Memorandum

To: Savannah River Site and SEC Issues Work Group
From: SC&A, Inc.
Date: June 3, 2020
Subject: Review of Multiple Imputation Methods Applied to Censored Bioassay Datasets

Introduction and Background

A primary issue in the formulation of co-exposure models concerns the treatment of monitoring records that have been censored at a set threshold level, such as the minimum detectable activity (MDA), reporting level or some other predetermined value. Of particular concern are datasets in which large portions of the available monitoring data are censored at a given threshold (in some cases greater than 90 percent). Because co-exposure models rely on fitting the available data to statistical distributions, datasets with a large portion of censored data pose a particular analytical challenge.

To address this issue, the National Institute for Occupational Safety and Health (NIOSH) produced a methodology based on multiple imputation to infer (or "impute") numerical values below the censoring limit. The basic concept assumes that the dataset follows a given distribution and that the uncensored values (i.e., the real numerical values reported at values above the censoring limit) follow the upper tail of that distribution.[1] Once the upper tail of the lognormal distribution has been established, the censored bioassay results can be inferred with the assumption that they follow the lognormal parameters established by the uncensored results. Random variability of the specific censored bioassay result is taken into account by imputing the numerical result multiple times and then averaging the result for the final numerical result used in co-exposure modeling. The specific mechanisms and calculation steps of this methodology are described in the NIOSH report, ORAUT-RPRT-0096, revision 00, "Multiple Imputation Applied to Bioassay Coworker Models" (NIOSH, 2019a; "RPRT-0096").

This methodology was first used in formulating co-exposure models for the Savannah River Site (SRS), as outlined in ORAUT-OTIB-0081, revision 04, "Internal Coworker Dosimetry Data for the Savannah River Site" (NIOSH, 2019b). SC&A reviewed this co-exposure model and released its report in September 2019 (SC&A, 2019). After discussion at the December 5, 2019, joint meeting of the Savanna River Site and SEC Issues Work Groups (SRS & SEC WG, 2019), SC&A issued revision 1 of this report on March 13, 2020 (SC&A, 2020). As part of that review, SC&A performed a preliminary review of imputation methods with a narrow focus on their

[1] In the case of bioassay data, the lognormal distribution is typically assumed.

application and results related to SRS co-exposure models.[2] During discussions with the SRS and SEC work groups and NIOSH, SC&A was tasked with specifically reviewing the technical aspects of multiple imputation in the broader context of its general use under the Energy Employees Occupational Illness Compensation Program Act (EEOICPA).

This memorandum summarizes SC&A's review of imputation methods applied to bioassay data as presented in RPRT-0096 (NIOSH, 2019a). The next section discusses the technical aspects of multiple imputation and its general use in inferring values when large portions of a dataset are censored. Following that section, SC&A discusses the practical significance of using multiple imputation in co-exposure modeling versus other dose reconstruction methods (i.e., missed dose assignment). SC&A's summary conclusions about multiple imputation are found in the final section of this memorandum.

Technical Evaluation and Discussion of Imputation Methods

In discussing incorporating nondetects in science, Helsel (2009) references a U.S. Geological Survey report by Miesch (1967) and concludes the following:

1. "In general, do not use substitution. . . . Substitution is NOT imputation, which implies using a model such as the relationship with a correlated variable to impute (estimate) values" (Helsel, 2009, p. 261).[3]
2. "Method evaluations for estimating a mean do not necessarily carry over to the more difficult issues of how to compute interval estimates, upper percentiles, a correlation coefficient, a regression slope and intercept" (Helsel, 2009, p. 261).
3. Right-censored data from the survival/reliability analysis can often be adapted and used for left-censored data as well.
4. "Commercial software should more easily incorporate left-censored data into its survival/reliability routines" (Helsel, 2009, p. 261).

Helsel (2020) further elaborates that substitution methods (such as replacing censored data with half the detection limit) creates problems associated with "invasive data," such as the artificial lowering of the standard deviation and resulting confidence intervals. In addition, substitution of censored values at a preset level such as MDA/2 may create artificial trends in the data that do not actually exist. This is especially problematic for datasets in which multiple censoring levels are present, such as bioassay measurements where the limits of detection improved over time.

Ideally, labs should report the data point for the measurement with an associated uncertainty in accordance with the International Organization for Standardization (ISO)/International Electrotechnical Commission (IEC) Guide to the Expression of Uncertainty and Measurement (ISO/IEC, 2008). The gain is that unbiased point estimates of means and standard deviation can be made in the usual way from the sample data. The bad news is that some of the data will be

[2] For a discussion of the imputation review findings in SC&A (2020), refer to the section "Implications for Dose Reconstruction and Co-Exposure Modeling" in this memorandum.

[3] The substitution method (such as using the lower limit of detection divided by two (LLD/2)) just assumes the data below the LLD are uniformly distributed between zero and the LLD. The interesting factor here is that the substitution method can be applied to an individual data point, based on the LLD for that measurement.

negative if there is a background subtraction necessary; i.e., if there is no real concentration, roughly half of the estimates will be negative, so the mean will be about zero.

Refer to Helsel (2012) for most recent advances in software and consult the Comprehensive R Archive Network (CRAN) package, Nondetects and Data Analysis for Environmental Data (NADA), which contains many relevant R-code functions (https://cran.r-project.org/).

RPRT-0096 further discusses the imputation method and concludes that a lognormal distribution is most appropriate. A common argument is that if the data are the result of the product of a large number of independent random variables, each of which is positive (but not necessarily lognormal), then the lognormal can be the result of considering the central limit theorem in the log domain.

Multiple imputation uses the information in the "detected data" to generate values below the detection limit by assuming they come from a common lognormal distribution. So, if correct, it uses more of the information in the data set (and therefore, from a statistical point of view, should be better).[4]

One issue that does seem to arise is what level of censoring is too much, and opinions vary. It is not entirely clear that there is such a point. A dataset with 50 percent censored data still allows the simple method of using the median and the 84th percentile to estimate the geometric mean and geometric standard deviation. Krishnamoorthy et al. (2009) suggest that the performance of the imputation approach may depend how many data points are censored rather than the percentage of such values.[5] SC&A does not believe it is appropriate to select a universal upper limit on the percentage of censored results (i.e., a percentage of censored results above which multiple imputation should not be used on the dataset). Alternately, SC&A believes each dataset should be evaluated individually, with the emphasis on the total number of uncensored results rather than the percentage of the total available datapoints.

Even in the case of no detects, there may be information that could be extracted from the LLD itself, even though this process is referred to as uninformative imputation because there are no uncensored data to impute a model. The LLD is itself a bearer of information that can be used to provide bounds on the data. This is discussed in Section 6 of RPRT-0096. There is no reason to select an arbitrary level of censoring a priori.

Finally, there is one other imputation model that could be considered. There are several cases in RPRT-0096 in which the data appear to be a mixture of positive and negative values. So, some values do not have a log. It may be that these are cases where there is mixture of data from unexposed workers (zero exposure) along with data from exposed workers. This is discussed in the third bullet of Section 1.1 of RPRT-0096. Aitchison and Brown (1957) discuss a distribution they call a delta distribution, which is a mixture of a lognormal and a discrete probability at zero. There is a method for estimating the three parameters of this distribution i.e., delta, mu and sigma (Gogolak, 1986) even when the data are censored. This is the case when there are

[4] As with any assumption about statistical data, the analysis benefits from robust assumptions; i.e., those that can that are largely unaffected by outliers or small departures from model assumptions in a given dataset. This is also given in Section 1.1 of RPRT-0096 as a reason for the choice of a lognormal model.

[5] For example, 90 of 100 censored may behave differently than 9 out of 10.

nondetects but there is also some chance that among the censored data there are cases where the concentration is truly zero (e.g., never exposed, never released).

This method was tested and described by Gogolak (1986) with such a data set: krypton-85 concentrations in air near a fuel reprocessing plant that was not always running. It turns out that there is a simple maximum likelihood estimate (MLE) for the probability of a true zero, using only the information on the "less than" data, The MLEs for the mean and standard deviation are the same as for a truncated lognormal distribution that does not use the information about the nondetects. This worked quite well. The MLEs outperformed the minimum variance unbiased estimator (MVUE) Cramer-Rao bound in terms of mean squared error (MSE), and the MSEs approached the MVUE value (Cramer-Rao bound from below) as the sample size increased and the bias went to zero. It turns out that this works for any distribution with positive support that is "delta-ized" by mixing with a discrete probability for zero. The MLEs are the same; just substitute, for example, gamma or Weibull for lognormal. This suggests that NIOSH might also consider the delta distribution for use in an imputation modeling. In particular, the delta distribution may be preferable for datasets with a large proportion of unexposed workers mixed with the exposed worker population. In these cases, the lognormal fits to the data may be problematic.

Implications for Dose Reconstruction and Co-Exposure Modeling

It is important to note that individual dose reconstruction methods for monitored and unmonitored workers are not the same regarding imputation. Specifically, dose reconstructions for monitored workers do not use imputation methods to replace censored data in the worker's monitoring file. Rather, "missed dose" methods are applied to censored results to evaluate the intake of radioactive material as outlined in the NIOSH technical information bulletin, ORAUT-OTIB-0060, revision 02, "Internal Dose Reconstruction" (NIOSH, 2018). To summarize this methodology, NIOSH (2018) treats the individual censored bioassay result at a value of one-half the MDA/censoring level and calculates a chronic intake rate assuming the bioassay result occurred at the midway point between two claimant-specific dates. The start date will typically be the first day of covered employment, the date of the previous relevant bioassay sample, or some other applicable date such as a change in job title from nonradiological to radiological work. The end date for evaluation is typically the actual submission date of the censored bioassay result. For estimation of probability of causation (POC), the resulting organ-specific dose estimate is treated as a triangular distribution with a mode of the calculated dose, a minimum value of zero, and a maximum value of double the mode (dose) estimate.

Unlike missed dose, unmonitored dose assignment uses co-exposure modeling of the available dataset from the monitored worker population to develop distributions of bioassay, intake, and, ultimately, dose values. As described in the introduction to the memorandum, these distributions often must be developed from datasets with a significant portion of data that has been censored. Applying imputation methods to datasets with large portions of censored data will often result in 50th percentile estimates that are necessarily below the MDA/censoring level and sometimes well below one-half of the MDA/censoring level.

Ostensibly when this occurs, one might logically conclude that the application of the co-exposure model based on imputed bioassay results would result in unreasonably low dose estimates. This

would be especially concerning when comparing cases where missed dose is applied to the monitored worker based on one-half of the MDA while unmonitored co-exposure dose assignment is based on imputed bioassay values that are significantly below one-half the MDA.

SC&A (2020) discusses this subject in observation 1 and finding 2 of that report:

> **Observation 1:** While the multiple imputation method is mathematically correct, it has the potential to result in biasing the simulated bioassay results unnecessarily low. Alternate approaches, such as the maximum possible mean method, which replaces censored data with the actual censoring limit (or alternately one-half the censoring limit), would solve the issues associated with datasets containing a large number of censored values in a claimant-favorable manner. [SC&A, 2020, p. 11]

> **Finding 2:** Use of imputed values that are less than one-half of the MDA raises a fundamental fairness issue in that monitored workers who have bioassay results that are less than the MDA are assigned a missed dose in accordance with ORAUT-OTIB-0060 (NIOSH, 2018). Per that guidance, bioassay values that are censored are assumed to be equal to one-half of the MDA rather than an alternate imputed value. [SC&A, 2020, p. 29]

While it is clear that the use of multiple imputation on censored bioassay results may result in estimates of bioassay values that are much lower than the simpler missed dose approach, the overall effect on dose, and ultimately the POC, is less clearly defined. To assess the practical differences in the two approaches, SC&A performed scoping calculations that are presented in Section 3.3 of SC&A (2020). The scoping calculations evaluated hypothetical scenarios that compared the calculated POC using the missed dose approach with co-exposure doses that used imputation methods. SC&A performed calculations for the following radionuclides: strontium-90, cobalt-60, neptunium-237, plutonium-239, and uranium-234. SC&A (2020) had two observations based on this comparative analysis:

> **Observation 2:** A scoping assessment of the use of coworker bioassay data that are significantly less than the MDA versus an alternate missed dose approach concluded that, while intakes and doses are significantly higher using a missed dose approach in most of the sample calculations, the overall effect on resulting POC values was relatively minor, and, in most cases, the coworker-derived POC bounded the missed dose evaluation. This appears to be due to the effect the statistical distribution has on resulting POC values, namely, the use of a triangular distribution for missed dose evaluation versus a lognormal distribution for coworker data. [SC&A, 2020, p. 36]

> **Observation 3:** The sample comparison of coworker intakes to a missed dose method for uranium showed that the coworker model derived intakes were a factor of 4 or more higher than the missed dose approach. This illustrates the potential for inequity between the treatment of unmonitored workers assigned coworker intakes and monitored workers with results less than the detection limit in some situations. [SC&A, 2020, p. 37]

Surprisingly, the two methods actually showed remarkable agreement for many of the hypothetical situations. In fact, co-exposure estimates significantly bounded the missed dose approach for some situations (e.g. uranium). Therefore, the practical effect on the POC using the two methods may be of limited significance.

It should also be noted that typically the 95th percentile of the co-exposure distribution is applied to radiological workers who should have been monitored and were not. In most applications, it is likely that the 95th percentile assignment is reflective of uncensored bioassay results except for situations where nearly all of the available data are censored.

Summary Conclusion

Based on the technical and statistical basis for the multiple imputation method discussed in this memorandum, SC&A concludes that it is a mathematically accurate method for assessing censored bioassay data in the absence of other information (such as the actual raw measurements). This was noted in SC&A's original review of the SRS co-exposure models (refer to observation 1 from SC&A (2020) shown above). In addition, SC&A recommends that the total number of uncensored results (rather than the percentage of total results) be the driving factor in evaluating whether multiple imputation can be appropriately applied to an individual dataset. Further, SC&A believes that NIOSH may want to consider the benefits of applying the delta distribution for cases where the dataset indicates a large proportion of unexposed workers and the lognormal fits to the available data are less than ideal.

Aside from the technical considerations, there is the more philosophical policy question of the use of multiple imputation under the auspices of EEOICPA. In a general sense, the issue of multiple imputation seeks to answer the question: *How do you treat bioassay data when the true value of the result could be anywhere from zero to the MDA/censoring limit?* When faced with this type of uncertainty, the EEOICPA program will typically err on the side of caution and choose claimant-favorable dose reconstruction approaches. As noted in SC&A (2020), such dose reconstruction approaches might include assigning each censored value at one-half the MDA (which is consistent with the missed dose approach) or, alternately, the maximum possible result (which would be the MDA/censoring level itself). However, the substitution approach also has a number of noted analytical drawbacks, as pointed out by Helsel (2020).

When a more scientifically defensible approach is available, then the best scientific practice should be considered appropriate for dose reconstruction. Furthermore, SC&A's (2020) scoping calculations presented in its review of the SRS co-exposure models indicated that there may be very little practical difference between missed dose approaches and co-exposure modeling evaluated at the 50th percentile. In addition, most unmonitored radiological workers would have the 95th percentile of the co-exposure model applied, which is likely reflective of uncensored bioassay results. In conclusion, SC&A finds that the use of multiple imputation in evaluation of bioassay datasets with censored results is technically appropriate, scientifically defensible, and likely of small practical significance when considering its effect on resulting POC calculations.

References

Aitchison, J., & Brown, J. A. C. (1957). *The lognormal distribution*. Cambridge: Cambridge University Press.

Gogolak, C. V. (1986). *Parameter estimation for censored samples from delta-ized distributions with application to air quality monitoring data* [Doctoral dissertation, Polytechnic University of Brooklyn]. Proquest.com Pub. ID 27948318

Helsel, D. R. (2009). Much ado about next to nothing: Incorporating nondetects in science. *Annals of Occupational Hygiene, 54*(3), 257–262.

Helsel, D. R. (2012). *Statistics for censored environmental data using Minitab® and R* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Helsel, D. R. (2020). *Why not substitute ½ DL for nondetects?* PracticalStats.com. Retrieved from https://www.practicalstats.com/resources/Webinar-pdfs/Why-Not-Sub.pdf

International Organization for Standardization/International Electrotechnical Commission (ISO/IEC). (2008). *Uncertainty of measurement —Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)* (ISO/IEC Guide 98-3:2008).

Krishnamoorthy, K., Mallick, A., & Mathew, T. (2009). Model-based imputation approach for data analysis in the presence of non-detects. *Annals of Occupational Hygiene, 53*(3), 249–263.

Miesch, A. T. (1967). *Methods of computation for estimating geochemical abundance* (Geological Survey Professional Paper 574-B). Retrieved from https://pubs.usgs.gov/pp/0574b/report.pdf

National Institute for Occupational Safety and Health (NIOSH). (2018). *Internal dose reconstruction* (ORAUT-OTIB-0060, rev. 02). Retrieved from https://www.cdc.gov/niosh/ocas/pdfs/tibs/or-t60-r2-508.pdf

National Institute for Occupational Safety and Health (NIOSH). (2019a). *Multiple imputation applied to bioassay coworker models* (ORAUT-RPRT-0096, rev. 00). Retrieved from SRDB Ref. ID 175396

National Institute for Occupational Safety and Health (NIOSH). (2019b). *Internal coworker dosimetry data for the Savannah River Site* (ORAUT-OTIB-0081, rev. 04). Retrieved from https://www.cdc.gov/niosh/ocas/pdfs/tibs/or-t81-r4-508.pdf

Savannah River Site and SEC Issues Work Group (SRS & SEC WG). (2019). *Advisory Board on Radiation and Worker Health Savannah River Site (SRS) and SEC Issues Work Groups Joint Meeting Thursday, December 5, 2019* [Transcript of meeting]. Hebron, KY. Retrieved from https://www.cdc.gov/niosh/ocas/pdfs/abrwh/2019/wgtr120519-508.pdf

SC&A, Inc. (2019). *Review of ORAUT-OTIB-0081, revision 04, "Internal Coworker Dosimetry Data for the Savannah River Site"* (SCA-TR-2019-SEC004, rev. 0). Retrieved from SRDB Ref. ID 178392

SC&A, Inc. (2020). *Review of ORAUT-OTIB-0081, revision 04, "Internal Coworker Dosimetry Data for the Savannah River Site"* (SCA-TR-2019-SEC004, rev. 1).