Journal of Exposure Science and Environmental Epidemiology (2017) 27, 1–6

www.nature.com/jes

ORIGINAL ARTICLE

Method for analyzing left-censored bioassay data in large cohort studies

Jeri L. Anderson[1] and A. Iulian Apostoaei[2]

In retrospective epidemiological studies of large cohorts of workers exposed to radioactive materials, it is often necessary to analyze large numbers of bioassay data sets containing censored values, or values recorded as less than a detection limit. Censored bioassay data create problems for all bioassay analysis methods, including analytical techniques based on least-squares regression to estimate intakes. A method is presented here that uses a simple empirically-derived equation for imputing replacement values for urine uranium concentration results reported as zero or less than a detection limit, that produces minimal bias in intakes estimated using least-square regression methods with the assumption of lognormally distributed measurement errors.

Journal of Exposure Science and Environmental Epidemiology (2017) **27,** 1–6; doi:10.1038/jes.2015.36; published online 13 May 2015

Keywords: left-censored data; detection limit; censored bioassay data; intake estimation

INTRODUCTION

The National Institute for Occupational Safety and Health (NIOSH) is currently studying mortality in a large pooled cohort of uranium-enrichment workers. As part of this study, organ dose from chronic internal exposure to uranium is being assessed for over 29,000 study subjects from three different gaseous diffusion plants, which includes analyzing over 17,000 individual bioassay data sets containing an average of 20 to 40 urine samples (range 1–664 urine samples). A majority (>80%) of these data sets contain at least one urine uranium result reported as zero or reported as the facility administrative decision limit or instrument detection limit, *L*. These data points are referred to as non-detects or censored data.

The InDEP computer program has been developed for NIOSH to use in the analysis of bioassay data to derive intakes from uranium or plutonium using least-squares regression techniques for large numbers of study subjects simultaneously.[1] Censored bioassay data create problems for most bioassay analysis methods, including analytical techniques based on least-squares regression to estimate intakes with the assumption of lognormally distributed measurement errors.

Several methods are currently used to analyze censored data sets, including ignoring the censored data, setting the censored values equal to zero, or substituting some constant value such as one half the *L* or $L/\sqrt{2}$.[2] However, these methods can cause significant bias when characterizing the data or performing least-squares regression to calculate intakes, especially with data sets with a large fraction of the data censored.[3] Other methods recommended for analyzing censored data sets are maximum-likelihood estimation,[4,5] log-probit regression,[6,7] and the Kaplan–Meier method,[8] but these methods tend to be computationally intense, especially when dealing with thousands of data sets of varying sizes. Several studies have been reported in the scientific literature comparing these various methods.[3,7,9] Although some methods are more robust, there appears to be no one method that is desirable in all situations. The success of the method depends on whether the distribution of the data is well known, the level of censoring, whether the data are censored using only one or multiple *L*, and the objective of the censored data analysis (i.e., descriptive statistics, hypothesis tests, correlation and regression, and multivariate procedures).[3]

The purpose of this paper is to describe a simple empirically derived method for imputing replacement values for urine uranium concentration results reported as zero, as *L*, or "below the *L*," that produces minimal bias when calculating intakes using least-square regression methods with the assumption of lognormally distributed measurement errors. The efficacy of this method was tested by comparing the radionuclide intakes estimated using the proposed method on simulated censored bioassay data with the "true" intakes used to generate the simulated data. The performance of the proposed method was also compared with results obtained with two other commonly used methods for analyzing censored data sets: substitution of non-detects with *L*/2 and a least-squares regression (probability plot) method, also known as regression on order statistics (ROS)[6] or log-probit regression (LPR).[7]

METHODS

Simulated bioassay data sets of uranium radioactivity concentration (in Bq/day) in urine were created with the InDEP computer program using realistic exposure and bioassay monitoring histories of five study subjects, A, B, C, D, and E, from whom 10, 25, 50, 100, and 263 urine samples were collected, respectively. The bioassay data points for each set were projected for a unit chronic intake (1 Bq/day) of a 5-μm activity median aerodynamic diameter (AMAD) natural uranium aerosol with Type F absorption. Each projected data point was assumed to be affected by uncertainties that can be described by a lognormal distribution with a geometric standard deviation (GSD) equal to 3.4. This lognormal distribution accounts for most sources of uncertainty in an individual bioassay sample including analysis uncertainty, day-to-day variability, background

[1]Division of Surveillance, Hazard Evaluations and Field Studies (DSHEFS), National Institute for Occupational Safety and Health (NIOSH), Cincinnati, Ohio, USA and [2]Oak Ridge Center for Risk Analysis, Oak Ridge, Tennessee, USA. Correspondence: Dr. Jeri L. Anderson, 1090 Tusculum Avenue, MS R-14, Cincinnati, OH 45213, USA.
Tel.: +513 841 4510. Fax: +513 841 4470.
E-mail: JLAnderson@cdc.gov
Received 11 December 2014; revised 3 March 2015; accepted 24 March 2015; published online 13 May 2015

interference, and uncertainty in model parameters used to describe distribution and retention in the radionuclides in the body.[10]

We simulated the projected data sets using Monte Carlo methods. For each simulated bioassay data set, 300 iterations were performed with samples for each data point being statistically independent from all other data points in that data set. Projected urine sample results produced by one selected iteration of each simulated data set (i.e., the first iteration was selected at this stage) were multiplied by 2, 4, 8, 16, and 32 to produce a total of six simulated bioassay data sets for assumed chronic intakes of 1, 2, 4, 8, 16, and 32 Bq/day for each of the five study subjects. In each bioassay data set, all bioassay values below a theoretical detection limit of activity L, in urine of 2 Bq/day were set as non-detects. This process resulted in bioassay data sets with different levels of censoring for each study subject. The assumed chronic intakes of 1, 2, 4, 8, 16, and 32 Bq/day are referred to as "true intakes".

All simulated data sets were modified by replacing censored data with one half the L (Method 1), a single value calculated using a new empirically derived method (Method 2), or a set of values calculated using linear regression to fit censored values into a lognormal distribution (Method 3).

Our proposed method (Method 2) relies on the assumption that, for a constant level of chronic exposure to uranium, urine excretion quickly reaches a steady state, with lower urine uranium excretion corresponding to lower levels of intake. Then a function was derived that is dependent on the fraction of samples that were censored in the data set such that greater the fraction of censored samples, lower is the value selected to replace the samples below L:

$$\mathrm{GM} = L(1-f) \tag{1}$$

where GM is the geometric mean of the distribution of samples below L and f is the fraction of samples below L in the data set.

For Method 3, linear least-squares regression was used to simulate replacement values for the censored data by fitting a lognormal distribution to the uncensored data. For each individual bioassay data set, data points were ordered and plotting positions, for a cumulative distribution function of a normal distribution in log space were calculated using the Gringorten formula:[11]

$$p_i = \frac{(i-0.44)}{(n+0.12)} \tag{2}$$

where p is the plotting position, i is the ith ordered bioassay data point, and n is the number of samples in the bioassay data set. The Microsoft Excel (© 2010 Microsoft Corporation) Solver Function (© 2014 Frontline Systems) was then used to estimate parameters a and b, which would minimize the sum of the squared errors (SSE) calculated by:

$$\mathrm{SSE} = \sum_{i}^{n} (y_i + (a + bx_i))^2, \tag{3}$$

where y_i is the inverse of the standard normal distribution for the plotting position $_{pi}$ and x_i is the natural logarithm of the measured bioassay data point, i. Fitted data corresponding to the bioassay data points below the detection limit were then substituted for the censored data.

Intakes were then calculated for all six data sets for each of the five study subjects, A–E, using the InDEP computer program. The intakes were assumed to be due to chronic exposure to a Type F, 5-μm AMAD aerosol of natural uranium. In the calculation of the intakes, the simulated bioassay data were assumed to have an uncertainty described by a lognormal distribution with a GSD of 1.6.[12] This GSD accounts for typical measurement error and day-to-day variability in excretion, and is used when the real bioassay data are used as an input, because the regression methods used to calculate intake in the InDEP program account for the uncertainties in the biokinetic model. Intakes were also calculated from the simulated bioassay data varying the uncertainty in the bioassay samples by setting the GSDs to 2.4 and 3.6 to account for cases with large measurement errors, which may have occurred at some facilities during certain periods of operation. Intakes were not calculated for simulated bioassay data sets where all samples were less than L (i.e., 100% censored). The performance of each of the tested methods was assessed by comparing the error in the calculated intakes relative to the true intake determined with the following relationship:

$$\text{Relative error } (\%) = 100 \times \frac{(\text{Calculated intake} - \text{True intake})}{\text{True intake}} \tag{4}$$

To examine the robustness of the proposed method using repeated testing, 15 additional study subjects, five from each study facility, were selected who had 25 urine sample results in each of their bioassay data sets. This selection is a reflection of the fact that the average number of urine uranium samples for each individual varied from 19 to 37 samples, with 29 (interquartile range 4–37) urine samples being the average per individual study subject across all three facilities. For each study facility, all study subjects with 25 urine samples in their bioassay record were sorted by average value of the urine uranium concentration in their record, and subjects with the five highest average value of uranium concentration were selected. The InDEP computer program was used to simulate urine uranium data for a unit intake of natural uranium (chronic exposure, 5-μm AMAD, Type F absorption) using a GSD of 3.4 and performing 300 iterations for each subject. The first iteration of each study subject's

Table 1. Descriptive statistics for urine uranium bioassay data sets for each study subject.

Subject	*Facility*	*Assumed length of exposure (days)*	*Samples (N)*	*Urine uranium concentration (mg/l)*	
				Average ± SD	*Range*
A	K-25	2685	10	0.003 ± 0.003	0.000–0.009
B	K-25	7608	25	0.003 ± 0.003	0.000–0.008
C	K-25	11,298	50	0.003 ± 0.005	0.000–0.033
D	K-25	4535	100	0.004 ± 0.004	0.000–0.022
E	K-25	14,864	263	0.008 ± 0.050	0.000–0.768
F	K-25	5276	25	0.031 ± 0.067	0.000–0.321
G	K-25	5888	25	0.033 ± 0.073	0.000–0.337
H	K-25	5720	25	0.015 ± 0.056	0.000–0.280
I	K-25	5835	25	0.022 ± 0.045	0.000–0.190
J	K-25	13,675	25	0.038 ± 0.084	0.000–0.382
K	Portsmouth	1709	25	0.006 ± 0.013	0.000–0.060
L	Portsmouth	2347	25	0.016 ± 0.062	0.000–0.300
M	Portsmouth	2048	25	0.006 ± 0.010	0.000–0.030
N	Portsmouth	1176	25	0.008 ± 0.020	0.000–0.080
O	Portsmouth	2769	25	0.018 ± 0.060	0.000–0.300
P	Paducah	2428	25	0.016 ± 0.042	0.000–0.200
Q	Paducah	711	25	0.029 ± 0.108	0.000–0.546
R	Paducah	1016	25	0.014 ± 0.015	0.001–0.076
S	Paducah	2905	25	0.011 ± 0.012	0.000–0.050
T	Paducah	9492	25	0.014 ± 0.027	0.000–0.139

simulated urine data were again scaled using multiples of 2, 4, 8, 16, and 32 to create multiple levels of censoring and bioassay data using Method 2 for censored data. Intakes were calculated for these 15 study subjects using a GSD of 1.6 and the first iteration of the data and calculated intakes were compared with true intakes by calculating relative errors. For three of the study subjects, one from each study facility, intakes were calculated using a GSD of 1.6 and five iterations of simulated bioassay data and results were compared.

RESULTS

Table 1 shows the characteristics of each study subject and the descriptive statistics for each subject's bioassay record. Average urine uranium concentration measurements range from 0.003 to 0.038 mg/l, or 0.12 to 1.52 Bq/day, assuming a urine excretion rate of 1.6 l/day and a specific activity of 25 Bq/mg for natural uranium. The average projected urine excretion rate for each of the 20 study subjects was 0.57 Bq/day (range 0.19 and 1.83 Bq/day) for a chronic exposure to 1 Bq/day of natural uranium. Maximum measured values were as high as 27 Bq/day and projected values were as high as 10 Bq/day.

A demonstration of the application of each of the three methods in calculating substitution values for censored data is given in Table 2. This table shows the bioassay data simulated for subject B assuming a "true" intake of 8 Bq/day and an administrative/detection limit, *L*, of 2 Bq/day. At this level of intake, 44% of the urine data were censored.

The percentage of simulated data censored for study subjects A–E varied from 8% to 100%. Comparison of calculated errors in intakes using both the popular substitution method (Method 1), the proposed imputation method (Method 2), and the least-squares regression method (Method 3; Figure 1) indicate that all the three methods give similar results for data sets with $< \sim 50\%$ of the data censored. Once the level of censoring exceeds 50%, Method 2 appears to perform much better (i.e., lower relative error), with the exception of the study subject A ($n = 10$ bioassay data points), although for this subject, only four data points were available because of 100% censoring at the two lowest intakes. Also, for study subject A, the level of censoring for the remaining data points increases from 50% to 90%, with no data points in between to determine the shape of the curve.

For all levels of censoring for study subjects A–E, average relative error was 88% using Method 1 (minimum to maximum range − 45% to 339%), 8% using Method 2 (range − 85% to 115%), and 56% using Method 3 (range − 27% to 305%). With the exception of study subject A, average relative error decreased for increasing number of samples in a data set for all the three methods. When intakes were calculated using larger GSDs of 2.4 and 3.6, the average relative error between true and calculated intakes for subjects A–E for proposed Method 2 at various levels of censoring remained at 8% (range − 85% to 115%). Also, for study subject A ($n = 10$ bioassay data points), Method 3 appears to perform the best.

Figure 2 shows calculated intakes *versus* true intakes for 15 study subjects (study subjects F–T) with 25 urinalysis data points in each of their records. Generally, average relative errors for this group decreased with decreasing levels of censoring. Average relative error for all levels of censoring was − 9% (range − 71% to − 103%). Again, average relative error at each level of censoring decreased with decreasing levels of censoring, varying from − 51% at censoring $> 90\%$ to 2% at censoring levels of 4–15%.

In Figure 3, the average calculated intakes are compared with true intakes for five study subjects from each study facility; that is, K-25 (study subjects F–J), Portsmouth (study subjects K–O), and Paducah (study subjects P–T). The relative error between the

Table 2. Modification of a data set (subject B) simulated assuming an intake of 8 Bq/day and an *L* of 2 Bq/day.

Sample date	*Simulated bioassay data*	*Censored values eliminated*	*Censored values replaced using*		
			Method 1[a]	*Method 2*[b]	*Method 3*[c]
02/27/1975	12	12	12	12	12
07/21/1975	3.9	3.9	3.9	3.9	3.9
08/18/1975	5.9	5.9	5.9	5.9	5.9
11/11/1975	1.3		1	1.12	1.7
02/11/1976	0.81		1	1.12	1.4
05/03/1976	4.8	4.8	4.8	4.8	4.8
09/03/1976	6.9	6.9	6.9	6.9	6.9
11/11/1976	2.9	2.9	2.9	2.9	2.9
12/20/1976	1.9		1	1.12	2.9
02/07/1977	6.5	6.5	6.5	6.5	6.5
05/03/1977	29	29	29	29	29
08/09/1977	15	15	15	15	15
12/12/1977	15	15	15	15	15
05/08/1978	14	14	14	14	14
08/20/1978	0.72		1	1.12	1.2
12/01/1978	0.36		1	1.12	0.40
03/07/1979	2.0		1	1.12	3.3
06/05/1979	1.8		1	1.12	2.6
09/06/1979	0.45		1	1.12	0.70
12/05/1979	18	18	18	18	18
03/03/1980	0.52		1	1.12	0.95
05/30/1980	18	18	18	18	18
09/06/1980	1.3		1	1.12	2.0
11/28/1980	1.6		1	1.12	2.3
06/27/1990	5.7	5.7	5.7	5.7	5.7

This demonstrates how each method was applied to replace censored data using each of the three methods. [a]Method 1 consists of replacing censored data with $L/2$. [b]Method 2 consists of replacing censored data with a value $GM = L(1\text{-}f)$, where GM is the geometric mean of the distribution of samples below *L* and *f* is the fraction samples below *L* in the data set. For this example (subject B), at an assumed intake of 8 Bq/day, 11 of 25 samples ($f = 0.44$) were below the *L*. [c]Method 3 consists of using linear least-squares regression to fit censored data to a lognormal distribution determined by the uncensored data.

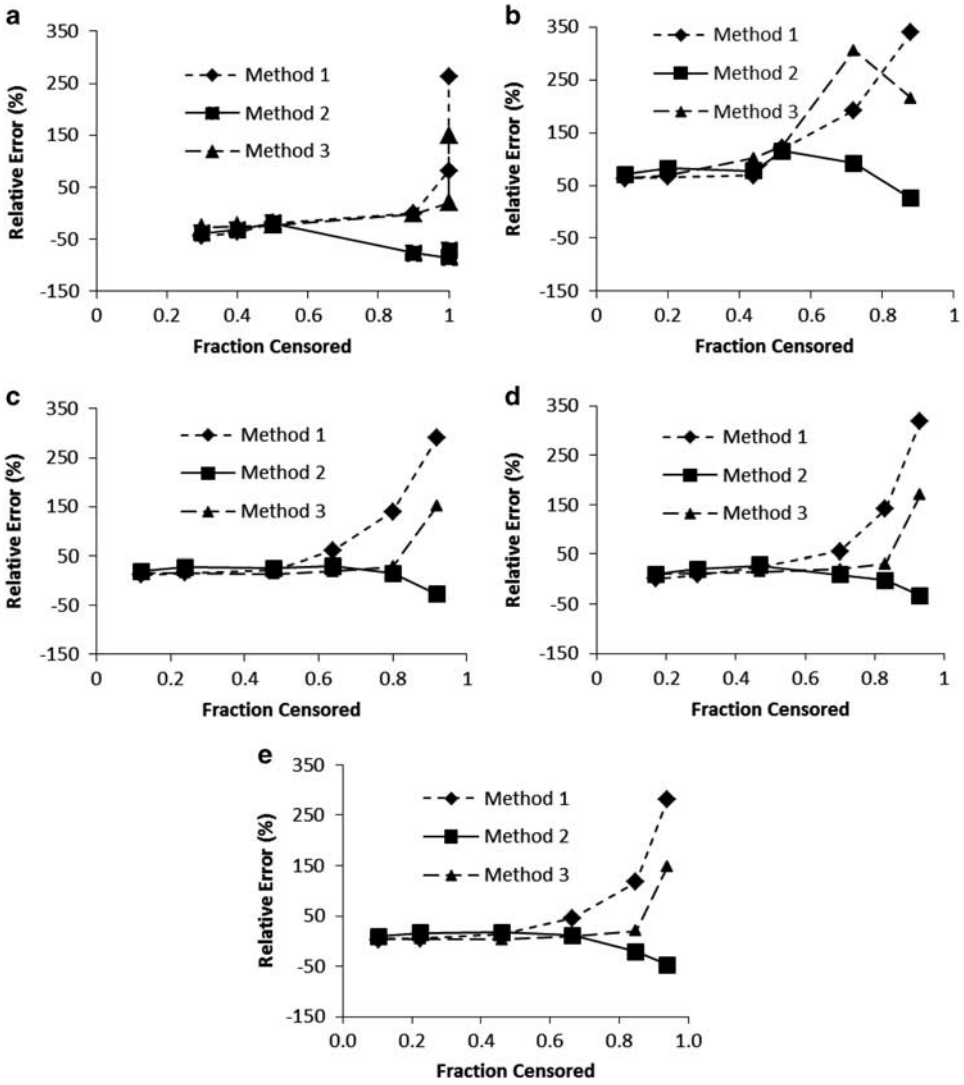

Figure 1. (a–e) Relative error between calculated intakes and true intakes *versus* fraction of urine samples censored in the bioassay data set for subjects A–E, respectively, using each of the two methods for handling censored data sets. Results were obtained assuming a GSD of 1.6 in the calculation in intakes using simulated data.

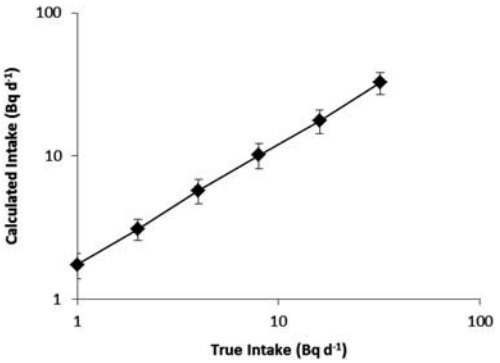

Figure 2. Comparison of true intakes and average intakes calculated using proposed imputation method (Method 2) for 15 study subjects (subjects F–T) for a range of intakes. Lower value intakes correspond to higher levels of data censoring compared with higher value intakes. Error bars represent one standard deviation.

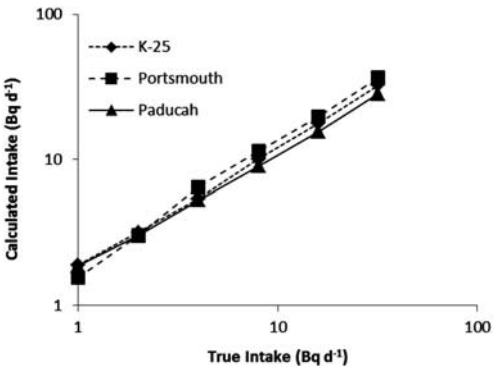

Figure 3. Comparison of calculated average intakes (using proposed Method 2) to true intakes for five study subjects from each study facility. Lower value intakes correspond to higher levels of data censoring compared with higher value intakes.

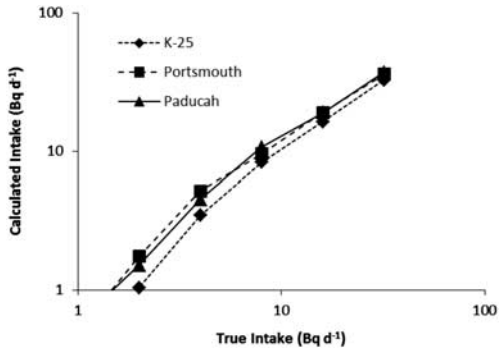

Figure 4. Comparison of calculated average intakes (using proposed Method 2) to true intakes for subjects F, K, and S (one from each facility) using five simulated sets of bioassay data for each study subject. Lower value intakes correspond to higher levels of data censoring compared with higher value intakes.

facility average (over five study subjects) calculated intake and the true intake decreased with decreasing levels of censoring, while the standard deviation (SD) in the average calculated intakes increased in the linear space but stayed the same in log space.

When multiple iterations were analyzed for one study subject from each facility (study subjects F, K, and S from K-25, Portsmouth, and Paducah, respectively), average relative errors between calculated intakes and true intakes for five iterations varied between − 19% (K-25) and 2% (Paducah; Figure 4). For all the three study subjects, the absolute value of the average relative error decreased in decreasing levels of censoring, whereas the SD in the five calculated intakes at each level of censoring increased with decreasing levels of censoring. Censoring varied between 4% and 96%. The overall average relative error of calculated intakes *versus* true intakes was − 4% (range − 61% to 33%).

DISCUSSION

The imputation method proposed here (equation (1)) provides the ease of calculation offered by the substitution method combined with accuracy similar to the maximum-likelihood and least-squares regression methods. Several other empirical equations for calculation of an imputation value for uranium concentrations below a detection limit were examined before deciding on equation (1), including:

$$\mathrm{GM} = \frac{L}{2}\sqrt{1-f}$$

and

$$\mathrm{GM} = L\left(e^{\frac{n}{500}}\right)\sqrt{1-f^2}$$

The last of these two equations defines a geometric mean of the data below a detection limit (L) that is dependent not only on the fraction of data censored (f) and the limit (L), but also on the number of samples in the data set (n).

All empirically derived imputation equations that we have tested relied on the same general assumption that a robust imputation value for uranium concentrations below the detection limit (L) can be determined as a fraction of the limit L, with this fraction being dependent on the fraction of data censored (f). The equations tested were selected in an attempt to improve the performance of the proposed imputation method (Method 2; equation (1)), especially for large fractions (f) of censored data (Figure 1). Our testing indicated that, out of all the imputation equations, Method 2 provides the best results (lowest overall relative errors). Although a few other imputation equations did produce smaller relative errors at large censored fractions (f), they also increase the relative error at lower censored fractions (f), being overall less satisfactory than equation (1) selected for proposed Method 2.

The general consensus is that maximum-likelihood methods or log-probit (least-squares) regression methods produce the least bias when analyzing data sets, particularly when estimating moment statistics such as the mean, median, and interquartile ranges. However, many researchers in the environmental and occupational sciences still utilize substitution methods because of their ease of use. Simulation studies have shown that all methods perform differently depending on the desired outcome or how well the underlying distribution of the data is known,[3] and there appears to be no "perfect" method for all situations.[7] In this study, both the substitution method and the least-squares regression method produce significant bias in censored data sets with relatively small numbers of samples ($n = 25$) and censoring $> 50\%$. A study by Marsh and Birchall[13] showed that maximum-likelihood methods perform well for data sets with censoring $< 60\%$ and samples sizes with $n > 30$.

When conducting exposure assessment in support of epidemiological studies, it is often necessary to analyze bioassay data sets to estimate intakes and organ doses for thousands of study subjects. These data sets often contain a large percentage of censored data, which, if ignored, could result in significant biases in the estimation of intakes of radionuclides of interest. A better treatment of these censored data points is to use maximum-likelihood methods, but these methods can be computationally intense and are often not feasible in large cohort studies. Another commonly used method, the least-squares regression method, is also cumbersome for large numbers of data sets, and does not provide significantly different results (i.e., less bias) than the substitution (Method 1) for data sets that are $> 50\%$ censored. Compared with the commonly used method of substituting censored data with $L/2$ (Method 1) or the least-squares regression method (Method 3), our proposed imputation method (Method 2) provides estimates of intake that are significantly less biased when data sets are censored at levels above 50%. All the three methods appear to work similarly when data sets are $< 50\%$ censored.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DISCLAIMER

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Institute for Occupational Safety and Health.

REFERENCES

1 Anderson JL, Apostoaei AI, Thomas BA. Estimation of internal exposure to uranium with uncertainty from urinalysis data using the InDEP computer code. *Radiat Prot Dosimetry* 2013; **153**: 64–73.
2 Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. *Appl Occup Environ Hyg* 1990; **5**: 46–51.
3 Helsel D. Much ado about next to nothing: incorporating nondetects in science. *Ann Occup Hyg* 2010; **54**: 257–262.
4 Cohen AC. On the solution of estimating equations for truncated and censored samples from normal populations. *Biometrika* 1957; **44**: 225–236.
5 Cohen AC. Tables for maximum likelihood estimates: singly truncated and singly censored samples. *Technometrics* 1961; **3**: 535–541.
6 Shumway RH, Azari RS, Kayhanian M. Statistical approaches to estimating mean water quality concentrations with detection limits. *Environ Sci Technol* 2002; **36**: 3345–3353.
7 Hewett P, Ganser GH. A comparison of several methods for analyzing censored data. *Ann Occup Hyg* 2007; **51**: 611–632.

8 Helsel DR. Less than obvious—statistical treatment of data below the detection limit. *Environ Sci Technol* 1990; **24**: 1766–1774.
9 Gilbert RO, Kinnison RR. Statistical methods for estimating the mean and variance from radionuclide data sets containing negative, unreported or less-than values. *Health Phys* 1981; **40**: 377–390.
10 Boecker B, Hall R, Lawrence J, Ziemer P, Eisele G, Wachholz B *et al.* Current status of bioassay procedures to detect and quantify previous exposures to radioactive materials. *Health Phys* 1991; **60**: 45–100.
11 Yahaya AS, Yee CS, Ramli NA, Ahmad F. Determination of the best probability plotting position for predicting parameters of the Weibull distribution. *Int J Appl Sci Technol* 2012; **2**: 106–111.
12 Castellani CM, Marsh JW, Hurtgen C, Blanchardon E, Berard P, Giussani A *et al.* IDEAS Guidelines (Version 2) for the Estimation of Committed Doses from Incorporation Monitoring Data. European Radiation Dosimetry Group e.V. Braunschweig, 2013 Report No. EURADOS-2013-01.
13 Marsh JW, Birchall A. Estimation of uptake from censored urine excretion data. *Radiat Prot Dosimetry* 1994; **53**: 187–190.