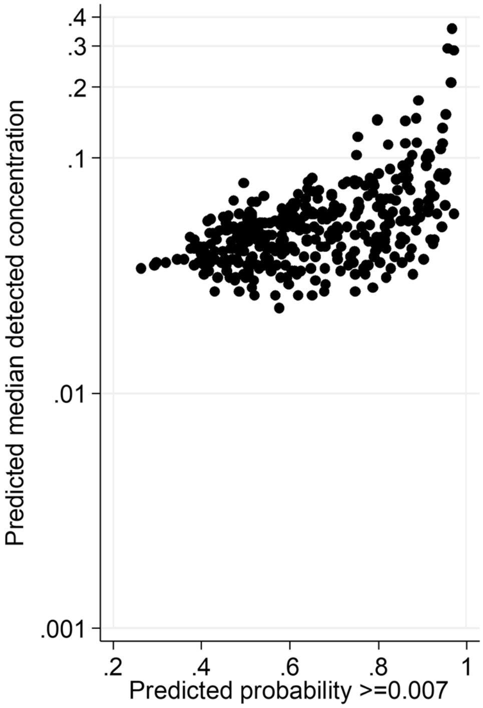

Fig. 3 Predicted probability of a measurement being ≥0.007 mg/m^3 vs. predicted median detected concentration. Each data point represents the predicted median detected concentration on the y-axis and the predicted probability on the x-axis for one of 575 industries.

Samples analyzed by ICP for a panel of metals were 80% less likely to be above our lead threshold and had median detected lead concentrations 20% lower than those analyzed by AAS for only lead. Samples analyzed by AAS for multiple metals were 18% more likely to be above the lead threshold and had median detected concentrations 28% lower than those analyzed by AAS for only lead. This analytical difference was expected given the higher likelihood that lead may not be the target analyte on panel samples compared to a targeted single analyte analysis. Also, for years 1984 through 1987 the AAS detection limit appeared to be one-third lower than the ICP detection limit. Analytical method was not a significant predictor in a meta-regression of published data on lead exposure for activities disturbing materials painted with or containing lead [24]; however, these were activities with high potential for lead exposure.

Lead measurements from states with a partial state OSHA plan were 15% less likely to be detectable and had detected concentrations 13% higher compared to measurements from states with an exclusively State or Federal OSHA plan. These small differences likely reflect the differing responsibilities and sampling strategy of the two separate OSHA regulatory units, potentially their choice of analytical laboratories, as well as differences in regulations. We observed only minimal differences in patterns for both model components for measurements collected under state versus federal OSHA plans; in contrast, analyses of the OSHA lead data in the IMIS database found a 1.18 (95%CI: 1.11–1.26) times higher odds of observing a lead concentration above a threshold limit value of >0.05 mg/m^3 and an 11% (95% CI: 0.75–1.06) decrease in relative intensity of lead exposure for measurements from state versus federal OSHA plans [25].

Broad industry differences in detection rate and median detected concentration were also observed, usually simultaneously within both model components. The highest detection rates and median detected concentrations occurred for lead samples from SIC1 = 0/1 (Agriculture, forestry, fishing, mining); these measurements predominantly represent mining scenarios. As with all broad industry group effects in unbalanced data, one must be cautious about generalize its effect to specific industries in the group that are poorly represented in the database. Industry-to-industry comparisons are beyond the goals of this paper, given our primary focus on applying semi-continuous models, as well as the breadth of industries covered, differences in how industries are grouped in other publications, and differences in evaluated measure of central tendency (e.g., GMs vs. arithmetic means). However, we refer the reader to a thorough literature review of published occupational lead exposure measurements in US workplaces, which includes tables of weighted arithmetic means [26].

The inclusion of industry as a random effect allowed for estimating industry-specific median lead concentrations while identifying population-level patterns. Many industries had no detectable lead concentrations; however, their retention in the models will contribute to future exposure assessment efforts that need to characterize lead across the range of U.S. industries. We noted that the number of measurements was a poor proxy for likelihood of detectable lead concentrations (rho = 0.44). For example, four of the 20 most frequently monitored industries had predicted median concentrations below our 0.007 mg/m^3 threshold (Table 2), which is likely because another metal on the analyte panel was the analyte of concern. The inclusion of broad industry group, SIC1, as a fixed-effect, and more detailed industry group, SIC4, as a random-effect, allowed the industry-specific estimate to shrink towards the broad industry effect when the more detailed industry group had a small sample size or were highly variable or pull towards the industry median when data were more plentiful and less variable [23]; however, as noted above, extrapolation to industries with few to no measurements should be made cautiously.

There were several limitations of the CEHD data. First, the workplaces monitored by OSHA do not represent a random sample of workplace or industries in the United States and selective biases in this data source have been hypothesized [11]. However, it remains one of the few sources of workplace air measurements over multiple decades and thus provides a reasonable proxy for population-level patterns over time. Second, the determinants evaluated in these analyses were limited to ancillary data reported to the OSHA laboratory and were primarily related to the measurement, such as sample duration and analytical method. Other variables related to the inspection (e.g., reason for measurement) or the jobs (e.g., job title) are available only in the IMIS data set. These analyses included ~31,000 lead measurements that were also found in IMIS, but excluded ~14,700 measurements only found in IMIS. Addressing the challenges with analyzing the CEHD data, in particular the issue of "present and not quantified" vs. "absent," was a crucial first step towards being able to jointly evaluate trends from both data sources.

Applying mixed-effect semi-continuous models to environmental measurements that are suspected to have a cluster at (or near) zero concentration addresses a large challenge in analysis of environmental data. We used previously developed code to apply the semi-continuous mixed-effect modeling to our data. However, we had to develop new code to combine the two model components to extract the median lead concentration for each industry and obtain 95% confidence intervals. The modeling approach had a few limitations. First, a single threshold value was required, thus changing limits of detection over time or by analytical method were not captured. This resulted in treating 3.6% of the measurements with a quantified concentration below the 0.007 mg/m^3 threshold as being treated as "not-detected". However, treating these measurements as "not-detected" has a negligible bias on the predicted medians because the predicted industry-specific estimates are based on the distributional parameters of the two components and thus does not require us to assume a distributional shape for the concentrations below the threshold value or to determine whether the sample result is a "true zero" or "present but unquantified." Second, semi-continuous models can include only a single random effect variable thus far. As a result, we were unable to account for the correlation between measurements collected within the same facility across multiple inspection dates. Only 1% of the inspections had measurements collected across multiple years (representing 4% of the measurements); however, 11% of the monitored establishments had repeated inspections (representing 34% of the measurements). This limitation can be addressed with further programming of the macro. Third, substantial computing power was necessary for obtaining 95% confidence intervals by bootstrap.

In summary, semi-continuous mixed-effects models represent the current best practice when the data has a cluster at zero or when deviations from a normal distribution is suspected, but to-date has not been applied to environmental air monitoring data. We have adapted the model to extend the coding to extract prediction estimates that combine both the occurrence and intensity models. To help other researchers apply this approach, we provide all code in on-line appendices.

Acknowledgements We thank Janet Tooze for providing the SAS macro to conduct the mixed distribution modeling and for providing permission to include that code in our supplemental materials. This work was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI.

Compliance with ethical standards

Conflict of interest The authors declare no competing interest.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Ganser GH, Hewett P. An accurate substitution method for analyzing censored data. J Occup Environ Hyg. 2010;7:233–44.
2. Helsel D. Much ado about next to nothing: incorporating nondetects in science. Ann Occup Hyg. 2010;54:257–62.
3. Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. Ann Occup Hyg. 2007;51:611–32.
4. Huynh T, Quick H, Ramachandran G, Banerjee S, Stenzel M, Sandler DP, et al. A Comparison of the beta-Substitution Method and a Bayesian Method for Analyzing Left-Censored Data. Ann Occup Hyg. 2016;60:56–73.
5. Huynh T, Ramachandran G, Banerjee S, Monteiro J, Stenzel M, Sandler DP, et al. Comparison of methods for analyzing left-censored occupational exposure data. Ann Occup Hyg. 2014;58:1126–42.
6. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004;112: 1691–6.
7. Creely KS, Cowie H, Van Tongeren M, Kromhout H, Tickner J, Cherrie JW. Trends in inhalation exposure-a review of the data in the published scientific literature. Ann Occup Hyg. 2007;51:665–78.
8. Koh DH, Nam JM, Graubard BI, Chen YC, Locke SJ, Friesen MC. Evaluating temporal trends from occupational lead exposure data reported in the published literature using meta-regression. Ann Occup Hyg. 2014;58:1111–25.
9. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics. 1999;55:625–9.
10. Taylor DJ, Kupper LL, Rappaport SM, Lyles RH. A mixture model for occupational exposure mean testing with a limit of detection. Biometrics. 2001;57:681–8.
11. Lavoue J, Friesen MC, Burstyn I. Workplace measurements by the U.S. Occupational Safety and Health Administration since 1979: Descriptive analysis and potential uses for exposure assessment. Ann Occup Hyg. 2013;57:681–3.
12. Lavoue J, Friesen MC, Burstyn I. Workplace measurements by the US Occupational Safety and Health Administration since 1979: descriptive analysis and potential uses for exposure assessment. Ann Occup Hyg. 2013;57:77–97.
13. Moulton LH, Halsey NA. A mixture model with detection limits for regression analyses of antibody response to vaccine. Biometrics. 1995;51:1570–8.

14. Tooze JA, Grunwald GK, Jones RH. Analysis of repeated measures data with clumping at zero. Stat Methods Med Res. 2002;11:341–55.
15. Su L, Tom BD, Farewell VT. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics. 2009;10:374–89.
16. Siega-Riz AM, Sotres-Alvarez D, Ayala GX, Ginsberg M, Himes JH, Liu K, et al. Food-group and nutrient-density intakes by Hispanic and Latino backgrounds in the Hispanic Community Health Study/Study of Latinos. Am J Clin Nutr. 2014;99:1487–98.
17. Wu X, Bennett DH, Lee K, Cassady DL, Ritz B, Hertz-Picciotto I. Longitudinal variability of time-location/activity patterns of population at different ages: a longitudinal study in California. Environ Health. 2011;10:80.
18. Tran V, Liu D, Pradhan AK, Li K, Bingham CR, Simons-Morton BG, et al. Assessing risk-taking in a driving simulator study: modeling longitudinal semi-continuous driving data using a two-part regression model with correlated random effects. Anal Methods Accid Res. 2015;5-6:17–27.
19. Biro FM, Pinney SM, Schwartz RC, Huang B, Cattran AM, Haslam SZ. Amphiregulin as a Novel Serum Marker of Puberty in Girls. J Pediatr Adolesc Gynecol. 2017;30:535–9.
20. Townsend JC, Steinberg DH, Nielsen CD, Todoran TM, Patel CP, Leonardi RA, et al. Comparison of lipid deposition at coronary bifurcations versus at nonbifurcation portions of coronary arteries as determined by near-infrared spectroscopy. Am J Cardiol. 2013;112:369–72.
21. Kreif N, Gruber S, Radice R, Grieve R, Sekhon JS. Evaluating treatment effectiveness under model misspecification: A comparison of targeted maximum likelihood estimation with bias-corrected matching. Stat Methods Med Res. 2016; 25:2315–36.
22. Sarazin P, Burstyn I, Kincl L, Friesen MC, Lavoue J. Characterization of the Selective Recording of Workplace Exposure Measurements into OSHA’s IMIS Databank. Ann Work Expo Health. 2018;62:269–80.
23. Friesen MC, Coble JB, Lu W, Shu XO, Ji BT, Xue S, et al. Combining a job-exposure matrix with exposure measurements to assess occupational exposure to benzene in a population cohort in shanghai, china. Ann Occup Hyg. 2012;56:80–91.
24. Locke SJ, Deziel NC, Koh DH, Graubard BI, Purdue MP, Friesen MC. Evaluating predictors of lead exposure for activities disturbing materials painted with or containing lead using historic published data from U.S. workplaces. Am J Ind Med. 2017;60:189–97.
25. Sarazin P, Burstyn I, Kincl L, Lavoue J. Trends in OSHA Compliance Monitoring Data 1979-2011: Statistical Modeling of Ancillary Information across 77 Chemicals. Ann Occup Hyg. 2016;60:432–52.
26. Koh DH, Locke SJ, Chen YC, Purdue MP, Friesen MC. Lead exposure in US worksites: A literature review and development of an occupational lead exposure database from the published literature. Am J Ind Med. 2015;58:605–16.