A COMPARISON OF DATA ANALYSIS OPTIONS WHEN SOME VALUES ARE BELOW THE LIMIT OF DETECTION (LOD)

Gillespie BW[1], Chen Q[1], Lee S-Y[1], Hong B[2], Garabrant D[2], Hedgeman E[2], Sima C[2], Lepkowski J[3], Olsen K[3], and Luksemburg W[4]

[1]Department of Biostatistics, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109; [2]Environmental Health Sciences, University of Michigan School of Public Health, 109 S Observatory, Ann Arbor, MI 48109; Institute for Social Research, 426 Thompson St., University of Michigan, Ann Arbor, MI 48104; [4]Vista Analytical, 1100 Windfield Way, El Dorado Hills, CA 95762

Keywords: Human samples, Blood, Environmental samples, Soil, Modeling

Introduction:

Analyses of chemical concentrations in disciplines such as toxicology and industrial hygiene often have a lower limit of detection (LOD) due to limitations of the measurement methods and instruments. Values below the LOD are often termed "nondetects", or nondetectable values. The analysis of data that includes some values below the LOD cannot employ standard statistical methods without replacing values below the LOD with an estimated value. Data with values below the LOD are technically left-censored, meaning that only an upper bound on the true value is known. We must also consider that estimation of the LOD is itself subject to error.

Common methods of handling values below the LOD include:

(1) setting such values to a constant. Common choices for this constant value are zero, the LOD, the LOD/2, and the LOD/$\sqrt{2}$; and

(2) using statistical methods for left-censored data, with the method of maximum likelihood used to estimate the parameters of the assumed distribution.

When the LOD is small compared with the distribution of the remaining values (i.e., when the proportion below the LOD is small), the method for handling values below the LOD makes little difference. The resulting analyses, whether estimating the distribution or making subgroup comparisons, are robust to the convention employed. When the proportion below the LOD is larger, e.g., greater that 5-10%, the method employed could have a substantial impact on the resulting analysis. This paper investigates the sensitivity of data analyses to the method chosen for handling values below the LOD. We consider both estimation of the concentration distribution, and regression analyses investigating predictors of the concentration. Our investigation considers (1) a range of proportions of the sample below the LOD, (2) two distributions assumed for sample values (lognormal and generalized (3-parameter) gamma). We also consider the variability in the LOD across samples, and error in estimation of the LOD itself.

This investigation was carried out using data from the University of Michigan Dioxin Exposure Study. This study collected samples of serum, house dust, soil, and vegetation from subjects (and their homes and property) living in Michigan, USA in areas potentially exposed to sources of dioxin-like compounds as well as areas presumably exposed only to background levels of these compounds. Chemical analysis was performed for 29 congeners of dioxin, furan, and PCB compounds. For each sample, the concentration was reported in parts per trillion (ppt). The LOD, and an indicator of whether the reading was below LOD, was also reported for each sample.

Methods: The University of Michigan Dioxin Exposure Study (UMDES) was carried out in Midland, Saginaw and parts of Bay Counties (potentially exposed areas) and Jackson and Calhoun Counties (control areas) of Michigan, USA. Contamination occurred when the Dow Chemical Company released into the Tittabawassee River by-products

of the manufacture of defoliants, including Agent Orange, during the Vietnam War era. The river sediment currently has elevated levels of dioxin-like compounds. This contamination, as well as possible contamination from other sources, has moved into the soil of adjacent areas through flooding, and into the food chain. The purpose of UMDES was to estimate the concentration of dioxin-like compounds in human serum for people living in the region of contamination and in control areas, and also to assess concentration levels in house dust, soil and vegetation, to investigate potential pathways of movement into human blood.

The area encompassing Midland/Saginaw/Bay Counties was divided into regions representing the Floodplain, Near-floodplain, Plume (near the former Dow Chemical plant incinerator), and all other areas. The populations in each region were sampled using a two-stage probability household sampling design.[1] Eligible subjects were at least 18 years of age, lived in their current residence for at least 5 years, and provided written informed consent to be administered a detailed exposure questionnaire. In addition, serum samples were collected from subjects who consented and were medically eligible to give blood as defined by the American Red Cross. Subjects who owned their home were eligible to provide house dust samples, and those who owned their property were eligible to provide soil and vegetation samples. Procedures for dust[2] and soil[3] sampling and processing were based on American Society for Testing and Materials (ASTM) methods. Chemical analyses were performed by Vista Analytical Laboratory, Inc. (El Dorado Hills, California, USA) for the World Health Organization designated 29 PCDD, PCDF, and PCB congeners[4] using US Environmental Protection Agency (EPA) methods 8290[5] and 1668.[6]

Statistical Methods

The percent of concentrations below the LOD were calculated for each congener and sample type (serum, dust, soil, and vegetation). The median and range of LOD values for each congener and sample type were also calculated. Of the 29 congeners, a few were selected for examination of the effect of LOD on statistical analyses. The selected congeners were chosen to represent high and low values of the percent below the LOD, and narrow and wide variability in the LOD values.

Five methods for handling values below the LOD were examined: the first four involve setting such values to a constant (zero, the LOD, the LOD/2, and the LOD/$\sqrt{2}$); and the fifth employs statistical methods for left-censored data, using the method of maximum likelihood to estimate the parameters of the assumed distribution.

Parametric and nonparametric estimation of the distribution function with the various "below LOD" conventions were first examined. The distributions were nonparametrically estimated using the Turnbull estimator, which is similar to the Kaplan-Meier estimator but allows left-censored data. The distributions were parametrically estimated assuming, in turn, the lognormal and generalized (3-parameter) gamma distribution. Parametric estimation used the method of maximum likelihood, and was performed using SAS Proc Lifereg.[7]

Parametric and nonparametric modeling of data with values below LOD was then considered. We began with models already developed to explain concentrations of the selected congeners. Each of the models, with covariates fixed, was fit using each of the conventions for values below the LOD. Only parametric models were used (lognormal and generalized gamma). Parameter estimates, standard errors and p-values are presented for each convention.

Results:

Over 1300 subjects were interviewed as part of the UMDES study, over 900 provided serum samples, and over 700 provided dust, soil and vegetation samples. For all samples, the reported LOD was estimated to be correct within 0.1 ppt.

The percent of values below the LOD for each congener are presented in Table 1 for serum, dust, soil and vegetation samples.

Table 1. Percent of sample concentrations below the LOD by congener and sample type.

	Serum (n=xx)	Dust (n=xx)	Soil (n=xx)	Vegetation (n=xx)
Dioxins				
2,3,7,8 TCDD				
…				
Furans				
2,3,7,8 TCDF				
…				
PCBs				
…				

The medians (and ranges) of LOD values for each congener are presented in Table 2 for serum, dust, soil and vegetation samples. These values are based on all LOD values, regardless of whether the sample was above or below the LOD.

Table 2. LOD median (range) by congener and sample type.

	Serum (n=xx)	Dust (n=xx)	Soil (n=xx)	Vegetation (n=xx)
Dioxins				
2,3,7,8 TCDD				
…				
Furans				
2,3,7,8 TCDF				
…				
PCBs				
…				

Variability of the LOD within a congener depended primarily on the sample volume. Although all samples were intended to have the same volume, logistical difficulties such as inability to draw the prescribed amount of blood, or insufficient house dust samples in spite of repeated vacuuming, occasionally occurred. Sample volume was quite consistent for soil samples, but was more variable with dust, serum and vegetation samples. The LOD was also affected by chemical interference.

Distribution function estimates are shown in Figure 1 for each method of handling values below the LOD. Panels 1-6 show nonparametric (Turnbull) estimates and Panels 6-12 show parametric estimates. Each plot shows two examples, one for a congener with few values below the LOD, and one with many values below the LOD.

Regression estimates, with standard errors and p-values, using the five “below LOD” conventions are shown in Table 3.

Table 3. Regression Coefficients (+/- standard error, p=value) sample type.

	Lognormal	**Generalized Gamma**
Congener #1	**	**
<LOD = 0		
<LOD = LOD/2		
<LOD=LOD/√2		
<LOD=LOD		
ML*		
Congener #2		
…		

*ML=Maximum likelihood; ** Values=0 are dropped from the analysis in SAS Proc Lifetest.

Conclusions:

Variability of the LOD between congeners is large. (Variability within congeners?) (Variability between serum, dust, soil, vegetation?) Sensitivity of statistical analyses to values below the LOD depends on both the proportion of values below LOD and the variability in LOD values for those below LOD.

When all LOD values are equal or nearly so, the distribution function estimate between zero and the (mean) LOD is a step function when values below the LOD are replaced by a constant, and a smoothed curve when the distribution has an assumed form and is estimated by maximum likelihood. It is recommended that distribution function estimates be performed using either the Turnbull estimator or by maximum likelihood estimation with an appropriate distribution such as the lognormal. Replacing values below the LOD with a constant, whether zero, LOD/2, LOD/√2, or LOD, provide crude step-function estimates. These conventions are not recommended unless the LOD is quite small. (Sensitivity of estimates to estimation of LOD?) (Sensitivity of regression coefficients, s.e.s, and p-values to method of handling values below LOD.) (Implications for future studies.)

References:

1. Lepkowski J, Olson K, Ward B, Ladronka K, Sinibaldi J, Franzblau A, Adriaens P, Gillespie BW, Chang S-C, Chen Q, Demond A, Gwinn D, Hedgemen E, Knutson K, Lee S-Y, Sima C, Swan S, Towey T, Zwica L, Garabrant D. *Organohalegen Comp* 2006 (forthcoming).
2. American Society for Testing and Materials (ASTM), Standard Practice for Collection of Floor Dust for Chemical Analysis, Designation D 5438-00, Reprinted from the Annual Book of ASTM Standards, Philadelphia, PA.
3. Soil sampling reference?
4. Van den Berg M, Birnbaum L, Bosveld AT, Brunstrom B, Cook P, Feeley M, Giesy JP, Hanberg A, Hasegawa R, Kennedy SW, Kubiak T, Larsen JC, van Leeuwen FX, Liem AK, Nolt C, Peterson RE, Poellinger L, Safe S, Schrenk D, Tillitt D, Tysklind M, Younes M, Waern F, and Zacharewski T. *Environmental Health Perspectives* 1998; 106:775-792.
5. United States Environmental Protection Agency. Method 1668, Revision A: Chlorinated biphenyl congeners in water, soil, sediment, and tissue by HRGC/HRMS. Washington, DC: Office of Water, 1999.
6. United States Environmental Protection Agency. Method 8290: Polychlorinated dibenzodioxins (PCDDs) and polychlorinated dibenzofurans (PCDFs) by high resolution gas chromatography/high resolution mass spectrometry (HRGC/HRMS). Washington, DC: Office of Solid Waste and Emergency Response, 1994.
7. SAS Institute. SAS/STAT User's Guide Version 9. Cary, NC: SAS Institute Inc., 2004.