Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator

Brenda W. Gillespie,[a,b] *Qixuan Chen,*[a,c] *Heidi Reichert,*[b] *Alfred Franzblau,*[d] *Elizabeth Hedgeman,*[d] *James Lepkowski,*[e] *Peter Adriaens,*[f] *Avery Demond,*[f] *William Luksemburg,*[g] *and David H. Garabrant*[d]

Background: Data with some values below a limit of detection (LOD) can be analyzed using methods of survival analysis for left-censored data. The reverse Kaplan-Meier (KM) estimator provides an effective method for estimating the distribution function and thus population percentiles for such data. Although developed in the 1970s and strongly advocated since then, it remains rarely used, partly due to limited software availability.

Methods: In this paper, the reverse KM estimator is described and is illustrated using serum dioxin data from the University of Michigan Dioxin Exposure Study (UMDES) and the National Health and Nutrition Examination Survey (NHANES). Percentile estimates for left-censored data using the reverse KM estimator are compared with replacing values below the LOD with the LOD/2 or LOD/$\sqrt{2}$.

Results: When some LODs are in the upper range of the complete values, and/or the percent censored is high, the different methods can yield quite different percentile estimates. The reverse KM estimator, which is the nonparametric maximum likelihood estimator, is the preferred method. Software options are discussed: The reverse KM can be calculated using software for the KM estimator. The JMP and SAS (SAS Institute, Cary, NC) and Minitab (Minitab, Inc, State College, PA), software packages calculate the reverse KM directly using their Turnbull estimator routines.

Conclusion: The reverse KM estimator is recommended for estimation of the distribution function and population percentiles in preference to commonly used methods such as substituting LOD/2 or LOD/$\sqrt{2}$ for values below the LOD, assuming a known parametric distribution, or using imputation to replace the left-censored values.

(*Epidemiology* 2010;21: S64–S70)

In toxicology and environmental science, estimates of population percentiles for environmental contaminants are often needed where the data include values below an analytical limit of detection (LOD). Data below an LOD are termed left censored in the statistical literature. Nonparametric methods for estimating the distribution function with left-censored data were first published in the early 1970s,[1,2] generalizing the Kaplan-Meier (KM) estimator[3] previously developed for right-censored data. However, in the environmental science literature, population percentiles are often reported using ad hoc methods such as substituting the LOD/2 or the LOD/$\sqrt{2}$ for values below the LOD.[4–6] Parametric estimators such as the lognormal distribution are sometimes used,[7] but nonparametric methods are desirable when population distributions do not follow a standard distribution.

The well-known KM estimator[3] generalized the empirical survival function to the setting of right-censored data. For left-censored data, an analogous estimator is obtained, and the formula mimics that for the KM with the scale reversed.[8] Turnbull[2] generalized the KM estimator to include both left and right censoring, and later to interval censoring.[9] The Turnbull estimator applied to left-censored data is equivalent to the reverse KM estimator.

The goal of this paper is to facilitate broader use of the reverse KM estimator by describing its desirable properties, illustrating its use with real data, and showing how it can be calculated using standard software. Examples are based on serum dioxin concentrations from 2 sources: the University of Michigan Dioxin Exposure Study (UMDES),[10] including only subjects from the control region exposed to background dioxin levels (Jackson and Calhoun counties in Michigan), measured in 2004–2005; and the National Health and Nutrition Examination Survey (NHANES), a population-based sample of the noninstitutionalized in the United States, measured in 2003–2004.[11]

Submitted 1 October 2008; accepted 15 October 2009; posted 8 April 2010.

From the [a]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI; [b]Center for Statistical Consultation and Research, University of Michigan, Ann Arbor, MI; [c]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY; [d]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, MI; [e]Institute for Social Research, University of Michigan, Ann Arbor, MI; [f]Department of Civil and Environmental Engineering, University of Michigan College of Engineering, Ann Arbor, MI; and [g]Vista Analytical Laboratory, El Dorado Hills, CA.

Supported by The Dow Chemical Company through an unrestricted grant to the University of Michigan.

SDC Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Correspondence: Brenda W. Gillespie, Center for Statistical Consultation and Research, 3550 Rackham Building, University of Michigan, Ann Arbor, MI 48109-1070. E-mail: bgillesp@umich.edu.

ISSN: 1044-3983/10/2104-0064
DOI: 10.1097/EDE.0b013e3181ce9f08

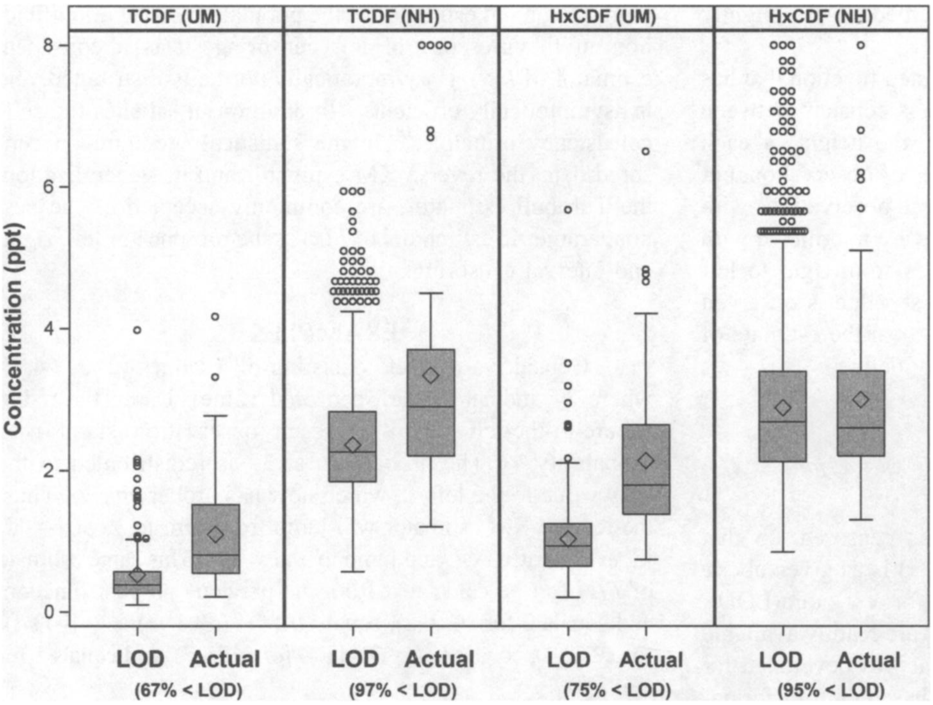

FIGURE 1. Box plots showing the distributions of the LOD (for values <LOD) and the actual value (for values >LOD) for serum concentrations of 2 congeners (2,3,7,8-TCDF, and 2,3,4,6, 7,8-HxCDF). The box covers the 25th–75th percentile; the line across each box represents the median, the diamond represents the mean; whiskers extend to the value within 1.5*interquartile range of each end of the box, and outliers are shown as individual circles beyond the whiskers. ppt = parts per trillion.

THE LIMIT OF DETECTION

Values below an LOD, sometimes called "nondetects," are common in environmental measurements. The LOD itself depends on the precision of the assay, the volume of the experimental sample, and adjustments made such as for blood lipid weight. In this paper, we ignore problems of measurement error (often up to 20%) and problems of estimation of the LOD itself, although these are relevant issues.

The LOD may be a fixed laboratory-specific or batch-specific number, but in some cases the LOD can vary widely from sample to sample, even within the same batch or study. The laboratory may report an LOD for each observation, whether censored or not. Here, however, we only consider the LOD for observations that are left censored.

Human serum concentrations of 2 polychlorinated dibenzofuran congeners (chemicals) with substantial proportions below the LOD in the UMDES and NHANES were chosen to illustrate the reverse KM method: tetrachlorodibenzofuran ([2,3,7,8-TCDF] UMDES: n = 251, 67% <LOD; NHANES: n = 1792, 97% <LOD) and hexachlorodibenzofuran ([2,3,4,6,7,8-HxCDF] UMDES: n = 251, 75% <LOD; NHANES: n = 1789, 95% <LOD). Estimation was only possible in these cases of more than 90% below LOD because of the large NHANES sample size, with at least 50 observed values for each congener. We also artificially censored a UMDES dioxin congener with complete data, octachlorodibenzodioxin ([OCDD] n = 251) for illustration, and we used an NHANES dioxin congener, tetrachlorodibenzodioxin ([2,3,7,8-TCDD] n = 1799, 62% <LOD), to illustrate a check of lognormal fit. All these congeners are considered to be toxic, and regular monitoring of serum levels in the US population is an important role of the NHANES. In the UMDES, congener levels in the control region (presented here) were compared with serum concentrations in an exposed region.[10]

As an exploratory step, the distributions of censored and exact (uncensored) values are shown separately. Figure 1 shows box plot distributions of both the LODs (for left-censored values) and the uncensored values for TCDF and HxCDF in UMDES and NHANES. In these data, the LOD values vary widely and substantially overlap the distributions of the respective uncensored values. More overlap indicates greater advantage to using the reverse KM method for estimating population percentiles, as will become clear later.

THE REVERSE KAPLAN-MEIER ESTIMATOR

The reverse KM estimator for left-censored data estimates the (right-continuous) cumulative distribution function, $F(x) = P(X \leq x)$. The calculation for left-censored data mimics that for the KM estimator.[8,9] Briefly, the calculation uses a product of probabilities of being lower than a given value, conditional on being lower than a slightly higher value. The uncensored values, x_j, are ordered from smallest to largest ($j = 1, \ldots J1$), n_j is the number of values (censored or not) less than or equal to x_j, and d_j is the number of uncensored values equal to x_j. Each factor of the product corresponds to a unique uncensored value, and it is easiest to consider the order of multiplication from the largest value to the smallest. If an uncensored value is equal to an LOD for another observation, the LOD is assumed to be slightly smaller than the uncensored value.

The resulting estimate of $F(x)$ is a step function that has a "jump" at each uncensored value and is constant between uncensored values. With no censoring, the height of each jump is 1/n (where n is the total number of observations) at each unique x_j, or d_j/n if there are multiple observations with the same value, x_j. If left-censored values are intermixed with observed values, the jump size increases from right to left after each censored value. If the smallest value is observed (not censored), so that $x_1 <$ min(LOD), then the estimate of $F(x)$ is zero to the left of x_1, and is fully defined:

$$\hat{F}(x) = \prod_{j:x_j > x} \frac{(n_j - d_j)}{n_j} \quad \textit{for } x < x_J, \textit{ and } \hat{F}(x) = 1 \ \textit{for } x \geq x_J.$$

However, if the smallest value is censored, so that min(LOD) $\leq x_1$, then the estimate of $F(x)$ is as given above for $x \geq$ min(LOD), and is undefined for $x <$ min(LOD). Standard errors of reverse KM estimates are readily available by using the Greenwood formula[2,12] that was developed for the KM estimator. Confidence intervals based on transformations that appropriately restrict the intervals to the range [0–1] are also available.[13] K-sample tests for right-censored data (eg, logrank or Wilcoxon tests) can be used for left-censored data by reversing the scale of the data (subtracting each value from a number larger than the maximum), and applying the test to the now right-censored data.

THE REVERSE KAPLAN-MEIER EXPLAINED AS A REDISTRIBUTION-TO-THE-LEFT ALGORITHM

The reverse KM estimator can be intuitively explained as follows, by analogy to the redistribution-to-the-right algorithm for right-censored data proposed by Efron.[14] Each observation begins with a probability of 1/n. Starting with the largest left-censored value, the associated probability (1/n) is spread equally to all observations (censored or not) to its left. If the largest LOD is larger than all exact values, the redistribution has the same effect as deleting the largest LOD from the data, thus reducing the sample size by 1. Move to the next largest censored value, and distribute its probability (now larger than 1/n) equally to all points to its left. Continue until all probability has been distributed, or the smallest left-censored value retains the remaining probability, which will be the point at which the curve "hangs" at the left end. The idea is that a left-censored value is best estimated by a distribution of its mass to all smaller data values.

ATTRIBUTES OF THE REVERSE KAPLAN-MEIER ESTIMATOR

The statistical properties of the KM estimator have been thoroughly studied, and these also apply to the reverse KM estimator. The reverse KM is the nonparametric maximum likelihood estimator of the population distribution function in the presence of left censoring. It is a consistent estimator of $F(x)$, is asymptotically normally distributed, and is asymptotically efficient.[15] In addition, it satisfies the self-consistency principle.[14] In the statistical literature on censored data, the reverse KM estimator and its generalization, the Turnbull estimator, are commonly accepted as the best nonparametric estimators of $F(x)$ in the presence of left, right, and interval censoring.[15,16]

EXAMPLES

Consider a dataset consisting of 4 values (2, 3^-, 4, 7, where 3^- indicates a left-censored value). Using the redistribute-to-the-left algorithm, each observation starts with probability ¼. The probability at 3^- is redistributed to the only value to the left, 2, which now has probability ½. Thus, the reverse KM estimator will jump from zero to ½ at $x = 2$, jump to ¾ at $x = 4$, and jump to 1 at $x = 7$. This same estimate of $F(x)$ can be calculated from the previous product function, and equals 0 for $x \in [0,2)$, equals 1/2 (= ((3–1)/3) × ((4–1)/4)) for $x \in [2,4)$, equals ¾ (= (4–1)/4) for $x \in [4,7)$, and equals 1 for $x \in [7,\infty)$.

An example using UMDES data for OCDD is presented in Figure 2. 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

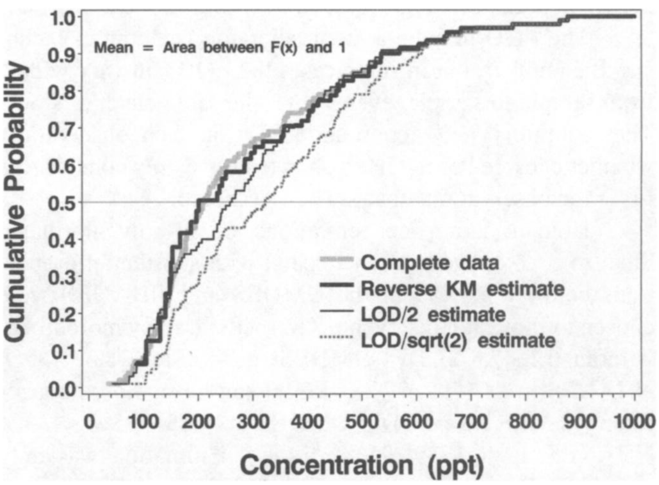

FIGURE 2. Complete data are 100 exact OCDD values from the UMDES (the gold standard). Random censoring was applied, resulting in 54 censored values. The cumulative distribution function ($F(x)$) estimates using the reverse KM estimator is compared with replacing values below LOD with either LOD/2 or LOD/$\sqrt{2}$. Reference line at $F(x) = 0.5$ shows the location of the median for each method as the x-value where the reference line intersects each curve. ppt = parts per trillion.

2 presents the estimates of $F(x)$ based on the complete data and based on the reverse KM estimator. On the right end of the reverse KM, the step size is 0.01 (= 1/100). Moving to the left, the step size increases after each censored value, as the probability is redistributed to the left. Note that the estimate "hangs" at x = 86 ppt, which is the min(LOD), although we know that for $0< x <86$, the estimate of $F(x)$ is bounded above by 0.063 (= F(86)) and bounded below by zero. 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 when using LOD/2 or LOD/$\sqrt{2}$, particularly when some LOD values are in the upper range of the true distribution. There is no bias in $\hat{F}(x)$ for x values larger than the highest censored value. Note that the bias will not always be in the direction shown.

The reverse KM estimates for the 2 example congeners are presented in Figure 3 for the UMDES and the NHANES data. For all 4, the min(LOD) is smaller than x_1, so the reverse KM estimate "hangs" (is undefined) at the left end of each curve. These examples were chosen to illustrate the surprisingly informative estimates despite high percentages below the LOD (eg, 97% for TCDF-NHANES). The UMDES concentrations are similar to the respective NHANES concentrations. In the UMDES and NHANES, the concentrations for HxCDF tend to be higher than those for TCDF.

ESTIMATION OF POPULATION PERCENTILES

Population estimates of the 100*pth percentile can be read from a graph or table as the value of x corresponding to $F(x)$ = p. For example, the median is estimated as the smallest value of x for which the estimate $F(x)$ is ≥0.5. If $F(x)$ = p exactly, corresponding to a step in the step function, then the midpoint of the step is often reported as the estimate of the pth percentile, rather than the smallest value on the step. All percentiles larger than the $F(x)$ corresponding x = min(LOD) can be estimated.

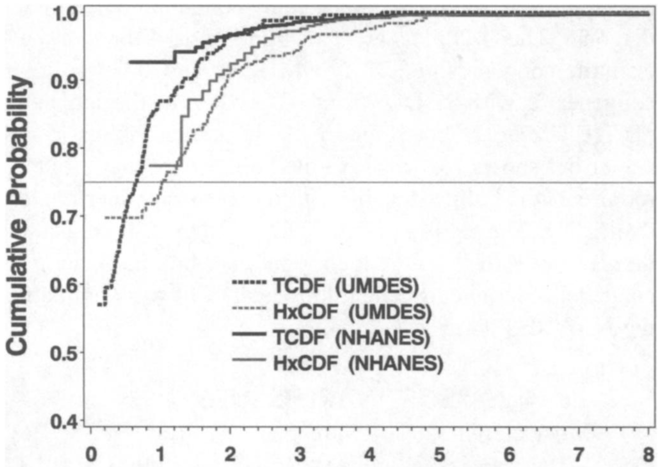

FIGURE 3. Reverse KM estimates of the distribution functions for serum concentrations of 2 furan congeners: 2,3,7,8-TCDF and 2,3,4,6,7,8-HxCDF in UMDES and NHANES. The 75th percentiles are shown by the x-values where the reference line at $F(x)$ = 0.75 intersects each curve (0.64 and 1.02 for TCDF and HxCDF, respectively, in UMDES). ppt = parts per trillion.

In Figure 2, the reference line at $F(x)$ = 0.5 shows that the estimate of the median from the reverse KM can differ from the estimates using LOD/2 or LOD/$\sqrt{2}$. In Figure 3, although the median cannot be precisely estimated for any of the congeners, the estimates are bounded above by min(LOD), which is 0.1 ppt for TCDF (UMDES). We can report, for example, that the median serum concentration of TCDF in UMDES was less than 0.1 ppt. The min(LOD) values, median and 75th percentile estimates for both congeners for UMDES and NHANES are given in the Table. The respective medians estimated using LOD/2 and LOD/$\sqrt{2}$ were uniformly higher than those estimated using the reverse KM (Table).

ESTIMATION OF POPULATION ARITHMETIC AND GEOMETRIC MEANS

The arithmetic mean is estimated as the sum of each x value multiplied by its associated probability, or equivalently, as the area between $F(x)$ and 1.0, for x >0. The latter is intuitive by considering the area divided into horizontal

TABLE. Comparison of Median, 75th Percentile, and Mean Estimates of Serum Concentrations in Parts per Trillion of TCDF and HxCDF by Using the Reverse KM Estimator Compared With Replacing Values Below the LOD With Either LOD/2 or LOD/$\sqrt{2}$

		TCDF		HxCDF	
Estimate	**Method**	**UMDES**	**NHANES**	**UMDES**	**NHANES**
Median	Reverse KM	<0.1	<0.6	<0.2	<0.8
	LOD/2	0.3	1.1	0.6	1.3
	LOD/$\sqrt{2}$	0.4	1.6	0.8	1.9
75th percentile	Reverse KM	0.6	<0.6	1.0	<0.8
	LOD/2	0.7	1.4	1.1	1.8
	LOD/$\sqrt{2}$	0.7	2.0	1.3	2.5
Mean					
Min(LOD)[a]		0.10	0.57	0.21	0.85
F(min(LOD))[a]		0.57	0.93	0.70	0.77
Product[b]		0.06	0.52	0.14	0.66
Lower bound	Reverse KM	0.41	0.18	0.60	0.45
Upper bound	Reverse KM	0.47	0.70	0.74	1.10
	LOD/2	0.54	1.24	0.93	1.53
	LOD/$\sqrt{2}$	0.61	1.72	1.09	2.09

Data are from the UMDES and the NHANES. In these examples, survey weights were not used. Chemical analyses were performed by Vista Analytical Laboratory (El Dorado Hills, California) for UMDES, and by the Agency for Toxic Substances and Disease Registry (ATSDR, Atlanta, Georgia) for NHANES.

[a] Min(LOD) = the smallest LOD, which in these examples is smaller than all uncensored values. F(min(LOD)) is the reverse KM estimate of $F(x)$ at min(LOD).

[b] The product, min(LOD) × F(min(LOD)), is the area of the rectangle enclosing the "undefined region" of the estimate of $F(x)$ and is the difference between the lower and upper bounds of the mean.

rectangular stripes, each with length equal to an x value and width equal to the respective probability (height of the step). The mean is obtained by summing the areas of the rectangles (x_i * $(F(x_i) - F(x_{i-1}))$). If the estimate of $F(x)$ is "hanging" at the left end, then this area is bounded by the lower and upper limits for "completing" the estimate of $F(x)$ in the undefined region. The upper bound for the mean assumes $F(x) = 0$ for $x <$ min(LOD), and the lower bound assumes $F(x) =$ F(min(LOD)) for $x <$ min(LOD). The difference between the lower and upper bounds for the mean is min(LOD) * F(min(LOD)). This value will be small if min(LOD) or F(min(LOD)) is small, in which case presenting a single estimate of the mean is a pragmatic choice. One could use half the area between the lower and upper bounds of $F(x)$, which is consistent with assuming a uniform distribution of values below min(LOD). Alternatively, one could assume a triangular density function in the undefined region, corresponding to a quadratic form for $F(x)$ in this region. However, when the upper and lower bounds for the mean differ substantially, it may be best to present both bounds.

In Figure 2, the lower and upper bounds for the estimated mean were 283.2 and 288.6 ppt, respectively, and differ by 86.4 × 0.063 = 5.4 ppt. Because both the min(LOD) and F(min(LOD)) are fairly small, the undefined region has little impact on the estimate. (The mean of the n = 100 values before censoring was 282.5 ppt; the estimates of the mean when using LOD/2 and LOD/$\sqrt{2}$ were 297.2 and 344.3 ppt, respectively.) Figure 3 illustrates the situation of small min(LOD) but large F(min(LOD)) for TCDF and HxCDF. For TCDF (UMDES), min(LOD) = 0.10, F(min(LOD)) = 0.57, and their product equals 0.057, which is the difference between upper and lower bounds. The lower and upper bounds are 0.41, and 0.41 + 0.06 = 0.47, respectively. In this case, the impact of the left censoring on estimation of the mean will be small, even though F(min(LOD)) is large. The estimate of the mean using LOD/2 is 0.54 and using LOD/$\sqrt{2}$ it is 0.61, both of which are larger than the upper bound of the estimate based on the reverse KM estimator. Table 1 shows these values for both congeners for UMDES and NHANES. For both NHANES congeners, where the percentage below LOD is 95% or greater, the range between the lower and upper bound is substantial (eg, 0.11–0.50 for TCDF NHANES). However, in all cases, the estimates of the mean using either LOD/2 or LOD/$\sqrt{2}$ are larger than the upper bound when using the reverse KM.

The geometric mean (GM) is sometimes used to provide an estimate of central tendency that is not as strongly affected by outliers as the arithmetic mean. It has a useful interpretation as e^{μ} for lognormally distributed data. However, nonparametric calculation of the GM is difficult with values below LOD. Compared with the GM, the median has a simple interpretation for all distributions, is also stable in the presence of outliers, and is easily estimated using the reverse KM method. Consequently, the median may be a better choice than the GM as a general estimate of central tendency.

ASSUMPTIONS

Replacing values below LOD with LOD/2 assumes that these values have a uniform (rectangular) distribution, for which the mean and median are both LOD/2. Use of LOD/$\sqrt{2}$ assumes that values below the LOD have a triangular distribution, for which the median is LOD/$\sqrt{2}$.[3] A triangular region may be reasonable when the LOD is less than the mode of the distribution; however, Figure 1 shows 4 examples in which LOD values are not limited to this range.

Random censoring is assumed for all commonly used parametric and nonparametric censored data methods, as well as for use of LOD/2 or LOD/$\sqrt{2}$. This assumption means that the true values associated with left-censored observations have the same distribution, $F(x)$, conditional on being less than the LOD. In the UMDES and NHANES samples, the LOD was mainly a function of sample volume and blood lipid weight, which was unlikely to be related to the dioxin level, so this assumption seems reasonable.

USING THE REVERSE KAPLAN-MEIER TO ASSESS GOODNESS-OF-FIT TO PARAMETRIC DISTRIBUTIONS

The reverse KM can be used to assess whether the data fit a given parametric distribution. A quantile-quantile (Q-Q) plot is used to check the fit of a parametric distribution by graphing the quantiles of the fitted distribution against the quantiles of the nonparametrically estimated (reverse KM) distribution. A good fit is indicated by points falling along the diagonal. The Q-Q plots are shown in Figure 4 for 1 of the 4 example congeners (TCDF in UMDES), and 1 other dioxin congener, 2,3,7,8-TCDD, from NHANES. In the left panel (TCDF-UMDES), the lognormal fit is reasonable in the center but shows lack of fit in the tails, where the left "tail" represents the bulk of the distribution, as most values are less than LOD. The right panel (TCDD in NHANES) illustrates dramatic lack of fit to the lognormal distribution. Among 29 congeners examined, such striking lack of fit was common in the NHANES data.

SOFTWARE ISSUES

Most statistical software packages (eg, SPSS, IBM Co, Chicago, IL; SAS and JMP, SAS Institute, Cary, NC; Stata, Stata Corp, College Station, TX; Minitab, Minitab, State College, PA; and S-Plus/R [spotfire.tibo.com and www.r-project.org]) include procedures to calculate the KM estimator. One can exploit these KM procedures by reversing the time scale (subtracting each value from a number larger than the maximum, thus turning the left-censored values into right-censored values), computing the KM on the reversed

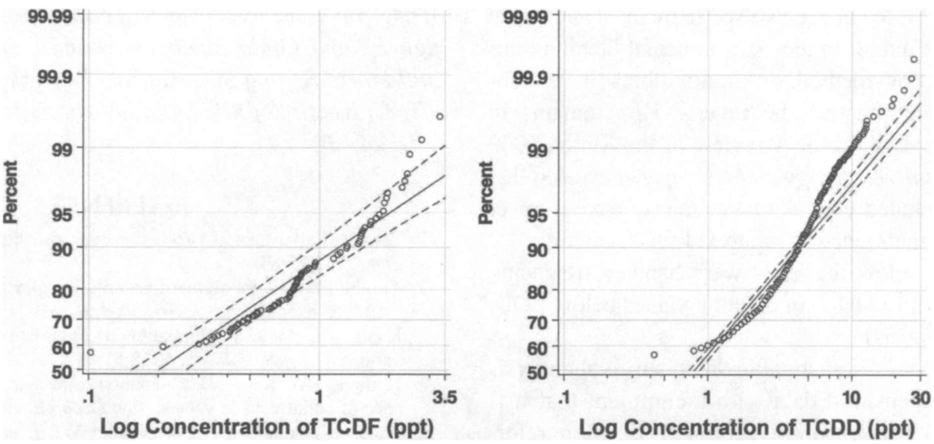

FIGURE 4. Probability plots comparing the cumulative distribution estimates based on the lognormal distribution (x-axis) with the reverse KM nonparametric estimates (y-axis) for serum concentrations of 2 congeners: 2,3,7,8-TCDF from the UMDES, and 2,3,7,8-TCDD from NHANES. Points near the diagonal reference line and between the dashed lines above and below the reference line indicate a good fit to the lognormal distribution. The left panel shows moderately good fit for the central part of the TCDF distribution; the right panel shows dramatic lack of fit to the lognormal distribution for TCDD. Plots were generated using SAS Proc Lifereg. ppt = parts per trillion.

scale, taking care when returning back to the original scale that the "steps" and "jumps" of the step function are in the correct positions. With environmental and chemical data, $F(x)$ is more commonly presented than $S(x)$ (estimated by the KM), and when the scale is reversed, $S(x)$ conveniently becomes $F(x)$. Although the calculation seems simple, there are pitfalls. When we reverse the x-axis, compute the KM estimate, and reverse back to the original scale, the points are associated with the wrong probabilities. The problem is solved by associating each value with the estimated $F(x)$ of the previous observation. This requires programming steps; SAS code to perform the reverse KM is given in the supplementary material (eAppendix, http://links.lww.com/EDE/A369).

For the user, software for the Turnbull estimator is easier to employ than reversing the x-axis and using a KM program. JMP, SAS, and Minitab software have procedures to calculate the Turnbull estimator directly. Example plots from each of these packages are presented in supplementary materials (eAppendix, http://links.lww.com/EDE/A369). The Turnbull procedures in JMP and Minitab software are easy to access using pull-down menus, and both produce plots. The JMP plots are correct. Unfortunately, the Minitab plot (as of this printing) does not appropriately show the "hanging" curve when the smallest value is below LOD and forces the right end to hang, even though the last point is not right censored. SAS calculates the Turnbull estimate in Proc Lifereg and Proc Reliability when a Q-Q plot for a parametric distribution is requested (eAppendix, http://links.lww.com/EDE/A369), but plots require further programming statements. In the S-Plus/R software, no Turnbull function is currently available; one R function published for interval-censored data does not yet work with left-censored or exact data.[17] A new feature that allows the free downloadable R software to be incorporated and used inside of Excel would make an R Turnbull function widely accessible.[19]

If the study design involves survey sampling weights, then the reverse KM estimator should incorporate these weights. Although no known software currently includes Turnbull estimates with survey weights, Stata and SUDAAN (RTI International, Research Triangle Park, NC) software can compute survey-weighted KM estimates, so the reverse KM calculation can be used. For the estimate itself, software for Turnbull or KM without survey weights can be used if each observation is repeated in the data a number of times proportional to its survey weight. However, standard errors produced with this method will not be correct. Although the UMDES and NHANES data were collected using population-based sampling designs, the survey weights were not used for these examples.

Current software does not provide estimates of the population mean in the case of left-censored data, even though giving the upper and lower bounds for an estimate would be a straightforward calculation. The KM software programs provide inconsistently defined estimates of truncated means,[18] which are not easily "reversed" to give means for left-censored data. The Q-Q plots for checking the fit of a parametric distribution using the reverse KM estimator are available in both SAS and Minitab.

DISCUSSION

This paper illustrates the reverse KM method for estimating population distribution functions, means, and percentiles in the presence of left-censored (below LOD) data. In the statistical literature, the reverse KM estimator is considered to be the best nonparametric estimator of the distribution function for left-censored data.[15,16] The reverse KM is the left-censored analog of the KM estimator, which is widely used in medical and social science research. One author in the environmental literature has long recommended this method,[20–22] and others have concurred.[23] However, in practice the reverse KM remains in limited use for below-LOD data. A review of the March 2009 issue of *Environmental Toxicology and Chemistry* revealed that of 13 articles that included data almost certain to have a lower LOD, only 4 articles mentioned or reported LODs, and only 1[24] reported how values below the LOD were handled (replacing values below LOD with LOD/2, or deleting values below LOD, depending on the analysis).

With 1 exception,[13] statistical texts on survival analysis quickly dismiss left-censored data with a comment that it is rarely encountered. The Turnbull estimator is often referenced, but without mention of its use in analyzing data with values below an LOD. Increasing awareness among statisticians of the wide applicability of this estimator would help facilitate its use and spur further software development.

We note that estimating the distribution function is often only the first step in a data analysis, comparable to plotting a histogram with complete data. Although estimating US population percentiles of environmental toxins in blood is an important end in itself of the NHANES study, for many researchers the reverse KM may simply be part of initial data exploration. Two-group tests, such as the nonparametric logrank or Wilcoxon test, are available by analogy with right-censored data methods. Semiparametric modeling can be performed using reversed-scale Cox models, although the interpretation of a reverse hazard function is less intuitive. If parametric modeling will be used, then Q-Q plots, in which reverse KM quantiles are used, can check the parametric fit.

The reverse KM estimator has 2 practical advantages over the methods of replacing values below LOD with either LOD/2 or LOD/$\sqrt{2}$. First, when the distribution of LOD values overlaps the distribution of uncensored values, the reverse KM takes advantage of all available information for estimating the values lower than the LOD through the redistribute-to-the-left algorithm. Second, in the common situation when the smallest value is below LOD, the reverse KM's undefined region in the left tail of the distribution clearly shows the limitations of the estimator. In contrast, this lack of information in the left tail is completely disguised by replacing left-censored values with LOD/2 or LOD/$\sqrt{2}$.

Although environmental data are often approximately lognormally distributed, distributional assumptions can fail, and nonparametric estimators are a safe choice. Given its strong statistical foundation and the availability of software, we recommend that the reverse KM estimator be used routinely in place of either ad hoc methods or parametric assumptions for estimating population distributions, means, and percentiles.

ACKNOWLEDGMENTS

The authors acknowledge Donald Patterson Jr, John A. Gillespie, and John D. Kalbfleisch for helpful comments, Jing Wang for assistance with R, Leslie Niziol for assistance with figures, and Linda Birnbaum, Ronald Hites, Paolo Boffetta, and Marie Haring Sweeney for their guidance as members of The University of Michigan Dioxin Exposure Study Scientific Advisory Board.

REFERENCES

1. Peto R. Experimental survival curves for interval-censored data. *Appl Stat.* 1973;22:86–91.
2. Turnbull BW. Nonparametric estimation of a survivorship function with doubly censored data. *J Am Stat Assoc.* 1974;69:169–173.
3. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53:457–481.
4. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg.* 1990;5:46–51.
5. Pavuk M, Patterson DG Jr, Turner WE, Needham LL, Ketchum NS. Polychlorinated dibenzo-*p*-dioxins (PCDDs), polychlorinated dibenzofurans (PCDFs), and dioxin-like polychlorinated biphenyls (PCBs) in the serum of US Air Force veterans in 2002. *Chemosphere.* 2007;68:62–68.
6. Niskar AS, Needham LL, Rubin C, et al. Serum dioxins, polychlorinated biphenyls, and endometriosis: a case-control study in Atlanta. *Chemosphere.* 2009;74:944–949.
7. Caudill SP, Wong L-Y, Turner WE, Lee R, Henderson A, Patterson DG Jr. Percentile estimation using variable censored data. *Chemosphere.* 2007;68:169–180.
8. Ware JH, DeMets DL. Reanalysis of some baboon descent data. *Biometrics.* 1976;32:459–463.
9. Turnbull B. The empirical distribution function with arbitrarily grouped, censored, and truncated data. *J R Stat Soc B.* 1976;38:290–295.
10. Hedgeman E, Chen Q, Hong B, et al. The University of Michigan Dioxin Exposure Study: population survey results and serum concentrations for polychlorinated dioxins, furans and biphenyls. *Environ Health Perspect.* 2009;117:811–817.
11. Patterson DG, Turner WE, Caudill SP, Needham LL. Total TEQ reference range (PCDDs, PCDFs, cPCBs, mono-PCBs) for the US population 2001–2002. *Chemosphere.* 2008;73(suppl 1):S261–S277.
12. Greenwood M. *The Natural Duration of Cancer. Reports on Public Health and Medical Subjects.* London: Her Majesty's Stationery Office; 1926:1–26.
13. Klein JP, Moeschberger ML. *Survival Analysis, Techniques for Censored and Truncated Data.* 2nd ed. New York: Springer-Verlag; 2003.
14. Efron B. The two sample problem with censored data. In: *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics IV.* Berkeley, CA: University of California Press; 1966:831–853.
15. Gu MG, Zhang C-H. Asymptotic properties of self-consistent estimators based on doubly censored data. *Ann Stat.* 1993;21:611–624.
16. Lindsey JC, Ryan LM. Methods for interval-censored data. *Stat Med.* 1998;17:219–238.
17. Giolo SR. Turnbull's nonparametric estimator for interval-censored data. Technical Report, August 2004. Available at: http://www.est.ufpr.br/rt/suely04a.pdf. Accessed November 10, 2009.
18. Barker C. The mean, median, and confidence intervals of the Kaplan-Meier survival estimate—computations and applications. *Am Stat.* 2009; 63:78–80.
19. Heiberger RM, Neuwirth E. *R through Excel.* New York: Springer-Verlag; 2009.
20. Helsel DR. More than obvious: better methods for interpreting nondetect data. *Environ Sci Technol.* 2005;39:419A–423A.
21. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere.* 2006;65:2434–2439.
22. Helsel DR. *Nondetects and Data Analysis.* Hoboken, NJ: John Wiley & Sons, Inc; 2005.
23. Antweiler RC, Taylor HE. Evaluation of statistical treatments of left-censored environmental data using coincident uncensored datasets: 1. Summary statistics. *Environ Sci Technol.* 2008;42:3732–3738.
24. Ollson CA, Koch I, Smith P, Knopper LD, Hough C, Reimer KJ. Addressing arsenic bioaccessibility in ecological risk assessment: a novel approach to avoid overestimating risk. *Environ Toxicol Chem.* 2009;28:668–675.