

City University of New York (CUNY)

CUNY Academic Works

Publications and Research

Brooklyn College

2011

Administrative Censoring in Ecological Analyses of Autism and a Bayesian Solution

Scott M. Bartell

University of California, Irvine

Thomas A. Lewandowski

CUNY Brooklyn College

[How does access to this work benefit you? Let us know!](#)

More information about this work at: https://academicworks.cuny.edu/bc_pubs/56

Discover additional works at: <https://academicworks.cuny.edu>

This work is made publicly available by the City University of New York (CUNY).

Contact: AcademicWorks@cuny.edu

Research Article

Administrative Censoring in Ecological Analyses of Autism and a Bayesian Solution

Scott M. Bartell¹ and Thomas A. Lewandowski^{2,3}

¹Program in Public Health and Department of Statistics, University of California, Irvine, 2241 Bren Hall, Irvine, CA 92697-1250, USA

²Gradient, Seattle, WA 98101-1248, USA

³Department of Health and Nutrition Sciences, Brooklyn College, Brooklyn, NY 11210, USA

Correspondence should be addressed to Scott M. Bartell, sbartell@uci.edu

Received 22 September 2010; Accepted 2 March 2011

Academic Editor: Pam R. Factor-Litvak

Copyright © 2011 S. M. Bartell and T. A. Lewandowski. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Widely cited ecological analyses of autism have reported associations with mercury emissions, with precipitation, and race at the level of counties or school districts. However, state educational agencies often suppress any low numerical autism counts before releasing data—a phenomenon known as “administrative censoring.” Previous analyses did not describe appropriate methods for censored data analysis; common substitution or exclusion methods are known to introduce bias and produce artificially narrow confidence intervals. We apply a Bayesian censored random effects Poisson model to reanalyze associations between 2001 Toxic Release Inventory reported mercury emissions and 2000-2001 autism counts in Texas. Relative risk estimates for autism decreased from 4.44 (95% CI: 4.16, 4.74) per thousand lbs. of air mercury emissions using a naive zero-substitution approach to 1.42 (95% CI: 1.09, 1.78) using the Bayesian approach. Inadequate attention to censoring poses a serious threat to the validity of ecological analyses of autism and other health outcomes.

1. Introduction

The prevalence of autism has been increasing in the US over the last few decades, but its causes are not well understood [1]. Several ecological analyses of autism have appeared in the last five years, reporting correlations with reported mercury emissions [2, 3], with precipitation [4], and with race [5]. These study findings have been cited extensively by other peer-reviewed publications [6] and widely circulated by the popular press [7, 8]. Although it is well known that ecological studies may be affected by aggregation bias when used to estimate individual level effects [9], ecological studies of medical conditions such as autism are typically susceptible to an additional hazard that is seldom recognized: censoring of low disease counts.

Autism is not a reportable disease, but states do track and report the total number of students receiving special education services for conditions such as autism to the US Department of Education under the 1990 Individuals

with Disabilities Education Act (P.L. 101–476, Section 618). Statewide totals are routinely published by the US Department of Education [10], but prevalence by school district or county must be obtained from each state’s department of education. Countywide environmental releases of mercury are also available publicly through the US Environmental Protection Agency Toxic Release Inventory, through a mandatory reporting system that includes any facility emitting 10 lbs. or more of mercury per year.

Educators have legal and moral obligations to protect the privacy of students and their educational records. Some states impose limits upon enumeration of small numbers of students receiving particular special education services in order to ensure privacy for these students. For example, the Texas Educational Agency provided us with exact autism counts for districts with 5 or more students with autism and for districts with no students with autism, but it used a special code (–999) for districts with 1–4 students with autism. 35% of the districts fell into the latter category requiring the

TABLE 1: Distribution of 2000-2001 autism counts* for Texas school districts.

| Students with autism | Number of school districts |
|----------------------|----------------------------|
| 0 | 451 |
| 1-4 (censored) | 362 |
| 5 | 18 |
| 6 | 16 |
| 7 | 16 |
| 8 | 15 |
| 9 | 7 |
| 10 | 10 |
| 11-20 | 53 |
| 21-50 | 44 |
| >50 | 37 |

* For brevity, this table summarizes counts greater than 10 using several categories though exact counts are available for districts with more than 4 students with autism. Only counts of 1 to 4 are censored.

special code (Table 1). Similarly, Waldman et al. [4] reported that Oregon censored the autism counts (i.e., substituted a special code for the actual counts) for counties with fewer than 10 students with autism at any one year of age.

These autism counts take the form of “administratively censored” outcomes. Likelihood-based statistical methods for administratively censored outcomes have long been available for classical parametric fixed effects models [11] and are widely implemented for well-known distributions in major statistical packages. However, previous ecological studies of autism have cited statistical methods such as mixed effects Poisson models and ordinary least-squares regression that are not designed to handle censored observations [2–5]. The absence of any explicit description of methods for handling censored observations in these previous publications suggests that censored values were deleted or substituted with fixed values such as zeros. Indeed, one might substitute zeros after incorrectly assuming that a missing autism count indicates that a school district had no children with autism to report.

Palmer et al. [2, 3] fit random effects Poisson models associating 2000-2001 school district autism counts and 2001 countywide total mercury releases, adjusting for percent white, district wealth, percent economically disadvantaged, and urbanicity, but they did not discuss censoring. We were able to compile these variables for 1029 of the school districts, and previously reanalyzed the data for these districts in various years using the model described by Palmer et al. [2]. We previously reported similar results to Palmer’s for 2000-2001 when using the zero substitution approach, with RRs of 1.29 to 2.03 per 1000 lbs of total mercury releases [12], depending on the estimation algorithm, versus Palmer’s originally reported RR of 1.61. However, when we substituted larger fixed values (threes) for the censored counts, our RR estimates decreased to values near unity [12]. Statistically significant associations between mercury emissions and autism did not persist in later years.

Substantial differences between the results of *ad hoc* fixed value substitution methods indicate that the method

of handling censored counts is critical in this setting, motivating the following analysis.

2. Methods

In general, the contribution of each interval censored observation to the likelihood is $\Pr(L < Y_i \leq U)$, where Y_i is a random variable describing the count in district i and (L, U) is the interval in which the censored value lies. The following statements comprise a censored Poisson random effects model for the ecological association between mercury releases and autism using the Texas data:

$$\Pr(Y_i = y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad \text{for } y_i = 0 \quad \text{or} \quad y_i > 4$$

$$\Pr(0 < Y_i \leq 4)$$

$$= \sum_{y_i=1}^4 \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!},$$

for censored counts (y_i coded as -999),

$$\begin{aligned} \ln(\lambda_i) = & \ln(\text{DistrictSize}_i) + \mu + \beta_{0j} + \beta_1 \text{Mercury}_i \\ & + \beta_2 \text{White}_i + \beta_3 \text{Taxbase}_i + \beta_4 \text{Econ}_i + \beta_5 \text{Urban}_i \\ & + \beta_6 \text{Suburban}_i + \beta_7 \text{Other Urbanicity}_i, \end{aligned}$$

$$\beta_{0j} \sim N(0, \sigma_0^2),$$

(1)

where y_i is the observed autism count for school district i , λ_i is the Poisson rate parameter modeled as an exponentiated linear combination of seven covariates and an offset for the log of the number of students in the district, and β_{0j} is a random effect of county j (many counties have more than one school district). μ , σ_0^2 , and the β s are unknown parameters, and β_1 is the parameter of interest: the log RR for autism associated with each 1000 lbs increase in total mercury emissions. $N(0, \sigma_0^2)$ denotes a normal distribution with mean 0 and variance σ_0^2 .

Although censored Poisson fixed effects regression is widely implemented by major statistical software packages, interval-censored Poisson random effects regression is a newer area of statistical research. Moreover, frequentist Poisson random effects regression algorithms rely on differing likelihood approximations that are known to produce disparate estimates for these particular data [12] and for rare outcomes in general [13]. We use a Bayesian approach for parameter estimation instead, requiring the specification of prior distributions for the unknown parameters. We chose relatively uninformative priors

$$\mu \sim N(-6, 10),$$

$$\beta_k \sim N(0, 100) \quad \text{for all } k, \quad (2)$$

$$\sigma_0^2 \sim \text{Uniform}(0, 50).$$

We implemented the model using WinBUGS 1.4.3, a generalized environment for conducting Bayesian analyses

TABLE 2: Effect estimates per 1000 lbs of Toxic Release Inventory reported total mercury releases using various censored data methods.

| Approach | RR | 95% CI |
|------------------------------|------|------------|
| Bayesian censored likelihood | 1.18 | 1.07, 1.32 |
| Zero substitution | 2.02 | 1.96, 2.09 |
| One substitution | 1.26 | 1.17, 1.36 |
| Two substitution | 1.18 | 1.08, 1.28 |
| Three substitution | 1.14 | 1.02, 1.27 |
| Four substitution | 1.11 | 0.96, 1.29 |
| Exclusion | 1.16 | 1.14, 1.18 |

using Markov chain Monte Carlo methods, as well as the R2WinBUGS library in R. WinBUGS code is provided (see Algorithm 1). At least, 100,000 iterations were used in each analysis for each of 3 chains; 50% of the iterations were discarded for burn-in. Convergence of the sampling chains was assessed using plots and the R-hat statistic, which was less than 1.01 for each posterior RR for mercury and less than 1.05 for all other parameters.

For comparison purposes, naive approaches relying on either exclusion of censored observations or fixed-value substitution for censored observations were conducted using mixed effects Poisson regression in R as described previously [12] but with a more recent version of the lmer library (version 0.999375-31).

We also repeated all analyses for air mercury emissions in place of total mercury emissions. Air emissions are arguably a better proxy for concurrent exposure than total mercury emissions for reasons described in our previous analysis [12] and later in this paper.

3. Results

The resulting Bayesian posterior mean RR estimate per 1000 lbs of reported total mercury releases using the 2000-2001 data is 1.18 (95% confidence interval (CI): 1.07, 1.32). In contrast, the naive zero-substitution method yields an RR estimate of 2.02 (95% CI: 1.96, 2.09) per 1000 lbs. of total mercury releases. Table 2 shows the complete set of parameter estimates for the Bayesian censored likelihood analysis and for the naive substitution approach using various fixed values and for the exclusion approach. Although some of the naive approaches produce reasonable central estimates, they produce artificially narrow confidence intervals.

For air mercury emissions, the posterior mean RR estimate is 1.42 (95% CI: 1.09, 1.78) per 1000 lbs. of mercury. In contrast, the zero-substitution approach yields an RR estimate of 4.44 (95% CI: 4.16, 4.74) per 1000 lbs. of mercury. Complete results for air mercury emissions are shown in Table 3.

We checked for sensitivity to the choice of the prior by repeating the air mercury analysis with the prior for β_1 centered on $\log(4.44)$. This analysis yielded nearly identical results (RR = 1.42; 95% CI: 1.10, 1.80) to the primary analysis, suggesting that our choice of prior distribution for the effect of mercury is not overly influential.

TABLE 3: Effect estimates per 1000 lbs of Toxic Release Inventory reported air mercury emissions using various censored data methods.

| Approach | RR | 95% CI |
|------------------------------|------|------------|
| Bayesian censored likelihood | 1.42 | 1.09, 1.78 |
| Zero substitution | 4.44 | 4.16, 4.74 |
| One substitution | 1.64 | 1.40, 1.93 |
| Two substitution | 1.39 | 1.15, 1.69 |
| Three substitution | 1.28 | 1.00, 1.63 |
| Four substitution | 1.2 | 0.87, 1.65 |
| Exclusion | 1.37 | 1.31, 1.42 |

4. Discussion

Our results using the full censored likelihood indicate that the ecological association between total mercury emissions and 2000-2001 autism in Texas school districts may be much smaller than previously reported. Published associations between precipitation and autism are also dependent on censored data and could be affected by similar biases. RR estimates for the ecological association using the censored likelihood were fairly similar to the results using the *ad hoc* substitution approach when values near the middle of the censoring interval were chosen. However, confidence intervals were (appropriately) wider for the Bayesian likelihood-based approach than for the substitution approaches. Exclusion of all censored values produced similar central RR estimates to the Bayesian approach but with artificially narrow confidence intervals.

However, all results should also be viewed as ecologic effect estimates that may not be representative of individual level effects due to aggregation bias. Another consequence of the group-level analysis is that uncertainty regarding individual-level associations between mercury exposure and autism in Texas is much greater than suggested by the group-level confidence intervals reported here and in previous publications [3, 5, 12].

These results were surprising to us given that the cutoff for censoring was so low (5 students), but they may be explained by an apparent association between mercury emissions and the presence of censoring, whereby those districts with small autism counts tended to have lower mercury emissions. This effect may be better understood through inspection of Figure 1, which shows the data and model fit using a variety of fixed value substitution approaches. Because many of the censored values occur in counties with low mercury emissions, the overall intercept from the mixed effects Poisson regression is highly sensitive to the choice of fixed value substituted for the censored observations. In contrast, there are relatively few censored observations in counties with higher mercury releases, leading to a more stable position of the right-hand side of the prediction curve. The result appears to be a forcing of the slope (the log relative risk) to higher values when the intercept is lower.

Although both total mercury releases and air mercury emissions are poor surrogates for individual mercury exposures, air mercury emissions may be a better proxy for

```

model {
  for (i in 1:counties) {
    b0[i] ~ dnorm(0,tau)
  }
  for (j in 1:districts){
    zeros[j] <- 0
    phiobs[j] <- - lam[j] - y[j] * log(lam[j]) + 10000
    phicens[j] <- - lam[j] - log(pow(lam[j],4)/24 + pow(lam[j],3)/6 + pow(lam[j],2)/2
      + lam[j] ) + 10000
    # use zeros method for censored for counts between 1 and 4 (see WinBUGS manual)
    # y[j] is autism count for district j. Note that censored values are recorded as - 999.
    phi[j] <- - step(y[j]) * phiobs[j] + (1- step(y[j])) * phicens[j]
    zeros[j] ~ dpois(phi[j])
    lam[j] <- - exp(mu + b0[county[j]] + b[1]*airChem[j] + b[2]*white[j] + b[3]*taxbase[j] +
      b[4]*econ[j] + b[5]*urban[j] + b[6]*suburban[j] + b[7]*other[j] + log(students[j]))
  }
  # Note that WinBUGS uses precision not variance to define normal distributions
  mu ~ dnorm(- 6,0.1)
  b[1] ~ dnorm(0,0.01)
  b[2] ~ dnorm(0,0.01)
  b[3] ~ dnorm(0,0.01)
  b[4] ~ dnorm(0,0.01)
  b[5] ~ dnorm(0,0.01)
  b[6] ~ dnorm(0,0.01)
  b[7] ~ dnorm(0,0.01)
  tau <- - 1/var
  var ~ dunif(0,50)
  RR <- - exp(b[1])
}

```

ALGORITHM 1: WinBUGS code.

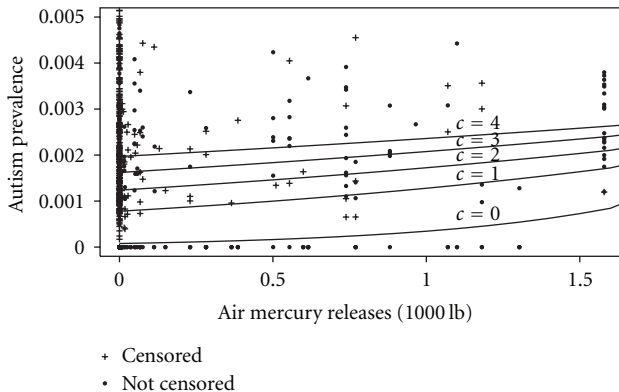


FIGURE 1: Autism prevalence versus Toxic Release Inventory reported air mercury emissions, with threes substituted for censored values ($c = 3$) for all data points. Lines show the covariate-adjusted random effects Poisson model predictions using five different fixed value substitution approaches (substituting different values of c for the censored counts).

concurrent exposures than total mercury releases. Total mercury releases include landfill disposals and other releases that may not lead to widespread or immediate public exposure, especially if compliant with the 1976 Resource Conservation Recovery Act (P.L. 94–580) and its 1984 amendment. In contrast, air mercury emissions are more

widely and rapidly dispersed with wind and rainfall and are, therefore, of the two, more plausible sources of human exposure in the short term. Surface water mercury emissions may also be quickly dispersed, but these emissions were relatively small in Texas in 2001: only 25% of Texas counties reported any surface water mercury emissions, and the largest annual release was 16 lbs. Air mercury releases occurred in much larger quantities, with reported releases in most Texas counties, a mean annual release of 288 lbs., and a maximum annual release of 1579 lbs. Although these analyses are limited by the ecological nature of the data and the crude measure of exposure, we believe that the air mercury emissions are the more relevant measure within these limitations. Researchers considering similar analyses should also consider time-lagged comparisons of autism counts with emissions/exposure estimates from previous years to allow enough time for environmental transport, autism development, and diagnosis to have occurred [12].

The three substitution and exclusion approaches happen to provide a rough approximation to the central RR estimate for these particular data, but likelihood-based approaches for handling censored data produce more realistic confidence intervals and have a stronger theoretical basis. All of the previous and current approaches, however, depend on common model assumptions such as Poisson distributed counts, log linearity of the predictors, and adequate control of confounding. The third assumption is of most concern

in the present analysis, given the limited understanding of autism risk factors and a lack of any individual-level data on mercury exposure or confounders.

Ecological studies of autism are not the only studies where censored disease counts are a threat to validity. All special education category counts reported by Texas are administratively censored, and the same practice appears to be followed by some other states. Other count data reported by states may be affected by similar issues. Moreover, efforts to protect privacy under the Health Insurance Portability and Accountability Act of 1996 (P.L.104–191) may lead to administrative censoring of medical surveillance data when counts are extremely low. With the widespread computerization of large data sets, ecological analyses are becoming ever easier to conduct and will likely appear more frequently in the scientific literature despite their limitations.

Censoring also arises in chemical concentration measurement, another common issue in environmental health, through reporting of low concentrations as below the “limit of detection” (LOD). If censored values are excluded, calculated means will generally be biased upwards in this setting. Environmental health researchers have a long tradition of substituting zeros, $LOD/2$, $LOD/\sqrt{2}$, or LOD for these censored values, treating the substituted values as if they were actually observed. The substitution approach is easy to implement but is inferior to formal likelihood-based censored data analysis in that it may also produce biased estimates and always fails to capture the uncertainty associated with measurements below the LOD, such as in calculating confidence intervals [14]. However, the adverse effects of substitution or exclusion may be negligible when both of the following conditions are met: (1) few samples are below the cutoff for censoring and (2) the cutoff for censoring is small relative to most of the measurements.

5. Conclusions

Our analyses indicate that previously reported ecological associations between mercury and autism in Texas are likely to have been overestimated due to inadequate statistical analysis of censored autism counts. Researchers should be aware of the issue of administrative censoring of disease counts and how ad hoc substitution methods can introduce bias and underestimate uncertainty in effect estimates. Bayesian methods offer a potential solution to the problem, do not rely on likelihood approximations, and are not difficult to implement.

Abbreviations

CI: Confidence interval
 LOD: Limit of detection
 RR: Relative risk.

Acknowledgment

Although this research was not funded by any research grants or contracts, the authors would like to acknowledge

the Electric Power Research Institute for funding the original data compilation and previously published analyses.

References

- [1] I. Hertz-Picciotto, L. A. Croen, R. Hansen, C. R. Jones, J. van de Water, and I. N. Pessah, “The CHARGE study: an epidemiologic investigation of genetic and environmental factors contributing to autism,” *Environmental Health Perspectives*, vol. 114, no. 7, pp. 1119–1125, 2006.
- [2] R. F. Palmer, S. Blanchard, Z. Stein, D. Mandell, and C. Miller, “Environmental mercury release, special education rates, and autism disorder: an ecological study of Texas,” *Health and Place*, vol. 12, no. 2, pp. 203–209, 2006.
- [3] R. F. Palmer, S. Blanchard, and R. Wood, “Proximity to point sources of environmental mercury release as a predictor of autism prevalence,” *Health and Place*, vol. 15, no. 1, pp. 18–24, 2009.
- [4] M. Waldman, S. Nicholson, N. Adilov, and J. Williams, “Autism prevalence and precipitation rates in California, Oregon, and Washington counties,” *Archives of Pediatrics and Adolescent Medicine*, vol. 162, no. 11, pp. 1026–1034, 2008.
- [5] R. F. Palmer, T. Walker, D. Mandell, B. Bayles, and C. S. Miller, “Explaining low rates of autism among hispanic schoolchildren in Texas,” *American Journal of Public Health*, vol. 100, no. 2, pp. 270–272, 2010.
- [6] “ISI Web of Knowledge,” February 2010, <http://apps.isiknowledge.com/>.
- [7] D. J. DeNoon, “Mercury in Air Pollution: A Link to Autism?” WebMD Health News, March 2005, <http://www.webmd.com/brain/autism/news/20050318/mercury-air-pollution-link-to-autism>.
- [8] M. Healy, “Link found between autism and rainfall,” *Booster Shots*, Los Angeles Times Blog, November 2008, http://latimesblogs.latimes.com/booster_shots/2008/11/link-found-betw.html
- [9] M. Szklo and F. J. Nieto, *Epidemiology: Beyond the Basics*, Jones and Bartlett, Sudbury, Mass, USA, 2nd edition, 2007.
- [10] Data Accountability Center, “State-Level Data Files,” Office of Special Education Programs, U.S. Department of Education, November 2009, <https://www.ideadata.org/StateLevelFiles.asp>.
- [11] D. Collett, *Modelling Survival Data in Medical Research*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 1999.
- [12] T. A. Lewandowski, S. M. Bartell, J. W. Yager, and L. Levin, “An evaluation of surrogate chemical exposure measures and autism prevalence in Texas,” *Journal of Toxicology and Environmental Health, Part A*, vol. 72, no. 24, pp. 1592–1603, 2009.
- [13] N. E. Breslow and X. Lin, “Bias correction in generalised linear mixed models with a single component of dispersion,” *Biometrika*, vol. 82, no. 1, pp. 81–91, 1995.
- [14] Y. Zhao and H. C. Frey, “Quantification of variability and uncertainty for censored data sets and application to air toxic emission factors,” *Risk Analysis*, vol. 24, no. 4, pp. 1019–1034, 2004.