Multiple Imputation Applied to Bioassay Co-Exposure Models

ORAUT-RPRT-0096 Rev. 01
Effective Date: 06/11/2021
Supersedes: Revision 00

Subject Expert(s): Thomas R. LaBone

Document Owner Approval:	Signature on File Thomas R. LaBone, Document Owner	Approval Date:	05/17/2021
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	05/17/2021
Concurrence:	Vickie S. Short Signature on File for Kate Kimpan, Project Director	Concurrence Date:	05/17/2021
Approval:	Signature on File Timothy D. Taulbee, Associate Director for Science	Approval Date:	06/11/2021

FOR DOCUMENTS MARKED AS A TOTAL REWRITE, REVISION, OR PAGE CHANGE, REPLACE THE PRIOR REVISION AND DISCARD / DESTROY ALL COPIES OF THE PRIOR REVISION.

☐ New ☒ Total Rewrite ☐ Revision ☐ Page Change

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
01/24/2019	00	New report initiated to discuss applying the multiple imputation model to bioassay coworker models. Incorporates formal internal and NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.
06/11/2021	01	Revision initiated to incorporate experience gained from applying multiple imputation to recent co-exposure models. Only the new multiple imputation methods are discussed, whereas, Rev. 00 included comparisons of the then current statistical methods for dealing with censored data with the new multiple imputation methods. In addition, Rev. 01 includes statistical methods (like the coworker R package) that were used for the first time in the INL co-exposure statistical analysis. Incorporates formal internal and NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.

TABLE OF CONTENTS

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

CL	censoring level
d	day
DL	decision level
dpm	disintegrations per minute
GM	geometric mean
GSD	geometric standard deviation
IL	imputation level
L	liter
MDA	minimum detectable amount
MI	multiple imputation
NIOSH	National Institute for Occupational Safety and Health
ORAU	Oak Ridge Associated Universities
ORAUT	ORAU Team
PDF	probability density function
q-q	quantile-quantile
SRDB Ref ID	Site Research Database Reference Identification (number)
TWOPOS	time-weighted one person–one statistic

1.0 INTRODUCTION

Bioassay datasets that are used to calculate chronic intake summary statistics for a co-exposure model can contain nonpositive values (values less than or equal to zero) or censored values (values reported as less than a censoring level [CL]). The presence of nonpositive or censored values in a dataset complicates the statistical analysis and application of the data. Multiple imputation (MI) is a Monte Carlo method used to analyze datasets with missing data by replacing them with simulated data [Harel and Zhou 2006]. This versatile technique can also be used to replace nonpositive and censored data with positive (greater than zero) uncensored values below the CL. The positive uncensored datasets can then be analyzed with standard statistical methods for complete data. This report provides four illustrative examples of applying an MI approach based on the work of Krishnamoorthy et al. [2009] and Lubin et al. [2004] to uranium and americium urine bioassay datasets. The example cases are:

- Total uranium activity in urine bioassay data containing censored data, analyzed using an informative lognormal imputation model (Section 3.0);
- Americium-241 in urine bioassay data consisting almost entirely of censored data, analyzed using an uninformative loguniform imputation model (Section 4.0);
- Total uranium activity in urine bioassay data containing nonpositive data, analyzed using a normal-lognormal mixture imputation model (Section 5.0); and
- Americium-241 in urine bioassay data containing nonpositive data analyzed using a lognormal imputation model (Section 6.0).

An imputation model is a probability distribution that describes the distribution of the data (nonpositive or censored) that needs replacement with positive results. Random values are drawn from the imputation model, conditioned on the value being less than the CL for censored data or the imputation level (IL) for nonpositive data. The imputation model can be informative, based on observed positive or uncensored results, or uninformative, based on prior experience with the distribution of similar data.

Perhaps the most difficult part of using MI is developing the imputation model. For censored data, the probability distribution chosen for the imputation model should accurately describe the distribution of the data below the CL and, at the same time, return only positive values. There is no a priori preferred distribution to use for an imputation model, and the best model is suggested by what is known about the subject dataset. However, after experimenting with different EEOICPA datasets and different distributions, two default approaches were developed for use on the Project. The preferred imputation model is derived from fitting a lognormal distribution to the uncensored bioassay data in a given year to estimate the distribution of the censored data in that year. This is referred to as an “informative” imputation model because it uses the uncensored data to estimate the distribution of the censored data. The imputed result is drawn from this lognormal distribution, conditioned on the result being less than the CL. The lognormal distribution was chosen because:

- It is arguable that this statistical distribution is the one most used in health physics to model right-skewed data, which means health physicists on the Project are comfortable working with it; and
- Less familiar, strictly nonnegative distributions, like the Weibull distribution, generate results similar to those obtained with the lognormal.

The second type of imputation model is generated by assuming a distribution for the censored data that is not based on the uncensored data. This is referred to as an "uninformative" imputation model and is used for datasets that consist primarily of censored data.

The concept of bioassay results that are measured to be less than zero is problematic because some stakeholders misinterpret such results as implying that the true result is negative (which is not the case). An approach to solving this problem that is favorable to claimants is to impute positive results for nonpositive results in a way that is analogous to how positive results are imputed for censored results. In this application (where there are no censored data), the imputation model is created by fitting a lognormal model to the positive data and then using it to impute nonpositive results. Imputation models for nonpositive data are always informative because the values for all the data are known.

2.0 PURPOSE

The purpose of this report is to present idealized examples that illustrate how MI can be used to convert datasets containing censored or nonpositive data into datasets that contain only positive uncensored data. This report is not intended to be a procedure that instructs how to perform MI with any given real-world dataset and all its inherent peculiarities. Therefore, specific parameters (i.e., CLs, distributions, number of iterations) used in the provided examples are for illustrative purposes only.

3.0 CENSORED DATA USING INFORMATIVE IMPUTATION

A quantile-quantile (q-q) plot of the daily excretion of total uranium activity in 2,208 urine samples submitted in 1 year by 427 workers is shown in Figure 3-1. This is a complete dataset that is not censored and contains nonpositive results. With a dataset consisting entirely of positive results with no censoring, the co-exposure modeling would proceed as follows:

1. Generate the time-weighted one person–one statistic (TWOPOS) results for each person for each year from the positive bioassay data.

2. Fit lognormal distributions to the TWOPOS data in each year to produce a geometric mean (GM) and geometric standard deviation (GSD) for each year.

The parametric 50th and 84th percentiles from the lognormal distributions in step 2 are the input parameters used to calculate the chronic intake rates. Calculating TWOPOS statistics with censored data is problematic, in the past requiring arbitrary substitutions that produce overly conservative results.[1] In practice, complete datasets were often not reported by the site, which instead reported a censored dataset. The censored dataset resulted from applying one or more CLs to the complete dataset, with the censored results being reported simply as "<CL."[2] For this example, a simulated censored dataset was created by applying a CL of 25 dpm/d to the dataset in Figure 3-1, resulting in the censored dataset in Figure 3-2.

[1] Such as the maximum possible mean, for example.

[2] Where CL could be a number or something like the decision level (DL) or minimum detectable amount (MDA), for example, <25, <DL, and <MDA.

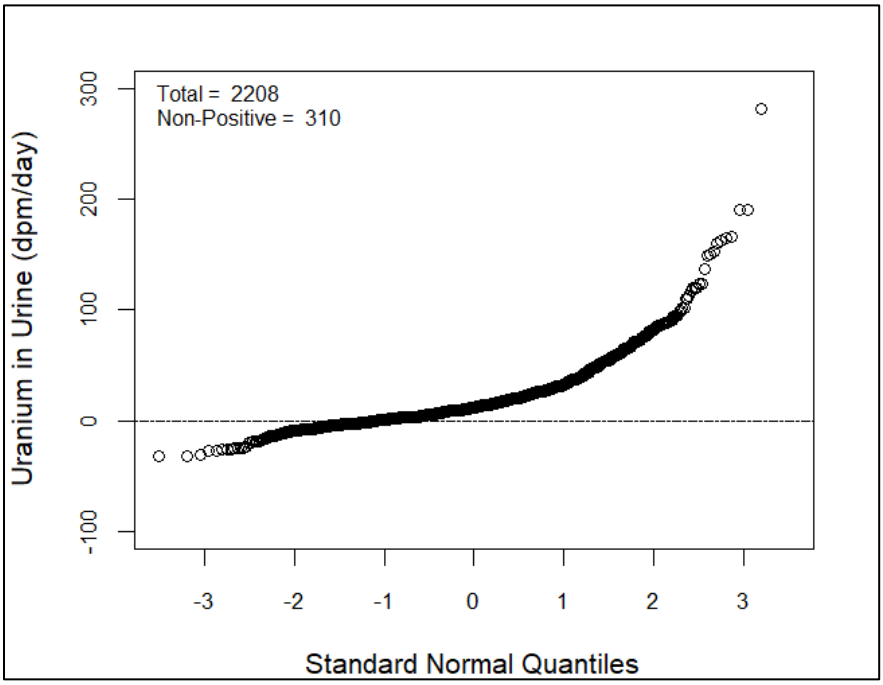

Figure 3-1. Normal q-q plot of 2,208 total uranium urine samples submitted by 427 workers in 1 year.

All the data below the dashed red line were reported as <25 dpm/d. That is, the actual values for the results represented by the grey dots were not known to the statistician analyzing the data.

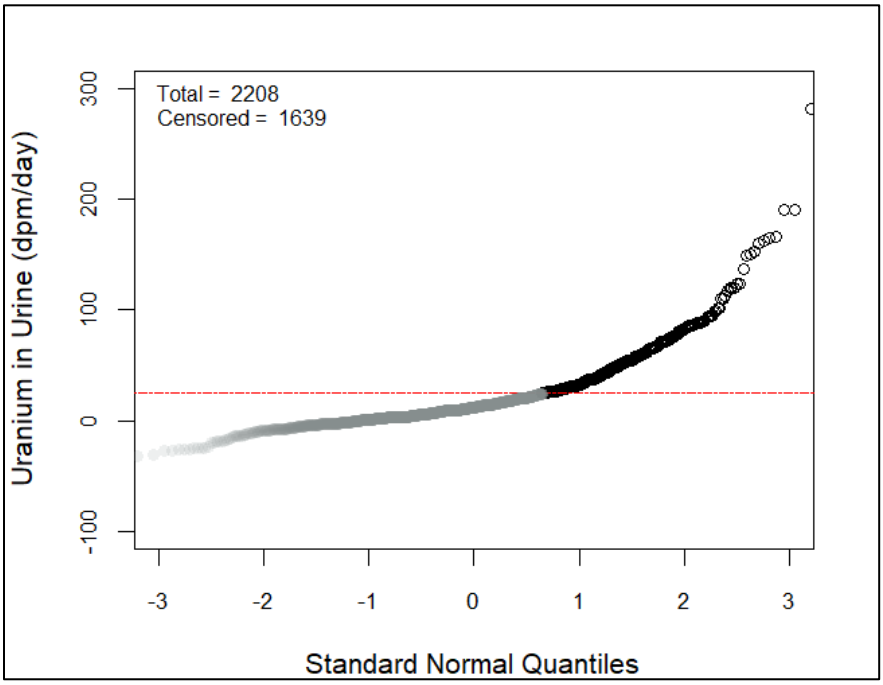

Figure 3-2. Data from Figure 3-1 with CL of 25 dpm/d (red line) applied, generating censored data (grey points).

The key idea behind MI is to first determine the probability distribution that best approximates the distribution of the true values of the censored data, which is called the “imputation model.” Note that the imputation model can be any statistical distribution, but it must have its support on the positive real

line.[3] Next, the censored data were replaced with random draws from the imputation model conditioned on the draw being less than the CL. The dataset now consists of positive results that can be analyzed with methods for complete datasets. Note that MI does not pretend to create information through use of the simulated values but rather to present the data in a way that makes them amenable to analysis with those methods [Rubin 1996].

An example of an informative imputation model is shown in Figure 3-3, where a lognormal distribution is fit to the uncensored uranium data using regression on order statistics. The resulting lognormal imputation model has a GM of 13.403 and a GSD of 2.518. The statistician is most concerned with the imputation model's goodness of fit in the region just above the CL because it is assumed that the distribution of censored results is best approximated by the fit in this region. Therefore, it is not necessary for the imputation model to have good agreement with the higher level values. In any event, this imputation model fits the entire uncensored dataset very well.

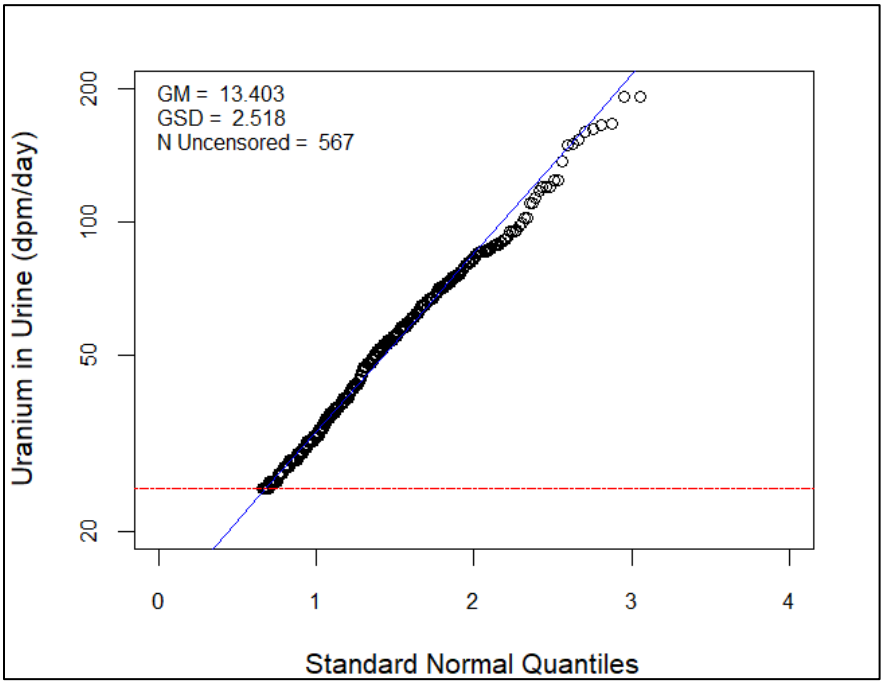

Figure 3-3. Lognormal q-q plot of uncensored data with lognormal quantile line (blue) of lognormal fit to the uncensored data overlaid. Two results greater than 200 dpm/d were excluded from the fit.

The q-q plot in Figure 3-4 shows the lognormal quantile line from the imputation model overlaid on the complete dataset. The imputation model and data agree very well until the data approach zero and take on negative values. The imputation model asymptotically approaches zero in this region, which means that imputed values will overestimate the true values (i.e., be favorable to claimants).

After the censored values in the dataset are imputed, TWOPOS statistics are calculated and a lognormal distribution fit to them. For this example, this was repeated $M = 10$ times and the 10 GMs averaged to give the mean GM of 15.84 dpm/d and the 10 GSDs averaged to give the mean GSD of 1.787.[4] The data from the 10 iterations and the final lognormal fit are shown in Figure 3-5.

[3] Note that this is not a constraint on the imputation model in general but rather one specific to the Project.

[4] The values of the log(*GM*) are averaged and then exponentiated to give the mean GM. The same procedure is followed for the GSD.

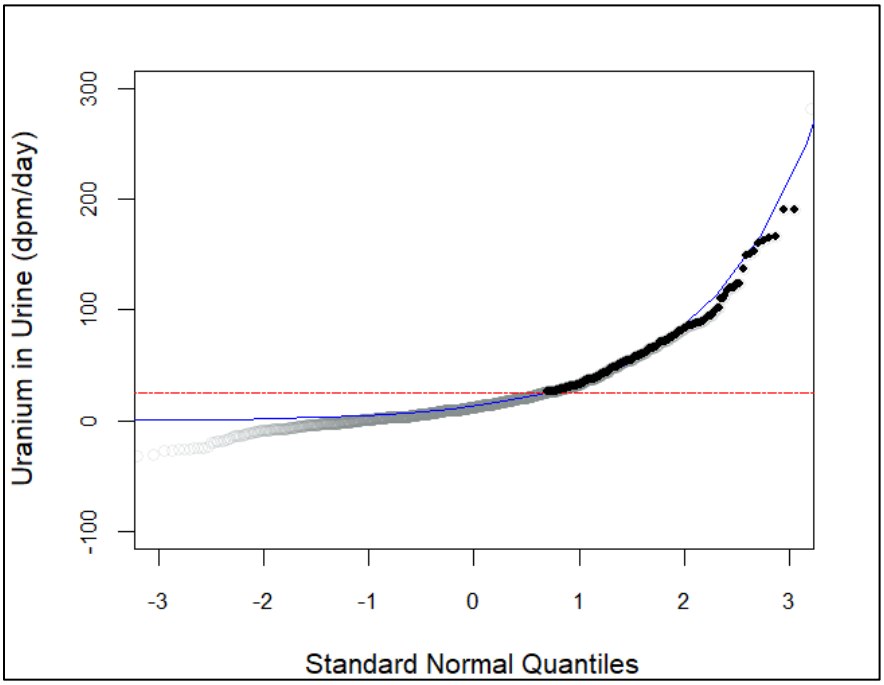

Figure 3-4. Normal q-q plot of all data with lognormal quantile line (blue) of lognormal fit to the uncensored data overlaid.

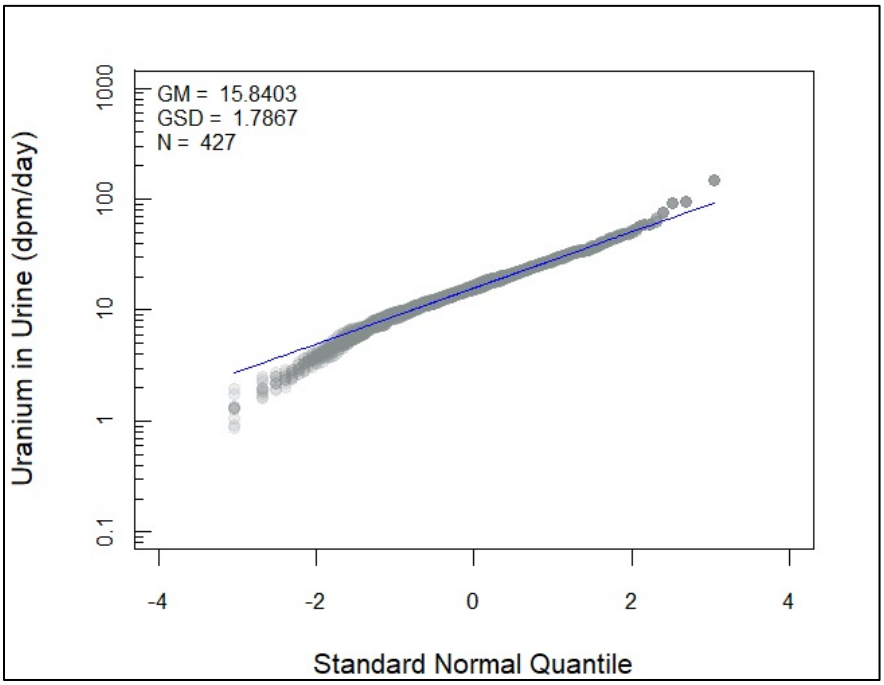

Figure 3-5. Uranium TWOPOS statistics derived from informative imputation for 427 workers generated from M = 10 iterations. The mean GM and mean GSD for the lognormal fit to the statistics are given along with the lognormal quantile line of the fit (blue line).

4.0 CENSORED DATA USING UNINFORMATIVE IMPUTATION

Bioassay datasets consisting primarily (or even entirely) of censored data are not uncommon. Although much is known about such datasets, they are difficult to use to assign a probability distribution to the data. For example, consider the uncensored ^{241}Am data shown in Figure 4-1.

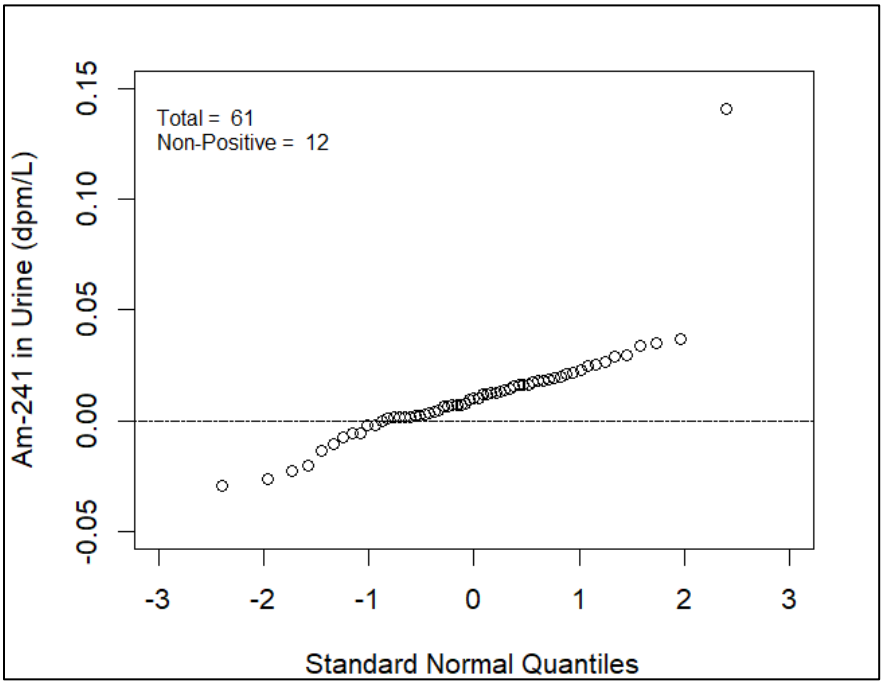

Figure 4-1. Normal q-q plot of 61 ^{241}Am results from urine samples submitted by 60 workers.

Applying a representative CL of 0.04 dpm/L to these data gives the plot in Figure 4-2. Note that the data below the CL (the dashed red line) are not observed (having been replaced by the CL). One way to approach this issue is to impute values for the censored results by drawing a random value from a probability distribution that is bounded at the upper end with a maximum equal to the CL. The loguniform *LU* distribution has this property and has proven to be very useful for this purpose. A random variable *X* is loguniformly distributed if its logarithm is uniformly *U* distributed:

$$\log(X) \sim U(\log(a), \log(b)) \tag{4-1}$$

where

log(*a*) = the lower limit of *U*
log(*b*) = the upper limit of *U*

This is expressed in terms of *X* using the notation:

$$X \sim LU(a,b) = \exp[U(\log(a), \log(b))] \tag{4-2}$$

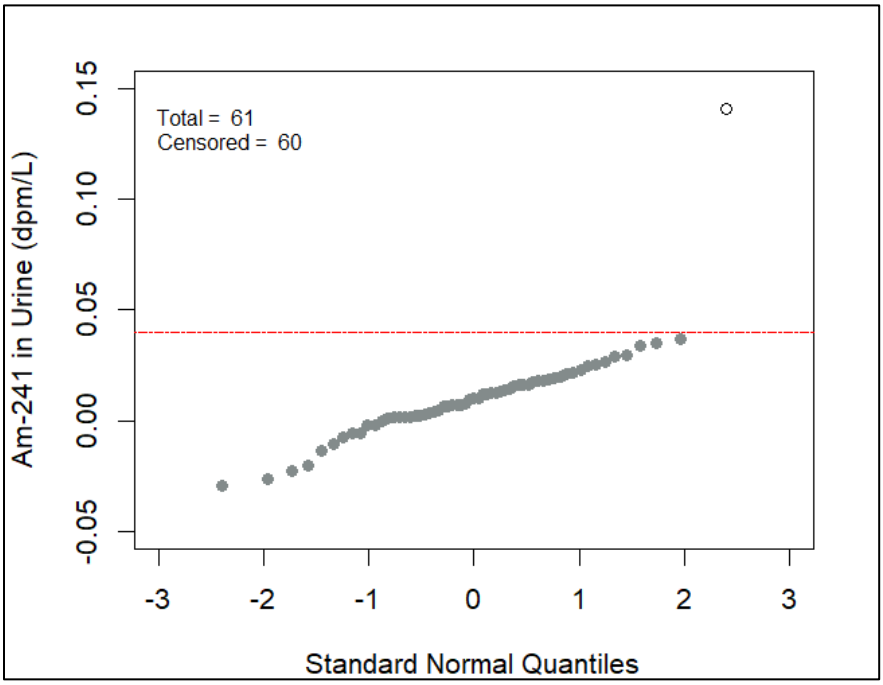

Figure 4-2. Data from Figure 4-1 with CL of 0.04 dpm/L (red line) applied, generating censored data (grey points).

The value of the upper limit *b* should clearly be the CL, but there is no obvious way to select the lower limit *a* based on statistical theory. In practice, the lower limit is taken to be the fraction of the CL that results in a GSD of approximately 3 in the TWOPOS dataset. This GSD was selected to be consistent with the guidance given in Section 2.5.3 in ORAUT-OTIB-0060, *Internal Dose Reconstruction* [Oak Ridge Associated Universities (ORAU) Team (ORAUT) 2018]. Therefore:

$$a = \frac{CL}{K}, \; a > 0 \tag{4-3}$$

where the choice of *K* gives a GSD of approximately 3 in the TWOPOS dataset. Therefore, *K* must be selected iteratively by choosing a value for *K*, examining how it affects the GSD of the TWOPOS data, and then selecting a new value of *K*.

As an example, after several iterations, $K = 50$ was chosen to use in $LU(0.04 \div 50, 0.04)$ to generate the quantile function denoted by the blue line in Figure 4-3, which matches the positive portion of the data very well and gives a GSD of approximately 3.

After imputing the censored data from the loguniform distribution, the TWOPOS statistics are calculated for each person and then a lognormal distribution is fit to the TWOPOS data. This is repeated $M = 10$ times, and the 10 GMs and 10 GSDs are averaged to give the final co-exposure model parameters for the year, as shown in Figure 4-4.

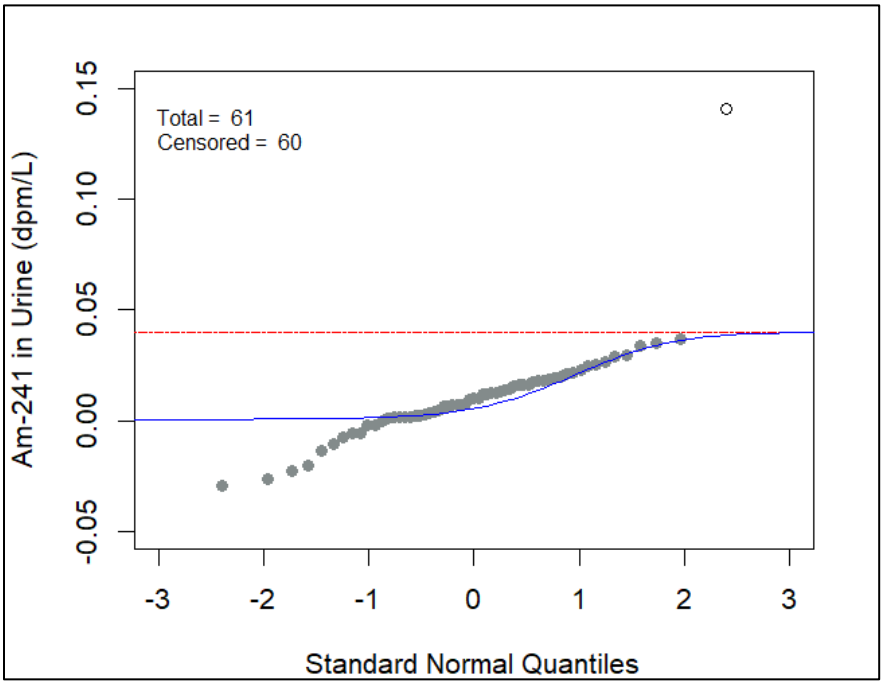

Figure 4-3. Figure 4-2 with loguniform quantile line (blue line) overlaid.

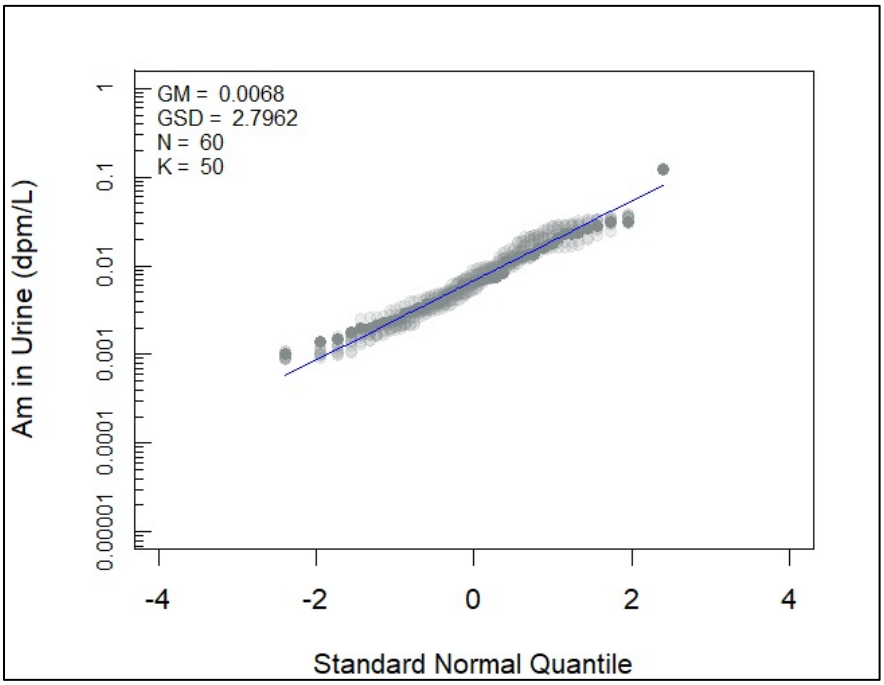

Figure 4-4. Americium TWOPOS statistics derived from uninformative imputation for 60 workers generated from M = 10 iterations. The mean GM and mean GSD for the lognormal fit to the statistics are given along with the lognormal quantile line of the fit (blue line).

5.0 NONPOSITIVE DATA USING MIXTURE MODEL IMPUTATION

Figure 3-1 shows the normal q-q plot for a urine bioassay dataset for total uranium activity that was complete in the sense that none of the data were censored, with nonpositive and positive data being

reported. One way of obtaining the imputation model for the complete uranium dataset is to model it as a mixture of a normal distribution and a lognormal distribution (i.e., a normal-lognormal mixture model). The normal component of the mixture can be viewed as the analytical noise generated when samples containing approximately the same levels of uranium are subtracted from each other, while the lognormal component is viewed as being representative of the real exposures to uranium. The data plotted in Figure 3-1 were fit with a normal-lognormal mixture using an expectation-maximization algorithm from the coworker R library [ORAUT 2020] with the following results:

- A normal mixing fraction of f_n = 0.712 (i.e., the probability of drawing a normally distributed random number from the mixture distribution is 0.712);
- A mean of μ = 8.04 dpm/d for the normal distribution;
- A standard deviation of σ = 9.99 for the normal distribution;
- A lognormal mixing fraction of 1 − f_n = 0.288;
- A logarithmic mean for the lognormal distribution of μ_{log} = 3.58, which equates to a GM of 35.9 dpm/d; and
- A logarithmic standard deviation for the lognormal distribution of σ_{log} = 0.618, which equates to a GSD of 1.86.

The quantile function of this mixture model is superimposed on the uranium data in Figure 5-1, showing excellent agreement with the data.

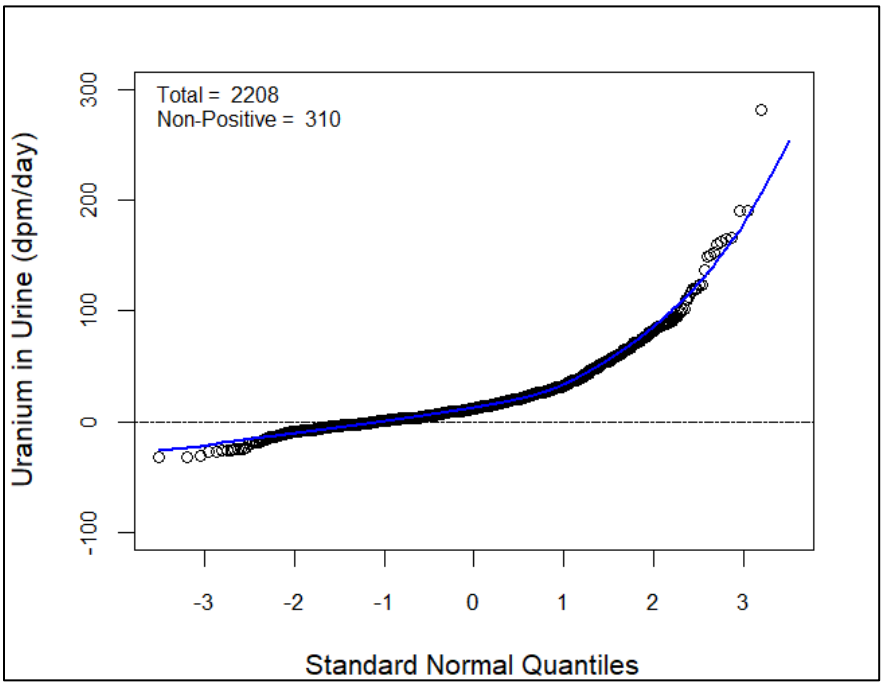

Figure 5-1. Figure 3-1 with the normal-lognormal quantile line (in blue) overlaid.

It is useful to examine the probability density functions (PDFs) of the mixture model and its normal and lognormal components as shown in Figure 5-2. The PDF of the mixture model is simply the pointwise sum of the normal and lognormal PDFs.

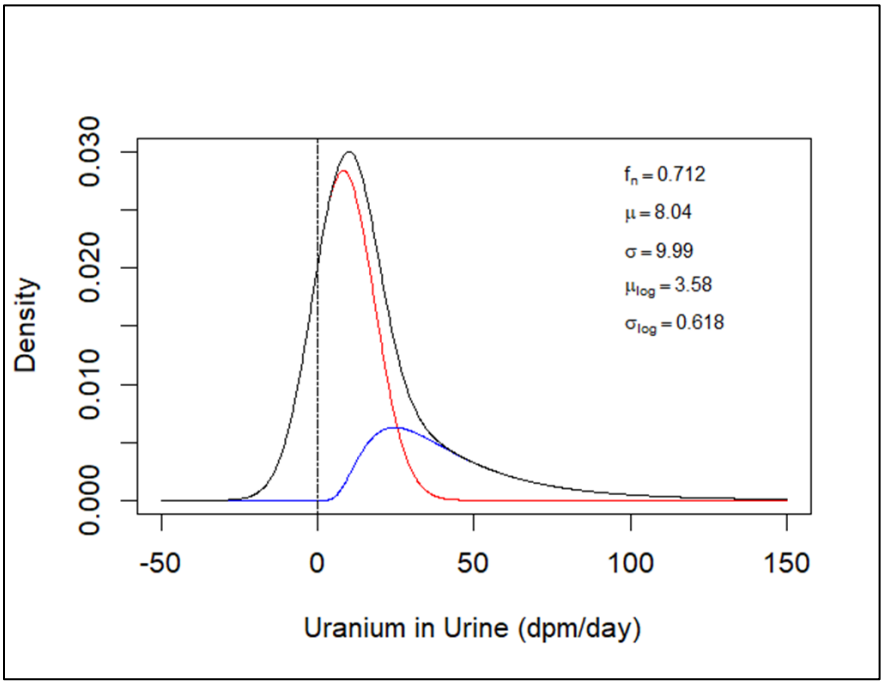

Figure 5-2. PDF plots of the mixture distribution (black line) along with the normal (red line) and lognormal (blue line) components of the mixture.

Once the mixture model is obtained, the normal component is used to define the IL and the lognormal component is used to define the imputation model. The IL is the level below which nonpositive results will be imputed (note that positive results below the IL are not usually imputed). The imputation model is the distribution from which random draws are made, conditioned on the draw being below the IL, with a nonpositive result below the IL being replaced (imputed) with the positive random draw. For co-exposure models, the IL is usually defined empirically as:

$$IL = 2\sigma + \mu \tag{5-1}$$

where μ is the mean and σ is the standard deviation of the normal component. In this case:

$$IL = 2(9.99) + 8.04 = 28.02 \tag{5-2}$$

Therefore, all the nonpositive results less than 28.02 dpm/d are replaced with random draws from a lognormal distribution with a GM of 35.9 dpm/d and a GSD of 1.86, conditioned on being less than 28.02 dpm/d.

After the nonpositive results are imputed, TWOPOS statistics are calculated and the process proceeds in exactly the same manner as for the censored data. The TWOPOS statistics for *M* = 10 iterations and the resulting lognormal fit are shown in Figure 5-3.

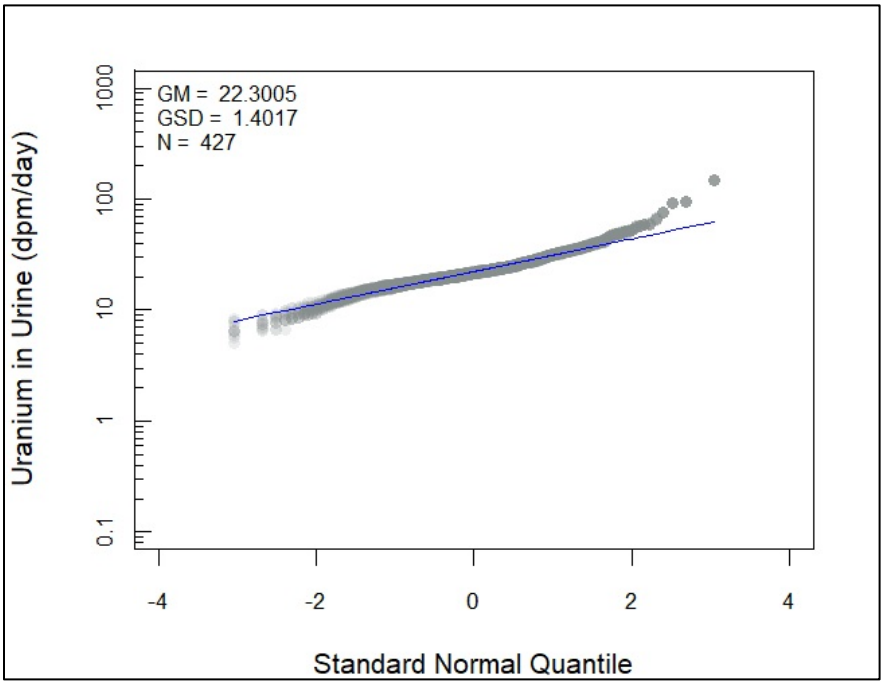

Figure 5-3. Uranium TWOPOS statistics derived from mixture model imputation for 427 workers generated from M = 10 iterations. The mean GM and mean GSD for the lognormal fit to the statistics are given, along with the lognormal quantile line of the fit (blue line).

6.0 NONPOSITIVE DATA USING LOGNORMAL IMPUTATION

The best imputation model for some bioassay datasets with nonpositive values is based on a normal distribution rather than a normal-lognormal mixture (i.e., the data are all noise with no signal). In such cases, the upper tail of the normal distribution is fit with a lognormal distribution, giving a lognormal imputation model. For example, assume the complete ^{241}Am dataset shown in Figure 4-1. As shown in Figure 6-1, there are 49 positive results of which 31 are greater than 0.01 dpm/L (the red line). These were fit with a lognormal distribution (the blue line). The lognormal imputation model has a GM of 0.0113 dpm/L and a GSD of 2.02.

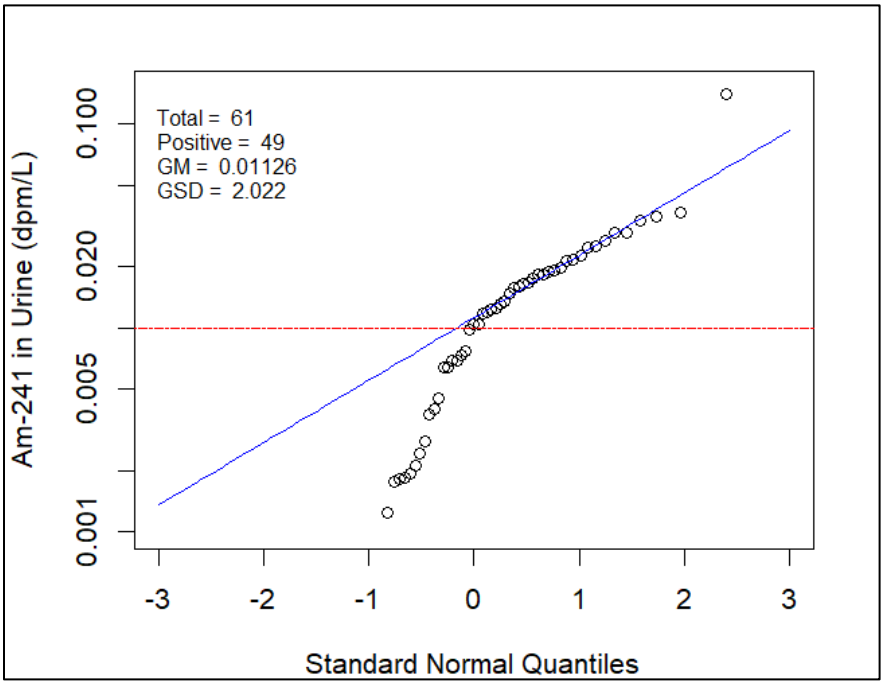

Figure 6-1. Lognormal q-q plot of the 49 positive ^{241}Am results. The 31 results above the red line at 0.01 dpm/L were fit with a lognormal distribution, which has the lognormal quantile line in blue.

This lognormal imputation model is compared with the complete dataset plot in Figure 6-2. The IL is typically an empirically chosen value between zero and the lowest datum used in the lognormal imputation model. Note that the imputed dataset is favorable to claimants regardless of the IL chosen because the normally distributed data (with nonpositive values) have been replaced with lognormally distributed data (with all positive values). The resulting TWOPOS model for 1994 is shown in Figure 6-3.

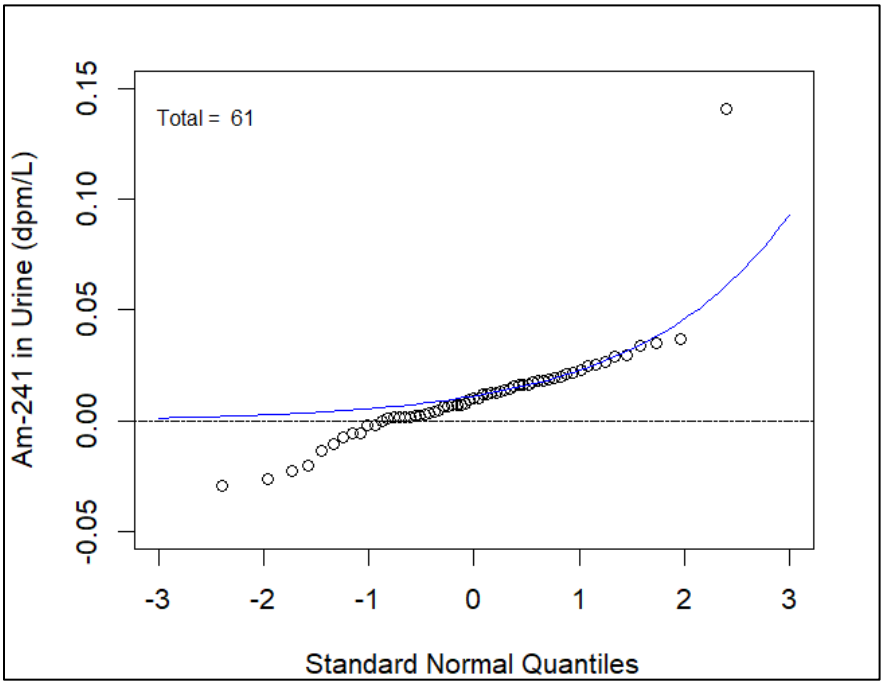

Figure 6-2. Normal q-q plot of the 61 ^{241}Am results. The lognormal quantile line of the fit to the 31 results above 0.01 dpm/L is shown in blue.

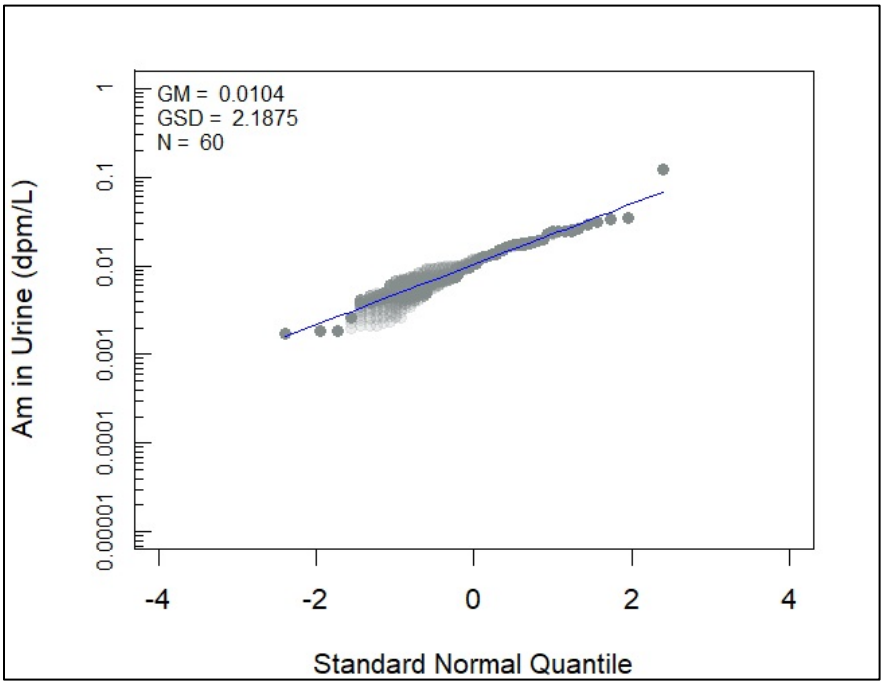

Figure 6-3. Americium TWOPOS statistics derived from log-normal imputation for 60 workers generated from M = 30 iterations. The mean GM and mean GSD for the lognormal fit to the statistics are given, along with the lognormal quantile line of the fit (blue line).

7.0 SUMMARY

This report discusses the use of MI to analyze datasets containing censored or nonpositive data. The key step in MI is the imputation model, which is used to predict what the censored data should look

like or give a plausible overestimate of nonpositive data. The end result is a positive dataset that is analyzed using statistical methods for complete datasets. The typical MI procedure can be broken down into seven steps, the first two of which are different in each example case.

Censored Data Using Informative Imputation

1. Estimate how the bioassay results in each year are distributed based on the uncensored data, giving the imputation models for each year.

2. Replace censored bioassay results below the CLs in all years with random draws from the appropriate imputation models, conditioned on the value being less than the CL.

Censored Data Using Uninformative Imputation

1. Estimate how the censored bioassay results in each year are distributed using a loguniform imputation model having an upper limit equal to the CL and a lower limit equal to the CL divided by a constant $K > 1$, which results in a lognormal fit to the TWOPOS data with a GSD approximately equal to 3.

2. Replace censored bioassay results below the CL in all years with random draws from the appropriate imputation models, conditioned on the value being less than the CL.

Nonpositive Data Using Mixture Imputation

1. Estimate how the bioassay results in each year are distributed using a normal-lognormal mixture model. The imputation models are given by the lognormal components of the mixture and the ILs by the normal components.

2. Replace nonpositive bioassay results below the IL in all years with random draws from the appropriate imputation models, conditioned on the value being less than the IL.

Nonpositive Data Using Lognormal Imputation

1. Estimate how the positive bioassay results in each year are distributed by fitting a lognormal model to the upper tail of the data. The ILs are chosen empirically.

2. Replace nonpositive bioassay results below the IL in all years with random draws from the appropriate imputation models, conditioned on the value being less than the IL.

The remaining steps are identical for all four cases.

3. Generate the TWOPOS results for each person for each year from the now positive bioassay data.

4. Fit lognormal distributions to the TWOPOS data in each year to produce a GM and GSD for each year.

5. Repeat steps 2 to 4 *M* times, which generates *M* GMs and *M* GSDs in each year.

6. The means of the *M* GMs and *M* GSDs in a given year are taken to be the final GM and GSD for that year.

7. Repeat step 6 for each year to produce the summary statistics that are modeled to calculate the intake rate.

In conclusion, MI is not a panacea for nonpositive or censored datasets and, just like all methods, it can be difficult to implement with some datasets. Nevertheless, it can provide:

- Parameter estimates for highly censored and completely censored datasets that are more realistic than those given by previous methods (i.e., binomial fits); and

- Parameter estimates for nonpositive datasets, a problem for which there was previously no accepted solution.

In addition, MI does provide parameter estimates for censored datasets that are consistent with those obtained with previous methods.

Analysis of real-world bioassay datasets is likely to require modifications of the standard MI approaches illustrated in this report. Such modifications will require that the analysis be performed by someone with formal training in statistics.

REFERENCES

Harel O, Zhou X-H [2006]. Multiple imputation - review of theory, implementation and software. Seattle, WA: University of Washington. September 7. [SRDB Ref ID: 183969]

Krishnamoorthy K, Mallick A, Mathew T [2009]. Model-based imputation approach for data analysis in the presence of non-detects. Ann Occup Hyg *53*(3):249–263. [SRDB Ref ID: 146843]

Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P [2004]. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect *112*(17):1691–1696. [SRDB Ref ID: 146841]

ORAUT [2018]. Internal dose reconstruction. Oak Ridge, TN: Oak Ridge Associated Universities Team. ORAUT-OTIB-0060 Rev. 02, April 20. [SRDB Ref ID: 171554]

ORAUT [2020]. Functions used in internal dose co-exposure modeling. Oak Ridge, TN: Oak Ridge Associated Universities Team. [SRDB Ref ID: 184010]

Rubin DB [1996]. Multiple imputation after 18+ years. J Am Stat Assoc *91*(434):473–489. [SRDB Ref ID: 184515]