Identifying industry and time trends in lead exposure in OSHA laboratory measurements using semi-continuous models to account for possible true zeros in left-censored data

Melissa Friesen
Occupational and Environmental Epidemiology Branch
Division of Cancer Epidemiology and Genetics, NCI
friesenmc@mail.nih.gov

NIH NATIONAL CANCER INSTITUTE

September 2018

Acknowledgments

NCI Biostatistics Branch

Bin Zhu

Hyoyoung Choo-Wosoba

Paul Albert

Université de Montreal

Philippe Sarazin

Jerome Lavoue

NCI OEEB

Pamela Dopart (now at Exponent)

Jooyeon Hwang (now at Western Kentucky University)

Nicole Deziel, Yale

Daniel Russ, NIH CIT

Missing exposure concentration data

- Common problem
- Concentration below method detection limit (<LOD)
  - Range from 0 to >90% of data
- Multiple solutions
  - Logistic regression (e.g. ≥ Threshold vs. < Threshold)
  - ß-Substitution
  - Maximum likelihood
  - Imputation (e.g., using PROC LIFEREG, Lubin et al. 2004)
  - Tobit models
  - Bayesian approaches

Assumptions:

- **Exposure is present but not quantified**
- **Single exposure distribution**

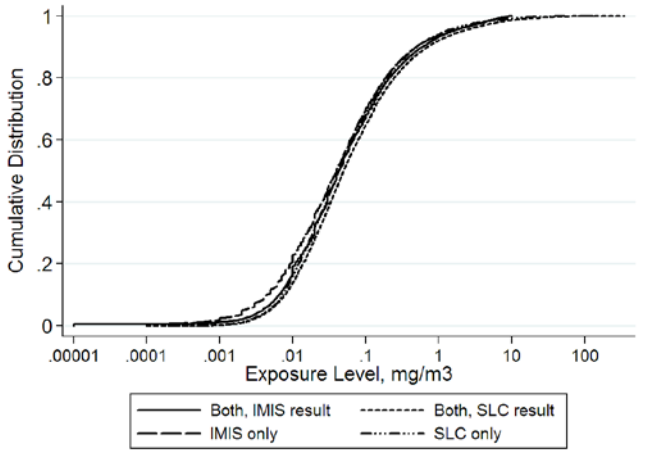
Analyte panels: Mixed distribution? True ‘zeros’?

- Measurement focus on X agent, but…
  - Multiple agents quantified and reported by lab
  - Primary agent(s) not identified in lab data set
  - E.g., metal panels

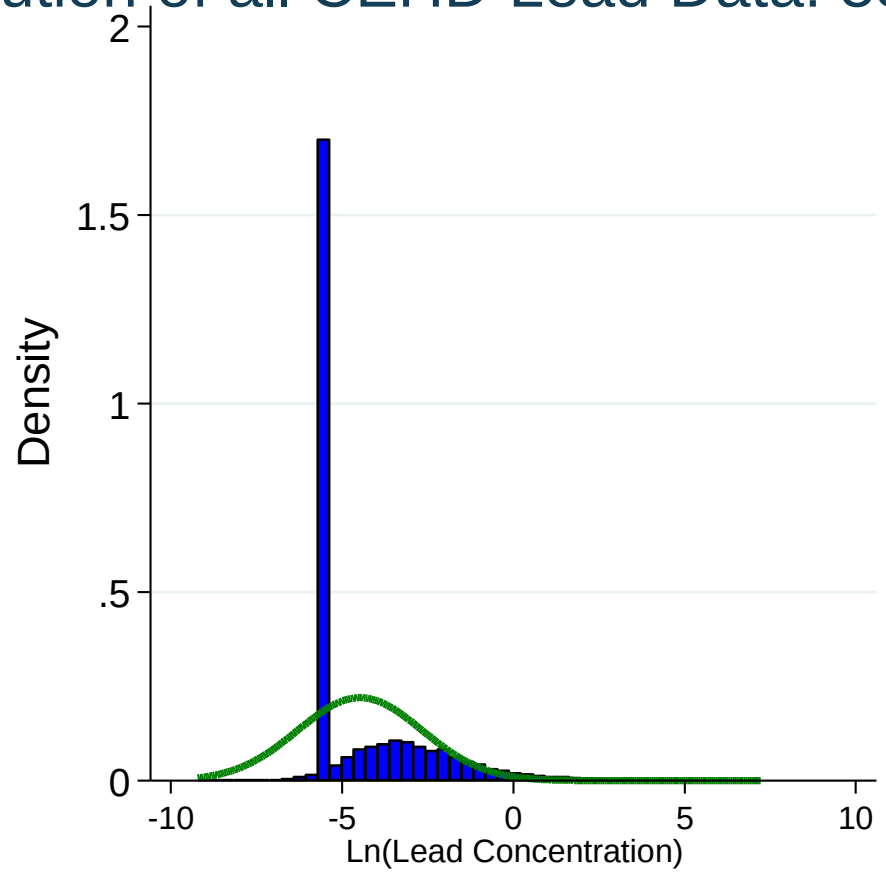
Case Study: Varying % LOD by data source

- OSHA inspection lead measurements (Lavoue et al. 2013)
  - Laboratory msmts: Chemical Exposure Hazard Database (CEHD) from Salt Lake City Laboratory
  - Inspection results: Integrated Management Information System (IMIS)

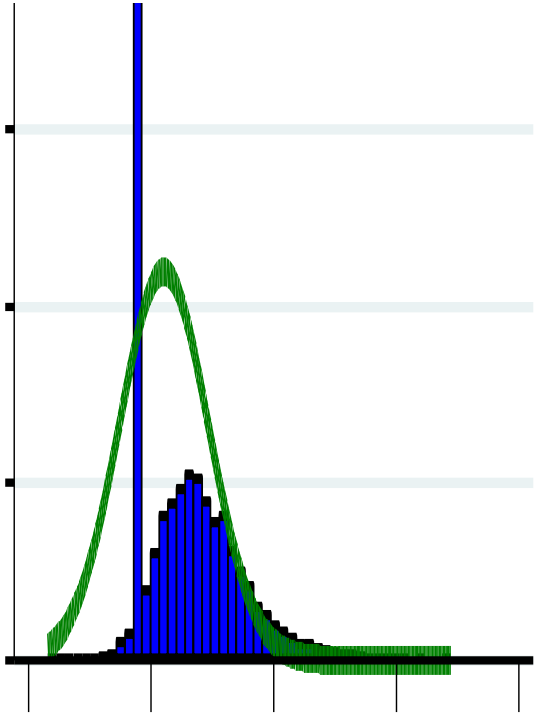
	% Sample	% < LOD
Both IMIS-CEHD	50	50
IMIS only	23	46
CEHD only	27	74



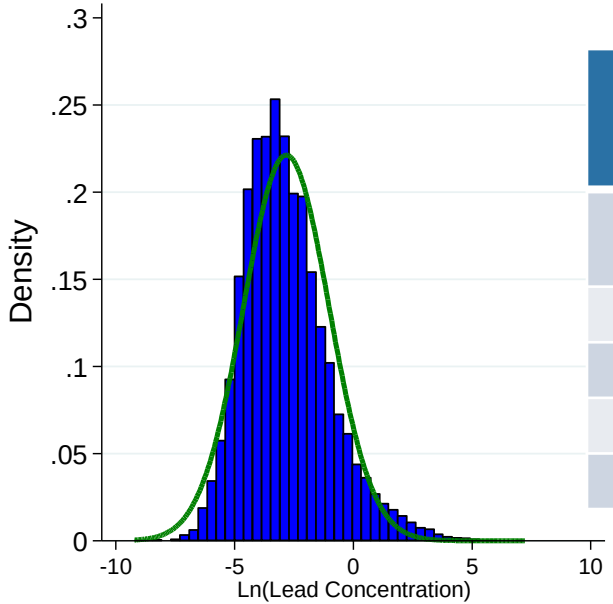
Distribution of all CEHD Lead Data: 58% not detected



CEHD Data



Assuming all ND's are absent



% True Zeros	GM (mg/m3)	Relative difference
0	0.011	**Ref.**
25	0.014	**1.3**
50	0.018	**1.6**
75	0.028	**2.5**
100	0.059	**5.4**

Can’t disentangle ‘present but not quantified’ vs. ‘true zero’

- Assume ‘present’ ?
  - E.g., imputation, MLE, substitution
  - Underestimate true exposure scenarios
  - Overestimate ‘true zero’ or ‘absent’ scenarios
- Restrict data set or two stage model ?
  - A priori knowledge about presence of exposure?
  - Base on likelihood of detection?
  - Ignores covariance, biased estimates
  - Narrows generalizability and inference

Semi-continuous modeling: CEHD lead data

- Joint modeling:
  1) Logistic regression:
     - Probability of detecting lead ≥7 μg/m^3.
  2) Linear regression:
     - Predictors of concentrations ≥7 μg/m^3

  - Set threshold at highest LOD over time period/method
  - Includes co-variance term for two model components
  - Allows one random-effect term per model (4-digit SIC code)
  - Allows different predictors for each model component

Descriptive evaluation

- 52,457 measurements
- 62% <0.007 mg/m^3
  - 75% ICP, 34% AAS
- 575 SIC4
  - Median 14 measurements (IQR 3-61)
  - Median 11% measurements ≥0.007 mg/m^3 (IQR 0-35%)
  - 203 SIC4 codes with no measurements ≥0.007 mg/m^3
  - N measurements does not predict % detected (rho = 0.44)

Determinants of exposure

	Probability ≥7 µg/m^3	Concentration
Year	↘ with time	≠
Analytical Method	ICP << AAS	ICP < AAS
Panel sample	No < Yes*	Yes < No
Sample duration	0-60 < 60-120 < 120-600	0-60 > 60-120 > 120-600
OSHA plan	State/Partial < Federal	State/Federal < Partial
Broad SIC grp (1-digit SIC)	RRs: 0.86-4.6	RRs: 1.0-2.4
Year-interactions	Yes	Yes

*Correlated with analytical method. All ICP samples were panel samples.

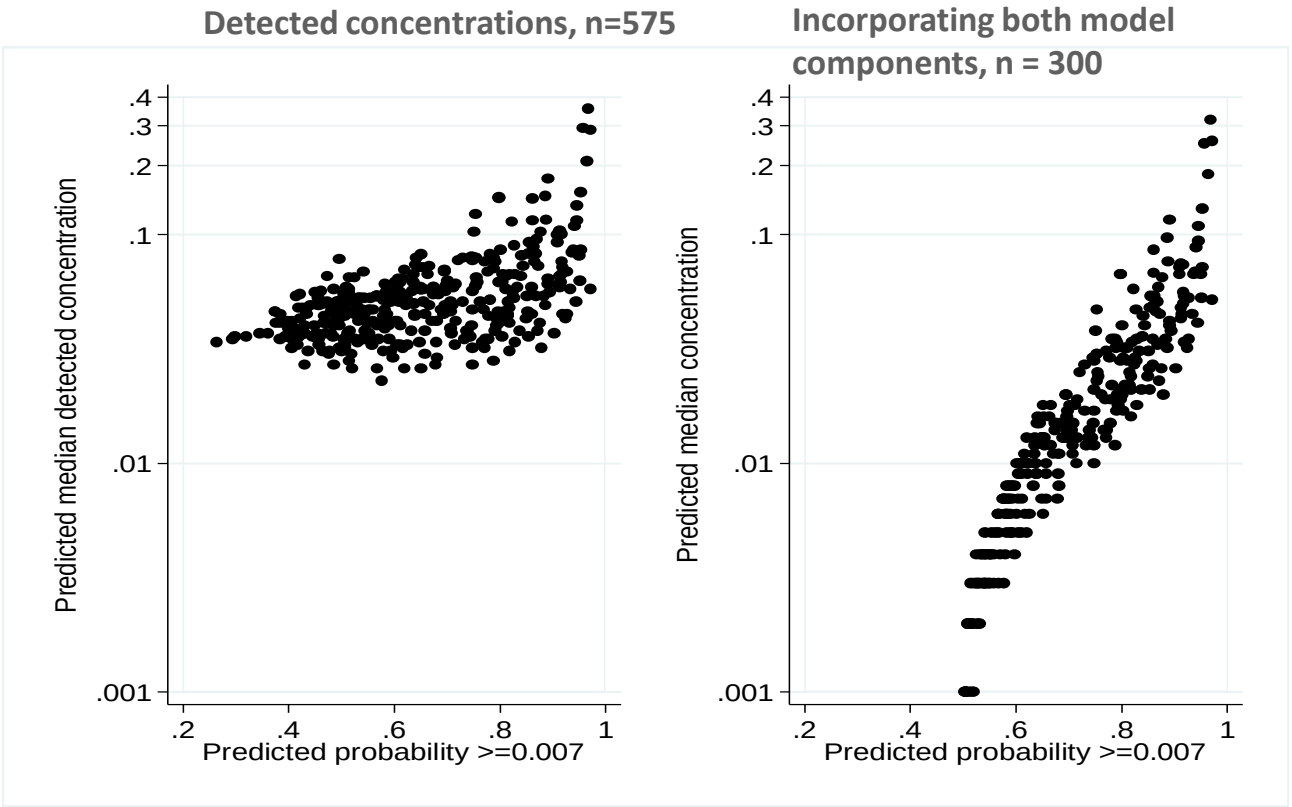
Obtaining SIC-specific estimate of exposure

- If SIC4 probability ≥0.007 mg/m^3 is <50%, then **SIC4 GM <0.007 mg/m^3**

- If SIC4 probability ≥0.007 mg/m^3 is ≥50%, then used distributional parameters from the cumulative distribution function to estimate exposure
  - 95% confidence intervals by bootstrap

For example, if the probability was 30% from the logistic component, we would obtain the ~20th percentile from linear regression component

Predicted SIC4 Pb concentration in mg/m^3, by predicted probability



Summary

Non-trivial covariance between models → joint modeling necessary

Strengths

- Inclusion of all data increases population-level generalizability
- Flexible modeling framework
- No assumptions on whether censored data were true zeros or present but unquantified

Limitations

- Limited to single threshold value,
- 1 random effect per component
- Some complexity in prediction using both components → code available

www.cancer.gov

www.cancer.gov/espanol