DEALING WITH DATA BELOW THE LIMIT OF DETECTION IN DIOXIN RESEARCH

Reichert H[1], Gillespie BW[1], Franzblau A[2], Jiang X[2], Hao W[2], Lepkowski J[3], Adriaens P[4], Demond A[4], Garabrant DH[2]

[1]University of Michigan Center for Statistical Consultation and Research, Ann Arbor, Michigan, USA;
[2]University of Michigan School of Public Health, Environmental Health Sciences, Ann Arbor, Michigan, USA;
[3]Univerisity of Michigan Institute for Social Research , Ann Arbor, Michigan, USA;
[4]University of Michigan College of Engineering, Civil and Environmental Engineering, Ann Arbor, Michigan, USA.

Introduction

Handling of values below a limit of detection (LOD) in the analysis of dioxin data has largely been *ad hoc*. Common conventions include replacing values below the LOD with zero, LOD/2 or LOD/sqrt(2). When LODs are very small, these conventions may give reasonable results. However, larger or varying LODs require other methods.

Dioxin data are particularly susceptible to wide variations in LODs for two reasons: First, the sample volumes of materials such as blood, soil and dust may differ. Second, even if sample volumes are uniform, because dioxins occur almost entirely in lipids, samples that include non-lipid material must be adjusted for lipid content. In blood, soil, and dust samples, the lipid content may vary widely. As a consequence, the LOD values may have substantial variation, and subsequent analyses of these data require special attention to the LOD issue.

Reporting of information about LOD values is a first step to facilitate interpretation of analyses. Such information includes the percent below LOD, the median LOD value and the range of LOD values. Data from the University of Michigan Dioxin Exposure Study[1] (UMDES) and the National Health and Nutrition Examination Survey[2] (NHANES) are used for illustration of these ideas.

Materials and methods

Statistical methods for appropriately handling values below detection are available in some statistical software packages, although these methods are not commonly used in dioxin research. Values below detection are termed 'left-censored', and software packages describe the methods as being for left-censored data, generally without mentioning LOD issues. Available methods for left-censored data include plotting the distribution function, performing parametric and nonparametric two-sample (or *k*-sample) tests, and regression analysis. The distribution function can be estimated either nonparametrically (using the Turnbull, or reverse Kaplan-Meier algorithm) or parametrically (assuming a lognormal, Weibull, or other distribution). Two-sample tests can be performed nonparametrically (using the logrank or Wilcoxon test for right-censored data after 'flipping' the data, described below), or parametrically (assuming a lognormal, Weibull, or other distribution). Regression analyses are most often performed using parametric (maximum likelihood based) methods, often assuming a lognormal distribution. The semi-parametric Cox regression model can be used after 'flipping', but interpretation is more difficult than in the right-censored case.

Flipping: Statistical methods for right-censored data are commonly implemented in software. An example of right-censored data is an outcome of time from disease diagnosis to death, when some patients are still alive at the time of analysis. For these patients, the time from diagnosis to date last known alive is a right-censored observation, and is a lower limit on the actual (unknown) time to death. Thus, right-censored data are the mirror image of left-censored data, where LOD values are an upper limit on the actual (unknown) dioxin concentration. To adapt procedures for right-censored data for use with left-censored (below LOD) data, we subtract each value from a number larger than the maximum data value. For example, if the original left-censored data had values of <1, 3, <4, 5, 6, and we subtract each value from 10 (or any number larger than 6), the resulting right-censored

dataset would have corresponding values >9, 7, >6, 5, 4, or re-ordering from smallest to largest, 4, 5, >6, 7, >9. This trick allows right-censored methods to be applied to left-censored data. In two- or *k*-sample tests, this method works well. For computing the distribution function, this method has a flaw in assigning probabilities to values shifted by one observation (e.g., in the data above, the probability for '4' would be assigned to '5'). In a large dataset, this shift will barely be noticeable, and it can be programmatically corrected. For regression analyses, use of maximum likelihood estimation for left-censored data is preferable to flipping because interpretation of regression coefficients on a reverse scale, particularly after log transformation, may not be straightforward.

Software: Reliable software that correctly implements these methods includes JMP (easy to use, moderately expensive), R (steeper learning curve, but free), and SAS (steeper learning curve, and fairly expensive). JMP and R will be considered here, including the R package NADA (Non-detects and Data Analysis), and the R-Excel tool to facilitate using R within Excel.

JMP Software

JMP is a user-friendly (point-and-click interface) statistical package that is known for being visual, interactive, and comprehensive. To plot a distribution function, the Turnbull estimator is found on the menu through Analyze -> Reliability and Survival -> Survival. If desired, click on "Plot Failure instead of Survival" to get F(x) rather than the survival function, S(x). Enter variables, e.g., C1 and C2 as defined below, in the box "Y, Time to Event". Press "OK". The plotting range is exactly the range of the data (3 to 12 in the dataset below), so the "hanging" curve on the left and the jump to 1.0 on the right are obscured by the plot frame. Double click on the x-axis numbers to open a window where a wider range can be specified.

The dataset formatted for **JMP** is given below. Two columns of values are required. When the values in the two columns (C1 and C2 below) are identical, the value is exact (uncensored). A missing value for C1 and the LOD value for C2 denotes a left-censored (below LOD) observation. The dataset below includes two left-censored observations.

Note the gap in the curve at the bottom left of the graph, indicating lack of knowledge about the distribution function shape below the smallest LOD (at 3).

Data in JMP:

C1	C2
.	3
4	4
6	6
8	8
.	10
12	12

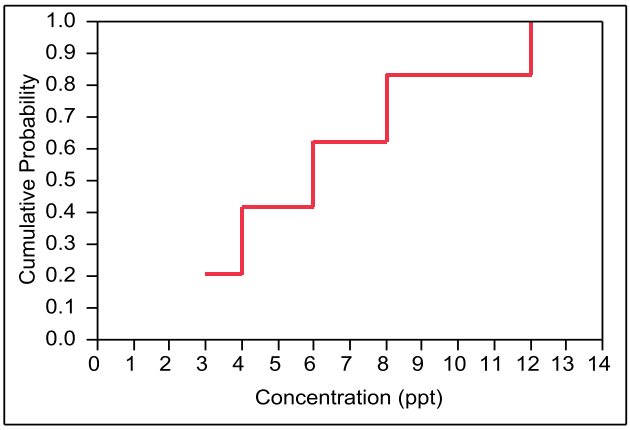

Figure 1. Plotting the distribution function using JMP software: Reverse KM Plot for dataset: 3⁻, 4, 6, 8, 10⁻, 12

R Software, the NADA package

[Adapted from the website http://www.practicalstats.com/nada/nadar/nadar.html] **R** is a free and freely-available implementation of the S-language for statistical computing originally developed by John Chambers and others at Bell Labs. It runs under the Unix, Windows and Macintosh operating systems. It is a powerful environment for statistical computing -- newly-developed methods are frequently released with an R implementation. R contributed packages are developed by volunteers worldwide.

NADA for R is a user-contributed package available from the CRAN(Comprehensive R Archive Network) site.

It implements the procedures found in the textbook Non-detects And Data Analysis by Dennis Helsel (2005)[3]. Methods are included for computing descriptive statistics, hypothesis tests, correlation and regression for left-censored (nondetect) data.

A user's guide to NADA for R is available on the website's NADA downloads page. Through example exercises, the use of the NADA for R package is demonstrated. Datasets used in the exercises come with the NADA for R package, so you can perform each exercise yourself to become familiar with all the NADA for R parametric and nonparametric functions.

The R package has also been adapted for use within Excel, which may be convenient for current Excel users. This implementation is described in the book by Heiberger and Neuwirth (2009)[4]. This implementation of R is also freely available by download from the CRAN website. The availability of free and excellent statistical software may be a huge advantage for researchers with limited funding.

Results and Discussion

In Figure 2, we illustrate the wide distribution of LOD values in data from human serum concentrations in two polychlorinated dibenzofuran congeners with substantial proportions below the LOD in both UMDES and NHANES: 2,3,7,8 TCDF (tetrachlorodibenzofuran; UMDES: n=251, 67% <LOD; NHANES: n=1792, 97% <LOD) and 2,3,4,6,7,8 HxCDF (hexachlorodibenzofuran; UMDES: n=251, 75% <LOD; NHANES: n=1789, 95% <LOD). Estimation was only possible in these cases of >90% below LOD because of the large NHANES sample size, with at least 50 observed values for each congener.

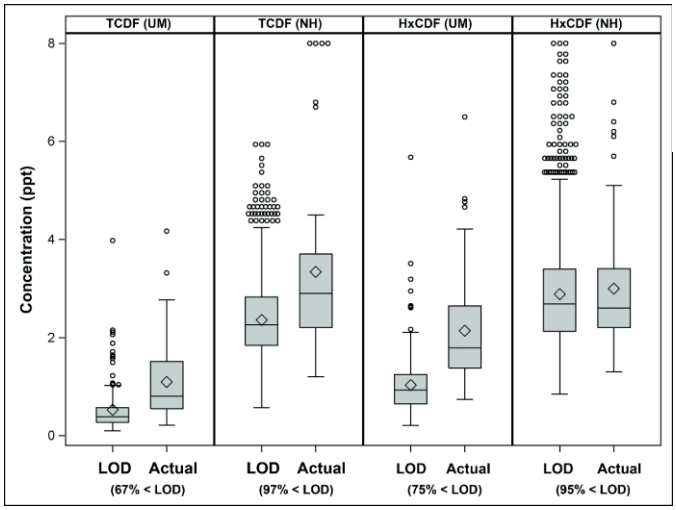

Figure 2. The boxplot distributions of the LOD values and the actual (above LOD) values for two furan congeners (TCDF and HxCDF) in the UMDES

We next illustrate how using the LOD/sqrt(2) can go wrong. An example using UMDES data for OCDD is presented in Figure 3. The original data included no left-censoring. We randomly selected 100 of the 251 observations for illustration. We then randomly left-censored the data by generating a lognormal censoring value ("LOD") for each observation, and retained only the "LOD" when it was larger than the actual value. This process resulted in 54 values below LOD.

Figure 3 (below) presents the estimates of the distribution function based on the complete data and based on the reverse KM estimator. For comparison, the distribution function estimates are also shown for the methods of replacing each left-censored value by either LOD/2 or LOD/$\sqrt{2}$. This comparison shows the potential for severe bias using LOD/2 or LOD/$\sqrt{2}$ when censoring is substantial, particularly when some LOD values are in the upper range of the true distribution. There is no bias for values larger than the highest censored value. Note that the bias will not always be in the direction shown.

The use of 'flipping' and performance of two-sample tests will be straightforward. As in any comparison, plotting of the distribution functions prior to testing is advised, as the logrank and Wilcoxon tests have the most power when the curves have a power relationship or are shifted relative to each other, or in any case do not substantially cross.

For regression modeling, the use of maximum likelihood methods for left-censored data (available in JMP and R) can be performed. As standard procedure, regression assumptions should be verified.

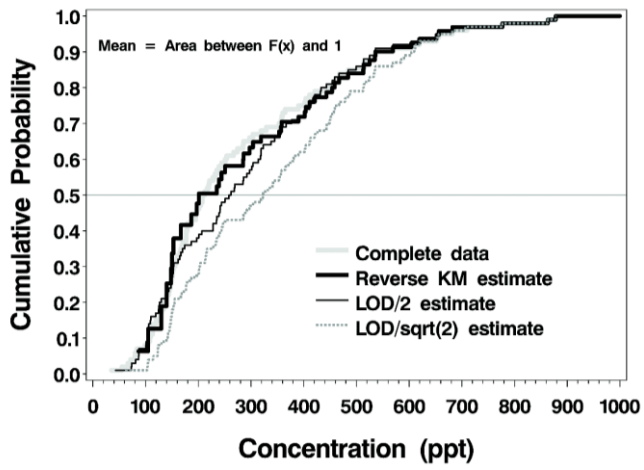

Figure 3. Cumulative distribution function (F(x)) estimates using the reverse Kaplan-Meier (KM) estimator compared with replacing values below LOD with either LOD/2 or LOD/√2.

Conclusions: This abstract has presented an overview of statistical methods appropriate for left-censored data. In particular, two statistical packages, JMP and R, were highlighted as providing excellent analysis options and correct algorithms.

Acknowledgements:

Financial support for this study comes from the Dow Chemical Company through an unrestricted grant to the University of Michigan. The authors acknowledge Drs. L. Birnbaum, R. Hites, P. Boffetta, and M. H. Sweeney for their guidance as members of our Scientific Advisory Board.
Disclosures: The author has no financial ties to either JMP or R software.

References:

1. Hedgeman E, Chen Q, Hong B, Chang C-W, Olson K, LaDronka K, Ward B, Adriaens P, Demond A, Gillespie BW, Lepkowski J, Franzblau A, Garabrant DH. (2009). The University of Michigan Dioxin Exposure Study: Population survey results and serum concentrations for polychlorinated dioxins, furans and biphenyls. *Environ Health Persp*, ;117(5):811-7.
2. Patterson DG, Turner WE, Caudill SP, Needham LL. (2008). Total TEQ reference range (PCDDs, PCDFs, cPCBs, mono-PCBs) for the US population 2001–2002. *Chemosphere* 2008; 73(1,Supp 1):S261-S277.
3. Helsel DR. (2005). *Nondetects and Data Analysis*. John Wiley & Sons, Inc., Hoboken, NJ.
4. Heiberger, RM and Neuwirth E. *R through Excel*. Springer-Verlag, New York (2009).