DOE Review Release 07/19/2012

Document Title:	Document Number:	ORAUT-RPRT-0053
Analysis of Stratified Coworker Datasets	Revision:	01
	Effective Date:	07/16/2012
	Type of Document:	Report
	Supersedes:	Revision 00

Subject Expert(s): Thomas R. LaBone, Nancy Chalmers, and Daniel Stancescu (NIOSH)

Approval:	Signature on File Thomas R. LaBone, Document Owner	Approval Date:	07/11/2012
Concurrence:	Signature on File John M. Byrne, Objective 1 Manager	Concurrence Date:	07/10/2012
Concurrence:	Signature on File Edward F. Maher, Objective 3 Manager	Concurrence Date:	07/10/2012
Concurrence:	Vickie S. Short Signature on File for Kate Kimpan, Project Director	Concurrence Date:	07/10/2012
Approval:	Signature on File James W. Neton, Associate Director for Science	Approval Date:	07/16/2012

☐ New ☒ Total Rewrite ☐ Revision ☐ Page Change

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

PUBLICATION RECORD

EFFECTIVE DATE	REVISION NUMBER	DESCRIPTION
10/28/2011	00	New report initiated to discuss methods for analyzing stratified coworker datasets. Incorporates formal internal and NIOSH review comments. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.
07/16/2012	01	Revision initiated to add methods for handling multiple comparisons and for comparing chronic intake rates. Incorporates formal internal and NIOSH review comments. Constitutes a total rewrite of the document. Training required: As determined by the Objective Manager. Initiated by Thomas R. LaBone.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS AND ABBREVIATIONS

AMAD	activity median aerodynamic diameter
CDF	cumulative distribution function
dpm	disintegrations per minute
EDF	empirical distribution function
FWE	family-wise error
GM	geometric mean
GSD	geometric standard deviation
HTO	tritium oxide (water or water vapor)
IMBA	Integrated Modules for Bioassay Analysis
IREP	Interactive RadioEpidemiological Program
IRF	intake retention fraction
KM	Kaplan-Meier
MPM	maximum possible mean
OPOS	one person – one sample
P-P	probability - probability
POC	probability of causation
ROS	regression on order statistics
TLD	thermoluminescent dosimeter
μm	micrometer

1.0 INTRODUCTION

Coworker models are used to estimate doses for workers who were not monitored for exposure to radioactive materials and external sources of radiation but, in retrospect, perhaps should have been (ORAUT 2007, p. 13). Such a dose is referred to as *unmonitored dose*. Coworker models are typically constructed using data from all monitored workers by fitting a lognormal probability distribution to the data (ORAUT 2005, 2006) to estimate the geometric mean (GM) and geometric standard deviation (GSD) of the doses. This procedure was extended to use a simple random sample of the monitored workers if a complete dataset was not available (ORAUT 2009a). Coworker models for external dose are usually constructed from external doses assigned to individuals with film badges and thermoluminescent dosimeters (TLDs). Coworker models for internal dose are calculated using bioassay data that are later evaluated in terms of chronic intake rates and, ultimately, internal doses.

It is reasonable to postulate that the population of all monitored workers is a conglomeration of a number of smaller subgroups of monitored workers, where the subgroups could receive significantly different average doses. For example, the group of workers who routinely make process line breaks in a heavy-water-moderated reactor might be expected to have a significantly different average tritium dose than the group of workers who worked in the control room of the reactor. In sampling theory, these relatively homogeneous subgroups are called *strata*.

Breaking a truly heterogeneous population into a number of relatively homogeneous strata is often desirable because the variance of the estimated parameters[1] will be smaller than the variance of the parameters estimated for the whole population of monitored workers (Lohr 2010, p. 74) and in general the parameter estimates will be more accurate. However, there are issues associated with stratified sampling. For example, criteria are needed to identify meaningful strata[2] and assign workers to the appropriate stratum. The term *meaningful* refers to the assumption that there are groups in the population of monitored workers that have significantly different average doses and that we know how to identify these groups. This is important because, if there is no real difference in the strata, the estimates of coworker dose made from the strata will be less precise than those obtained by using all monitored workers (i.e., from simple random sampling) because of the smaller sample size. Thus, it is important to decide if strata constructed from a population of monitored workers are significantly different before constructing coworker models for each stratum.

The purpose of this report is to detail statistical tests that can be used to decide if two strata from a given group of monitored workers are significantly[3] different. Significantly different strata could warrant coworker models based on the strata rather than the entire population of monitored workers if the difference is of practical significance.

2.0 INTERNAL DOSE COWORKER MODELS

Consider the process of assigning unmonitored internal dose to an individual, as shown in Figure 2-1. The coworker model starts with all the bioassay data for the monitored workers, which are then summarized with the "one person, one sample" (OPOS) bioassay statistics (see the discussion in the next section). The OPOS bioassay statistics are modeled to give the 50th- and 84th-percentile bioassay values, which are derived from the GM and GSD of the fit to the bioassay statistics. In the final step of the coworker modeling (Step 4), the 50th and 84th percentiles are modeled with the Integrated Modules for Bioassay Analysis (IMBA) computer program to give 50th- and 84th-percentile

[1] Like the variance of GM estimate and the variance of the GSD estimate.

[2] The strata need to be identified before seeing the bioassay data from the monitored population. To do otherwise can be considered to be a form of "data dredging" that invalidates the tests used to decide if the strata are different.

[3] In this report the term "significant" used by itself refers to statistical significance, which is not the same as practical significance.

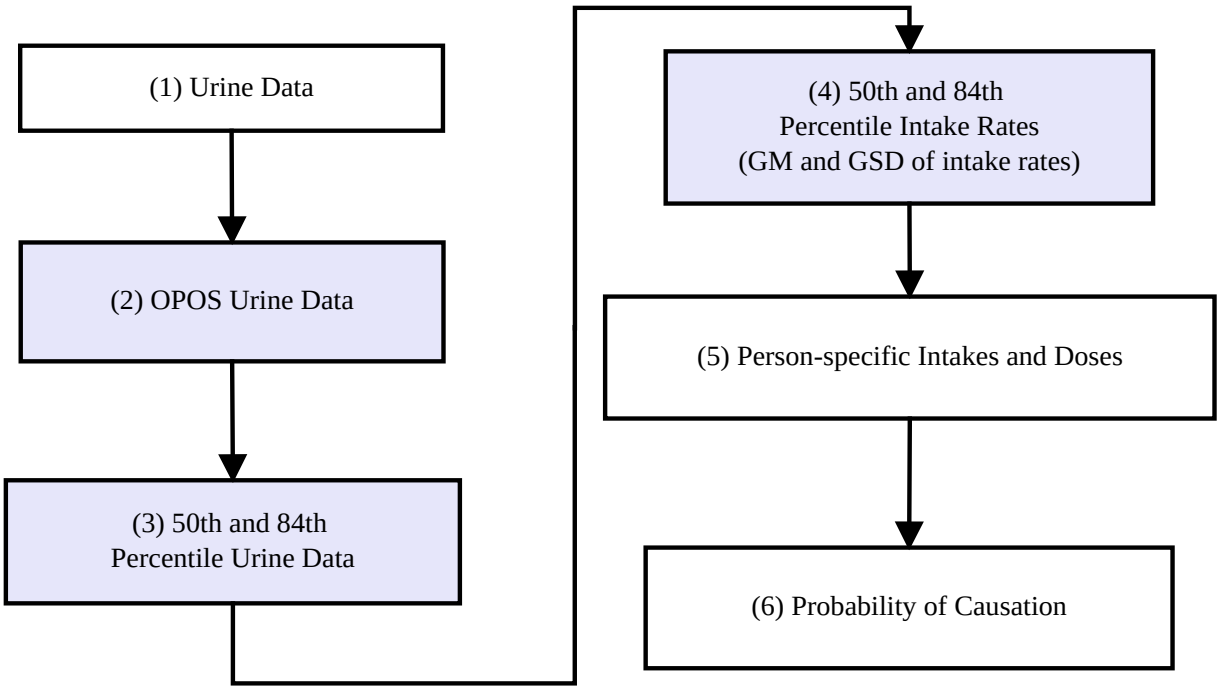

Figure 2-1. Summary of calculations performed for internal dose coworker modeling. Steps 2, 3, and 4 (shaded) are the stages of the process selected to compare strata.

chronic intake rates for each period of interest. These intake rates are used to calculate the GM and GSD of the intake rates, which are published for use by dose reconstructors who examine an individual's work history and assign chronic intakes to periods that are appropriate for that individual. The chronic intakes are then used to calculate organ-specific doses for each year, which are added to the Interactive RadioEpidemiological Program (IREP) along with doses from all other sources (e.g., external, environmental, medical) to calculate the probability of causation (POC) for the type of cancer in question.

The ultimate question here is: *Would the decision to compensate the individual be different if unmonitored dose was assigned using a coworker model based on Stratum A versus a coworker model based on the entire population of monitored workers*? The answer to this question depends on the POC assigned to the individual (Step 6 in Figure 2-1), which in turn depends in part on the intakes and doses assigned (Step 5). The compensation decisions in Steps 6 and 5 are very specific to the individual and, therefore, it is not feasible to compare strata in a general way at these stages of the process. At the other end of the process (Step 1), the comparison of strata is complicated because of the significant dependences that can exist in the data because one individual can have multiple bioassay results. This leaves Steps 2, 3, and 4 (the shaded boxes in Figure 2-1) as the best points at which to compare strata. Before discussing the pros and cons of performing the comparisons at each step, OPOS data will be discussed.

3.0 OPOS DATA

Monitoring for workers potentially exposed to external radiation typically consists of a dosimeter (e.g., TLD or film badge) worn on the upper torso of the body. The dosimeter indicates the cumulative dose to the whole body received between readings of the dosimeter, creating monitoring intervals that can be anywhere from days to months in duration. A key property of the doses reported for each interval is that, for a given individual, the doses are statistically independent of each other. For example, by itself, the dose received in a given month provides no information about the dose that will be received

in the next month and, similarly, the dose received in a given year provides no information about the dose that will be received in the next year.

Internal dose monitoring programs are in many ways similar to the external dose programs described above, except internal dose programs tend to be based on bioassay rather than dosimeters. This is an important difference because, if an individual had an intake of radioactive material such that a bioassay is positive at a given time, subsequent bioassay results could also be positive even if there are no additional intakes. This means bioassay results can be correlated with each other and are no longer statistically independent. This correlation could exist for a brief period, following a small intake of tritium oxide (HTO), for example, or for the length of a career, following a large intake of plutonium, for example.

Operational bioassay programs can generate multiple results for an individual in a given period (e.g., a year), which creates a related problem if an individual is involved in an incident and has more (potentially many more[4]) bioassay results than other workers. If these are not accounted for, the problems of correlated data and unequal number of samples per person can violate the assumptions on which the linear regression used to model the data and the statistical tests used to compare strata in the population are based (see Fox 2008, p. 100).

The ultimate solution to this problem is to convert bioassay data to committed effective doses. Unfortunately, this is usually feasible only for workers exposed to materials such as HTO that have simple biokinetic models and short retention times in the body. A bigger problem in practice is that organ doses, not effective doses, are required for dose reconstruction. This significantly complicates coworker models because there is no easy way to go from effective dose to organ dose, which means a coworker model would be required for each organ of interest. The next best solution is to generate a single statistic that characterizes multiple bioassay results for each person in a given period. This is referred to as an OPOS statistic. Obvious choices for such a statistic are the mean bioassay result, the median bioassay result, and the maximum bioassay result. The choice of statistic is complicated by the presence of censored data (i.e., bioassay results reported as less than some analytical detection level).

When we are presented with more than one bioassay result, all of which are uncensored, it is reasonable to take the mean to be the best point estimate to characterize the data (see Example A below). If there are censored data, we can take the mean using the face value of the censored results and consider it to be *maximum possible mean* (MPM). If there are uncensored results in the dataset, the MPM would be used as an uncensored result (Example B), whereas if all data are censored, the MPM would be used as a censored result (Example C).

Example A: 10, 3, 5, 6
Mean = 24/4 = 6 (report as 6)

Example B: 10, <3, <5, 6
Maximum Mean = 24/4 = 6 (report as 6)

Example C: <10, <3, <5, <6
Maximum Mean = 24/4 = 6 (report as <6)

For lognormally distributed data, the MPM will be greater than the true median and mean of the dataset. In short, the MPM has the desirable properties of being easy to calculate and applicable to all datasets and, at the same time, being conservative if censored data are present.

[4] One dataset we observed had 150 bioassay results in a given year, with 100 results from one individual.

4.0 METHODS FOR MODELING DATA AT STEP 3

Step 3 is where the 50th and 84th percentiles of the OPOS data are calculated using a lognormal parametric model. The lognormal parametric model is used to fit the bioassay data because:

- Experience has shown that occupational dose and bioassay data are frequently fit well by a lognormal model.
- The lognormal model is one of the standard options in the IREP software; that is, the GM and GSD derived from the coworker model plug directly into IREP.

Three approaches used for fitting a lognormal model to bioassay data are discussed below. Detailed examples for all three methods, implemented in the R computer code (Helsel and Lee 2010), are in Attachment A. As a general guideline, the minimum sample size used for coworker modeling is 30 individuals (i.e., 30 OPOS results) in a given period.[5] This minimum[6] can be relaxed if, in the judgment of the statistician performing the analysis, the uncertainty in the resulting parameter estimates is not excessive. The minimum sample size applies to a *sample of the monitored workers* and does not apply to the *population of monitored workers*; for example, if the entire population of monitored workers at a given site in a given year is 25, the coworker model derived from this population is valid.

4.1 REGRESSION ON ORDER STATISTICS

The method used in ORAUT-PROC-0095 (ORAUT 2006) to estimate the GM and GSD of a censored dataset is referred to as *regression on order statistics* (ROS) (Helsel 2005, Chapter 6; D`Agostino and Stephens 1986, Chapter 11; Cullen and Frey 1999, Chapter 5). In ROS the logs of the empirical quantiles (the observed bioassay results) are plotted against the theoretical quantiles from a standard normal distribution (the z-scores or standard deviations). If the points fall approximately along a regression line, the data are taken to be lognormally distributed with log of the GM given by the y-intercept of the line and the log of the GSD given by its slope. ORAUT-PROC-0095 uses Hazen plotting points (Cullen and Frey 1999, p. 94), which are referred to as "percentile midpoints" in the procedure. Hazen plotting points are valid only for datasets with a single left-censoring level (i.e., where a single decision level is applied to all data in the dataset). Here, Helsel-Cohn plotting points are used (Helsel and Cohn 1988), which are suitable for single and multiple left-censoring (i.e., where multiple distinct decision levels are applied to the data in the dataset). With the exception of the plotting points, the ROS used here is the same as the ROS used in ORAUT-PROC-0095. ROS is the method of choice for estimating the GM and GSD of a dataset, but it can generate misleading results if the dataset is highly censored (more than ~80%; see Helsel 2005, p. 78) or consists of two or more distributions with significantly different GMs and GSDs. The following two methods should be considered for use on such datasets.

4.2 EFFECTIVE FIT (MAXIMUM LIKELIHOOD)

The use of ROS on some datasets can yield excellent fits that nevertheless give estimates of GSD that are very large.[7] These datasets are assumed to consist of two distinct groups: an analytical

---

[5] Data from multiple years (usually no more than 3) can be combined to achieve this minimum if the conditions in the workplace are reasonably constant over the period in question.

[6] The U.S. Environmental Protection Agency (Singh, Armbya, and Singh 2010, p. 27) discusses minimum sample size required for performing statistical tests on censored datasets and recommends ~15 results per sample (stratum) as a minimum. Here we are estimating parameters from the data, so the default minimum has been increased to 30.

[7] A GSD in excess of 6 is considered to be large based on ORAUT experience with developing coworker models. GSDs in excess of 1,000 have been observed.

background distribution that might be censored (in part or in whole) and an exposed worker distribution that has a GM considerably higher than that of the analytical background. In ORAUT-RPRT-0044 (ORAUT 2009b), maximum likelihood methods were used to decompose such datasets into the analytical background and the true exposed worker data. Here, the analytical background distribution determined from the maximum likelihood fit is used to impute the censored data (see Helsel 2005, p. 68). The combination of these imputed data and the observed uncensored data is modeled using standard ROS assuming a single lognormal distribution. This is referred to as an *effective fit*, and it gives estimates of the GM and GSD that reflect the relative proportions of censored and uncensored data. An effective fit should be considered if the GSD calculated with ROS exceeds 6.

4.3 BINOMIAL FIT

Some datasets have a significant level of censoring, up to and including complete censoring (i.e., all bioassay results are reported as less than the detection level). Methods for estimating the GM and GSD of such highly censored datasets are presented in ORAUT-RPRT-0044 (ORAUT 2009b). The binomial fit should be considered for datasets if more than ~85% of the data are censored or the absolute number of uncensored data is small (e.g., in the range of 0 to 10). However, because it is difficult to give precise guidelines for when to apply the binomial fit that are appropriate for all datasets, the final decision on when to use it is determined on a case-by-case basis.

Only one censoring level can be used in the binomial fit because it dichotomizes the data into the number of data above the censoring level and the number of data below the censoring level. For datasets in which there are multiple censoring levels, ORAUT-RPRT-0044 suggests that the highest censoring level be used. This practice can be unreasonable if there is a single unusually high censoring level in the dataset.[8] Here it is proposed to use the smallest censoring level such that at least 95% of the censored data are below that value. This ensures the use of at least 95% of the data in the binomial fits.

5.0 TESTS FOR COMPARING STRATA AT STEP 2 AND STEP 3

Assume the monitored population is stratified into Stratum A and Stratum B and it is to be determined if Stratum A warrants having its own coworker model. To answer this question

- Compare Stratum A to Stratum B, or
- Compare Stratum A to the monitored population (the union of Stratum A and Stratum B).

The second option has the advantage of comparing one option to its alternative (i.e., for the individuals in Stratum A either a coworker model based on all the monitored workers will be used or a coworker model based on the workers in Stratum A will be used). The problem with this approach is that Stratum A is not independent of the population of all monitored workers (because the stratum is a subset of the population) and this dependence complicates the test. The first option does not have this complication because Stratum A and Stratum B are assumed to be independent of each other. For this reason, the statistical tests discussed below compare the strata to each other rather than comparing a stratum to the population of monitored workers from which it was drawn.

5.1 COMPARISON OF STRATA AT STEP 3: MONTE CARLO PERMUTATION TEST

It is practical to compare strata at Step 3 as long as the bioassay data are fit using the ROS method. For bioassay data that require an effective fit, the comparison of strata is pushed upstream to Step 2 (for reasons discussed below). Datasets that are modeled using the binomial fit are considered to not

[8] For example, there can be a variety of censoring levels between 5 and 10 and a single censoring level at 100.

contain enough information to decide if strata are different, and it is recommended that such datasets not be stratified.

The Monte Carlo permutation test described in ORAUT (2009a, 2010) and Noreen (1989, p. 43) is used here because it simultaneously compares the GM and GSD of the fits to the two strata and provides a very informative graphic that enables one to see if the strata are different. The technical specifications of the test are given below and it is implemented from first principles in the R code associated with this report (LaBone 2012).

Null Hypothesis H_0:
The GM and GSD of fits to Stratum A and Stratum B are the same. In other words, the coworker model derived from Stratum A is the same as the coworker model derived from Stratum B.

Alternate Hypothesis H_A:
Opposite of Hypothesis H_0 – the GM or the GSD of the two strata is different.

Test statistic
The coworker model consists of the GM and GSD of a lognormal model fit to the ordered OPOS statistics for a given year. The test statistics used here are the differences between the GM and GSD from the two strata:

$$\Delta_{gm} = GM_A - GM_B \quad (5\text{-}1)$$

$$\Delta_{gsd} = GSD_A - GSD_B \quad (5\text{-}2)$$

The joint distribution of the Δ_{gm} and Δ_{gsd} under the null hypothesis is calculated using a Monte Carlo permutation test that looks at a large number of fits of the two strata (10,000 fits typically). For this reason, the permutation test can be performed only for datasets that can be modeled using ROS, which does not require manual intervention.

Rejection Region
For a single test (see section below on multiple comparisons), the null hypothesis is rejected if the confidence ellipse (based on the observed Δ_{gm} and Δ_{gsd} pair) is greater than 95%. The Δ_{gm} and Δ_{gsd} pair represents x and y coordinates of a point on a scatter plot. The simulation consists of 10,000 such points with a bivariate normal confidence ellipse (Fox 2008, p. 203; Monette 1990) constructed through the point at the observed Δ_{gm} and Δ_{gsd} pair. The probability of the ellipse is presented on the plot if it is less than 99%. The null hypothesis is rejected if this ellipse is greater than 95%. If the point defined by the observed Δ_{gm} and Δ_{gsd} pair falls outside the 99% confidence ellipse, a 99% confidence ellipse is presented on the plot and the null hypothesis is rejected.

Alternate Rejection Region
Some datasets containing censored data can generate an asymmetric cloud of Δ_{gm} and Δ_{gsd} points that is clearly not bivariate normal. In these cases a nonparametric bagplot (Rousseeuw, Ruts, and Tukey 1999), is used to construct an ~95% confidence polygon. If the point defined by the observed Δ_{gm} and Δ_{gsd} pair falls outside the ~95% confidence polygon, the null hypothesis for a single test is rejected. Examples of the Monte Carlo permutation test are given in Example 1 in Attachment A.

5.2 COMPARISON OF STRATA AT STEP 2: PETO-PRENTICE TEST

Datasets that are highly censored can cause the Monte Carlo permutation test discussed above to fail because one or more of the 10,000 random draws results in a completely censored stratum that cannot be fit by ROS. In addition, strata modeled using an effective fit cannot be compared using the

Monte Carlo permutation test because the effective fitting procedure usually requires a considerable degree of manual intervention by the statistician (i.e., it is not feasible to perform 10,000 effective fits on random draws from the monitored population). Finally, it can be difficult to adjust for multiple comparisons for the tests performed at Step 3. For these reasons, a test for comparison of strata at Step 2 is presented.

A family of survival-analysis tests used to compare two sets of data to determine if they have the same statistical distribution includes the Gehan Test, Logrank Test, Peto-Prentice Test, and Tarone-Ware Test (Helsel 2005, Chapter 9; Millard and Deverel 1988; Klein and Moeschberger 2003, Section 7.3). These tests compare the survival curves[9] of the two strata (for example, see Figure A-6 in Attachment A) at each point where there are uncensored data. The primary difference in these tests is that they give different weights to various portions of the survival curves for the two strata. For reasons discussed in Attachment B, the Peto-Prentice Test was selected to use for comparing strata at Step 2. The Peto-Prentice Test is implemented in the R package *NADA* (Helsel and Lee 2010) as well as in the SAS System software (SAS 2011). The formal specification of the test is shown below:

Null Hypothesis H_0:
The distribution of the OPOS bioassay data is the same in Stratum A and Stratum B.

Alternate Hypothesis H_A:
Opposite of Hypothesis H_0 – the distribution of data is not the same in Stratum A and Stratum B.

Test statistic
The test statistic is either a Z-value or a chi-square value of the observed versus expected number of uncensored results in the two strata.

Rejection Region
If the two-sided p-value obtained from either the Z statistic or the chi-square statistic with 1 degree of freedom is less than $\alpha = 0.05$, the null hypothesis is rejected.

The Peto-Prentice Test typically leads to the same conclusion as the Monte Carlo permutation test for cases in which both are applicable. Examples of using the Peto-Prentice Test are in Examples 1 and 2 in Attachment A.

5.3 MULTIPLE COMPARISONS

In practice, the two strata are compared for a number of different consecutive time periods (e.g., years). When simultaneous multiple hypothesis tests like this are performed, an adjustment for multiple comparisons should be used. Let α_{single} be the Type I Error rate (or false positive rate) for each of the k individual tests, and let α_{family} be the false positive rate of the family of tests. A *family of tests* is defined as any set of tests for which it makes sense to consider a combined measure of error (Hochberg and Tamhane 1987, pg. 5). For simplicity, assume all k tests are independent. Then,

$$\begin{aligned}\alpha_{family} &= P(\text{at least one false positive among the } k \text{ tests}) \\ &= 1 - P(\text{no false positive among the } k \text{ tests}) \\ &= 1 - [1 - P(\text{false positive for single test})]^k \\ &= 1 - [1 - \alpha_{single}]^k .\end{aligned} \tag{5-3}$$

[9] The term *survival curve* is used in survival analysis for right-censored data like the lifetimes of light bulbs or survival of patients after medical treatment for a disease. For bioassay data, which are left-censored, we refer to these curves as the empirical distribution function (EDF) plots.

For example, suppose 20 independent hypothesis tests are conducted simultaneously using a 0.05 significance level, meaning $\alpha_{single} = 0.05$. The probability of at least one false positive among the family of tests is

$$\alpha_{family} = 1 - [1 - 0.05]^{20} = 0.6415 . \quad (5\text{-}4)$$

This family of tests has a false positive rate of more than 64%, which is much larger than the single-test false positive rate of 5%.

Bonferroni Correction

To control the false positive rate for a family of tests, family-wise error (FWE) methods are employed. These FWE methods control the overall Type I Error rate for all the comparisons (i.e., the false positive rate for the family of tests). The most widely used FWE method is Bonferroni's correction (Rice 2007, p. 458). This correction uses Boole's inequality to control the FWE. Assuming each of the single tests has the same significance level and the *k* single tests not being necessarily independent,

$$\begin{aligned} \alpha_{family} &= P(\text{at least one false positive among the } k \text{ tests}) \\ &= P\left(\bigcup_{i=1}^{k} \text{false positive for Test } i\right) \\ &\le \sum_{i=1}^{k} P(\text{false positive for Test } i) \\ &= k\alpha_{single}. \end{aligned} \quad (5\text{-}5)$$

Therefore, the Bonferroni correction is

$$\alpha_{single} = \frac{\alpha_{family}}{k} . \quad (5\text{-}6)$$

The p-values from each test are compared to α_{single}. Although it is simple, Bonferroni's method is very conservative due to the use of Boole's inequality in its derivation. The method's conservatism means the chance of Type II Errors (false negatives) is increased, which means the statistical power of the test is reduced. This means that legitimately significant results might not be detected. Note that the Bonferroni method and Holm method (discussed below) require the p-value for each test. In practice, this usually means that these methods can only be used with the Peto-Prentice test.

Holm Method

An alternative FWE procedure is the Holm method, a stepwise Bonferroni procedure that is less conservative and more powerful than the Bonferroni correction (Holm 1979). The Holm method still controls the probability of a Type I Error (false positive), but it is less likely to produce Type II Errors (false negatives). Because this is a stepwise procedure, the p-values from the *k* tests must be sorted from least to greatest. These ordered p-values are denoted as

$$p_{(1)} \le p_{(2)} \le \ldots \le p_{(k)} . \quad (5\text{-}7)$$

In Step 1, the smallest p-value is compared to the ordinary Bonferroni cutoff above. If $p_{(1)} > \alpha_{family}/k$, then $p_{(1)}$ is declared insignificant, and the procedure ends. If $p_{(1)} \le \alpha_{family}/k$, then $p_{(1)}$ is declared significant and the stepwise procedure continues to Step 2.

In Step 2, the next smallest p-value is compared to the Bonferroni cutoff assuming there are only $k - 1$ tests in the family, because the test in Step 1 has already been declared significant. If $p_{(2)} > \alpha_{family}/(k - 1)$, then $p_{(2)}$ is declared insignificant, and the procedure ends. If $p_{(2)} \leq \alpha_{family}/(k - 1)$, then $p_{(2)}$ is declared significant and the stepwise procedure continues to Step 3.

This stepwise process continues until a p-value is declared insignificant. Once an insignificant p-value is discovered, all larger p-values are also insignificant, so the procedure ends.

Most often in coworker analyses, several periods are simultaneously tested using the Peto-Prentice test to determine whether the two strata are significantly different. The tests from these periods can be correlated because the individuals being monitored in one year could also be monitored in following years. Both the Bonferroni and Holm methods control the FWE at a given significance level α_{family}, even if the tests are correlated (Schochet 2008, pg 10).

Example

The p-values for Peto-Prentice tests for a coworker analysis of ^{241}Am are listed in Table 5-1. With no multiple comparisons adjustment, if the significance of each of these tests is determined by comparing the p-values to $\alpha_{single} = 0.05$, the three periods with shaded p-values (years 1981-1982, 1983, and 1985) are considered significant.

Table 5-1. p-values for 19 Peto-Prentice tests performed on coworker data using ^{241}Am.

Year(s)	p-value	Year(s)	p-value
1966–1968	0.375	1978	0.2186
1969	0.1787	1979	0.9082
1970	0.8923	1980	0.2646
1971	0.813	1981–1982	0.0315
1972	0.7053	1983	0.0074
1973	0.3383	1984	0.093
1974	0.8684	1985	0.0005
1975	0.5465	1986	0.1782
1976	0.3021	1987–1989	0.6796
1977	0.4499		

The Holm method, with $k = 19$ tests, gives the Holm cutoff values in Table 5-2 (sorted by p-value on left and sorted by year on right). With no adjustment for multiple testing, from Table 5-1, three periods (1981 to 1982, 1983, and 1985) were deemed significant. After the Holm method adjustment for simultaneously performing 19 Peto-Prentice tests, from Table 5-2, only 1985 is significant. In general, the Holm method is the preferred approach for adjusting for multiple comparisons, but the Bonferroni method is acceptable if it is not practical to use the Holm method.

Table 5-2. p-values and Holm cutoff values (sorted and unsorted) for Peto-Prentice tests performed on coworker data using ^{241}Am.

Year(s)	p-value	Holm cutoff	Year(s)	p-value	Holm cutoff
1985	0.0005	0.0026	1966–1968	0.375	0.0056
1983	0.0074	0.0028	1969	0.1787	0.0036
1981–1982	0.0315	0.0029	1970	0.8923	0.025
1984	0.093	0.0031	1971	0.813	0.0125
1986	0.1782	0.0033	1972	0.7053	0.01
1969	0.1787	0.0036	1973	0.3383	0.005
1978	0.2186	0.0038	1974	0.8684	0.0167
1980	0.2646	0.0042	1975	0.5465	0.0071
1976	0.3021	0.0045	1976	0.3021	0.0045
1973	0.3383	0.005	1977	0.4499	0.0062

Year(s)	p-value	Holm cutoff	Year(s)	p-value	Holm cutoff
1966–1968	0.375	0.0056	1978	0.2186	0.0038
1977	0.4499	0.0062	1979	0.9082	0.05
1975	0.5465	0.0071	1980	0.2646	0.0042
1987–1989	0.6796	0.0083	1981–1982	0.0315	0.0029
1972	0.7053	0.01	1983	0.0074	0.0028
1971	0.813	0.0125	1984	0.093	0.0031
1974	0.8684	0.0167	1985	0.0005	0.0026
1970	0.8923	0.025	1986	0.1782	0.0033
1979	0.9082	0.05	1987–1989	0.6796	0.0083

6.0 COMPARISON OF STRATA AT STEP 4: COMPARISON OF INTAKE RATES

In the ^{241}Am example in the previous section, in one year out of the 19 years tested the strata were determined to be significantly different. The question at this point becomes *are the differences between the strata of practical significance*? One aid to help judge practical significance is to see if the constant chronic intake rates of the two strata calculated from the 50th and 84th percentiles of OPOS data (see Step 4 of Figure 2-1) are significantly different. Continuing with the ^{241}Am example, a test for equal intake rates is presented below.

The annual coworker ^{241}Am urinary excretion data for two strata (Group A and Group B) are listed in Table 6-1 along with the chronic intake retention fractions (IRFs) calculated for type M ^{241}Am. The regression through the origin of the median ^{241}Am excretion rate (in dpm/day) on the associated chronic intake retention fraction gives a line whose slope *b* is the chronic ^{241}Am intake rate (in dpm/day). The standard error of the slope is the uncertainty in the intake rate given that the uncertainty in the excretion rate is the only significant source of uncertainty.[10] The regressions for Group A and Group B are shown in Figures 6-2 and 6-3, respectively. The slopes of the regression lines can be tested using a simple *t* test with the following null and alternate hypotheses:

H_0: The slopes of the two regression lines are equal.
H_a: The slopes of the two regression lines are not equal.

The test statistic is

$$t = \frac{b_A - b_B}{\sqrt{s_A^2 + s_B^2}} = \frac{0.9861 - 0.731}{\sqrt{0.2222^2 + 0.1598^2}} = 0.9321479 \qquad (6\text{-}1)$$

Table 6-1. Median ^{241}Am urinary excretion rates (dpm/day) and associated chronic intake retention fractions for 5-μm AMAD type M material for two different groups of workers.

Year	Group A Median	Group A IRF	Group B Median	Group B IRF
1973	9.52E-03	5.58E-03	7.99E-03	5.58E-03
1974	1.37E-02	8.28E-03	8.59E-03	8.27E-03
1975	9.28E-03	9.95E-03	1.03E-02	9.94E-03
1976	1.53E-02	1.13E-02	1.23E-02	1.13E-02
1977	2.83E-03	1.24E-02	3.73E-03	1.24E-02
1978	1.51E-02	1.33E-02	3.01E-02	1.33E-02
1979	2.64E-02	1.41E-02	2.42E-02	1.41E-02
1980	8.99E-03	1.49E-02	9.62E-03	1.49E-02

[10] In other words we are assuming that the uncertainties in the biokinetic model that was used to generate the IRFs accurately describes the biokinetics of the "individual," the exposure to ^{241}Am was actually a constant chronic, etc.

Year	Group A Median	Group A IRF	Group B Median	Group B IRF
1981–1982	6.66E-03	1.59E-02	2.50E-03	1.59E-02
1983	1.23E-02	1.67E-02	4.25E-03	1.67E-02
1984	1.22E-02	1.73E-02	7.06E-03	1.72E-02
1985	5.14E-02	1.77E-02	9.30E-03	1.77E-02
1986	4.87E-03	1.82E-02	4.89E-03	1.82E-02
1987–1989	9.43E-03	1.90E-02	1.86E-02	1.90E-02

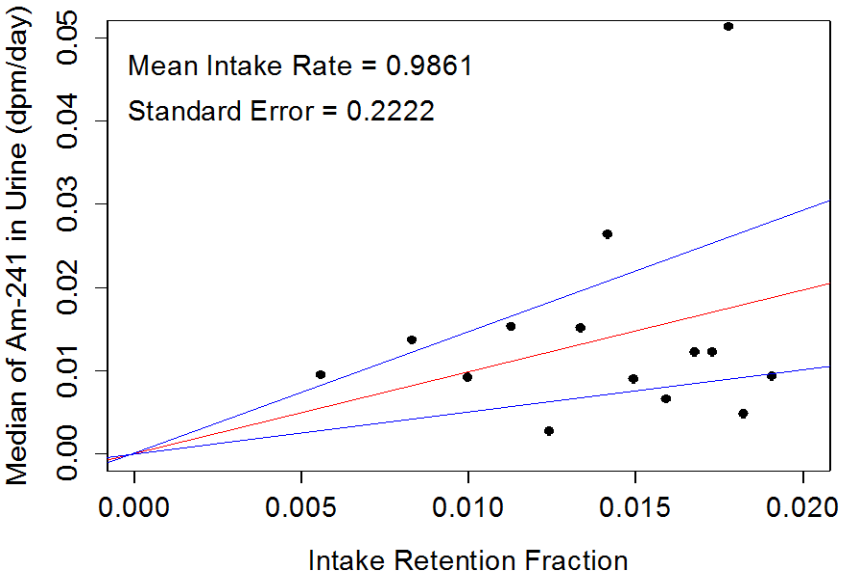

Figure 6-1. Regression of median excretion rate on chronic intake IRF for Group A. The center (red) line is the line of best fit and the outer (blue) lines are the 95% confidence interval of the mean.

For (14 + 14 - 2) = 26 degrees of freedom, the p-value (two-sided) associated with this value of the test statistic is

$$2[1-P(t_{df=26}<0.9321479)]=0.3598359 \quad (6\text{-}2)$$

The null hypothesis is not rejected at a significance level of α = 0.05, which means that there is nothing here to make us believe that the slopes of the two regression lines (the intake rates of the two groups) are significantly different. This test assumes that Group A and Group B have the same number of data, which would be the case most of the time.

Probably the easiest way to get the standard errors for the intake rates is with IMBA. IMBA offers three different regression methods for estimating the intake from a given set of bioassay data: *least squares*, *maximum likelihood*, and *Bayesian*. The IMBA menu for selecting the method is shown below in Figure 6-3.

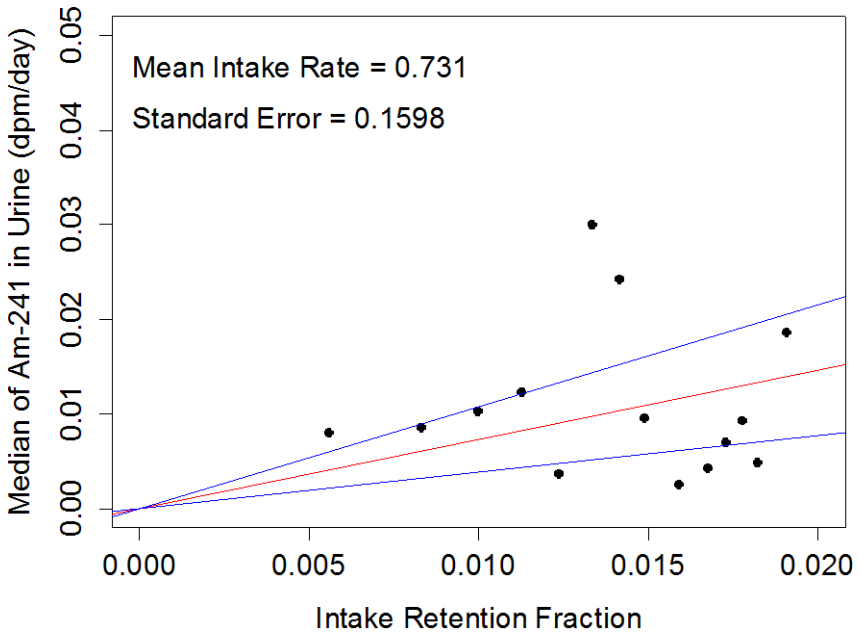

Figure 6-2. Regression of median excretion rate on chronic intake IRF for Group B. The center (red) line is the line of best fit and the outer (blue) lines are the 95% confidence interval of the mean.

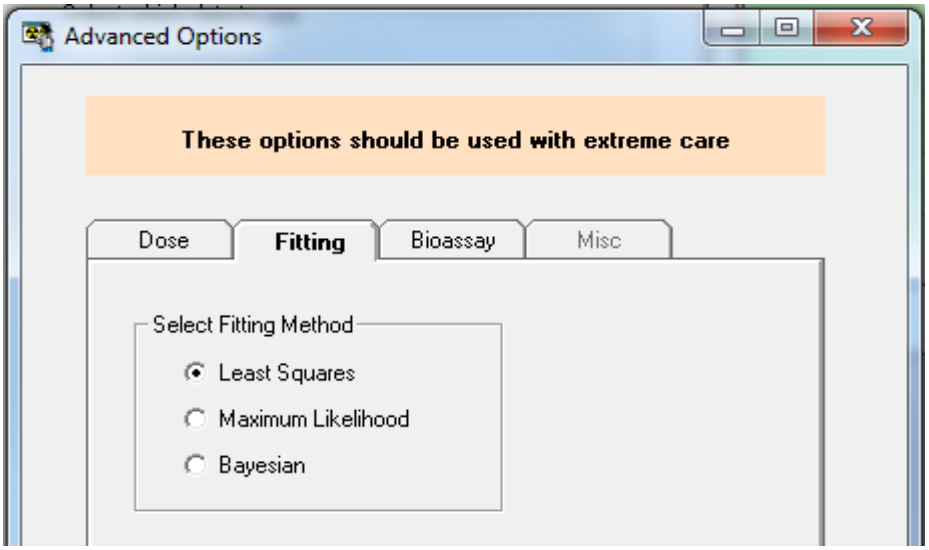

Figure 6-3. IMBA menu for selecting a regression method.

The default method is *maximum likelihood*. The standard errors in the intake are not reported with this method. To get the standard errors the *least squares* method must be selected.[11] For example, the least squares results for the Group A data are shown in Figure 6-4. The intake rate and standard error reported by IMBA (Figure 6-5) are the same as those calculated in R. The primary advantage of using IMBA here is that there is no need to export the chronic intake IRFs from IMBA that are required to do the calculation in R.

[11] The maximum likelihood estimates of intake rate are the same as those obtained from least squares as long as there are no censored bioassay data, a single intake, and normally distributed bioassay measurement errors.

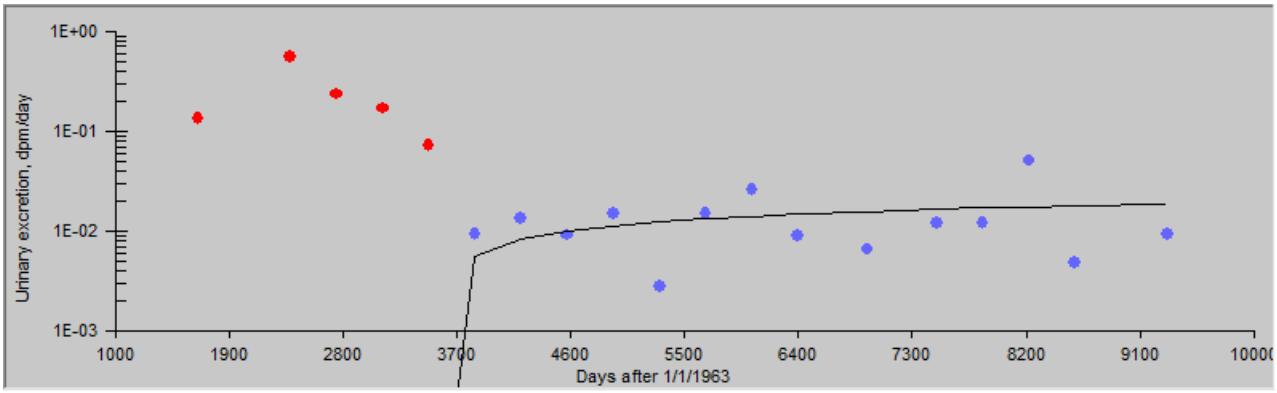

Figure 6-4. Regression of median excretion rate on chronic intake IRF for Group A (time series plot).

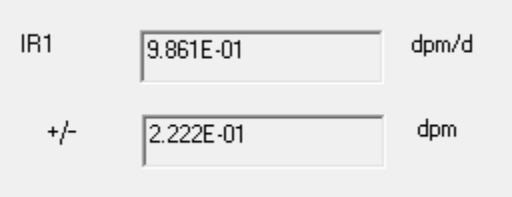

Figure 6-5. Chronic intake rate and uncertainty reported by IMBA.

It is probably fair to ask at this point why the test on chronic intake rates is not the only test performed. The main reasons are:

- The fitting of bioassay data to calculate intake rates is the most subjective of the steps discussed in this report. For example, the internal dosimetrist chose to fit the 1973 through 1989 data as one chronic intake period and the 1966-1972 data as another (see Figure 6-4). Thus, if the strata can be demonstrated to be basically equivalent at Step 2 or Step 3, tests at those steps are preferred.

- To use the methodology presented here the conditions of the simple-regression model must be met (Fox 2008, p 100). When these conditions are grossly violated the fits cannot be tested as described here.

7.0 SUMMARY

Statistical tests are presented in this report that can be used to decide if two strata from a given group of monitored workers are significantly different. Strata that are *statistically and practically* different might warrant coworker models based on the individual strata rather than the entire monitored group of workers. Specifically, the Monte Carlo permutation test described in ORAUT-RPRT-0049 (ORAUT 2010) is used to compare strata based on the parameters from fits of lognormal distributions to the strata. For cases in which the Monte Carlo permutation test cannot be used, the Peto-Prentice Test performed on the OPOS bioassay statistics from the two strata is offered.

Several important modifications to the existing coworker modeling methodology, which are needed to support the testing of strata, are introduced in this report. These include:

- The concept of summarizing multiple bioassay results for each worker in one statistic was formalized. This is the *one person - one sample* approach, and the *maximum possible mean* is offered as the statistic to use. Use of OPOS statistics for all individuals in a given period

eliminates dependencies in the data and improves the validity of statistical tests performed on the data.

- The methods in ORAUT-PROC-0095 (ORAUT 2006) for calculating the GM and GSD from a given set of bioassay data were extended to accommodate multiply left-censored data (i.e., bioassay data reported with different decision levels applied).

- The effective fit is described, which uses methods from ORAUT-RPRT-0044 (ORAUT 2009b) to model bioassay data that have excessively large GSDs.

- The issue of multiple comparisons (e.g., testing strata for a number of years) is addressed.

- A simple method for comparing two chronic intake rates is offered.

REFERENCES

Brown, M., 1984, "On the Choice of Variance for the Logrank Test", *Biometrika*, volume 71, number 1, pp. 65–74.

Cullen, A. C., and H. C. Frey, 1999, *Probabilistic Techniques in Exposure Assessment: A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*, Plenum Press, New York, New York.

D'Agostino, R. B., and M. A. Stephens, 1986, *Goodness-of-Fit Techniques*, Marcel Dekker, New York, New York.

Desu, M. M., and D. Raghavarao, 2004, *Nonparametric Statistical Methods for Complete and Censored Data*, Chapman & Hall/CRC, Boca Raton, Florida.

Fox, J., 2008, *Applied Regression Analysis and Generalized Linear Models*, Sage Publications, London, England.

Gehan, E. A., 1965, "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, volume 52, number1/2, pp. 203–223.

Helsel, D. R., 2005, *Nondetects and Data Analysis: Statistics for Censored Environmental Data*, Wiley Interscience, Hoboken, New Jersey.

Helsel, D. R., and T. A. Cohn, 1988, "Estimation of Descriptive Statistics for Multiply Censored Water Quality Data," *Water Resources Research*, volume 24, number 12, pp. 1997–2004.

Helsel, D. R., and L. Lee, 2010, *R Package NADA Version 1.5-3*, December 22.

Higgins, J. J,, 2004, *An Introduction to Modern Nonparametric Statistics*, Brooks/Cole-Thompson Learning, Pacific Grove, California.

Hochberg, Y. and A. C. Tamhane, 1987, *Multiple Comparison Procedures*, John Wiley & Sons, New York, New York.

Holm, S., 1979, "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, volume 6, number 2, pp. 65–70.

Klein, J. P., and M. L. Moeschberger, 2003, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, New York, New York.

LaBone, T. R., 2012, *R Code Examples in ORAUT-RPRT-0053*, Oak Ridge Associated Universities Team, Oak Ridge, Tennessee, May 18. [SRDB Ref ID: 113404]

Latta, R. B., 1981, "A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data," *Journal of the American Statistical Association*, volume 76, number 375, pp. 713–719.

Lawless, J. F., 1982, *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York, New York.

Letón, E., and P. Zuluaga, 2001, "Equivalence Between the Score and Weighted Tests for Survival Curves," *Communications in Statistics: Theory and Methods*, volume 30, part 4, pp. 591-608.

Letón, E., and P. Zuluaga, 2002, "Survival Tests for r Groups," *Biometrical Journal*, volume 44, number 1, pp. 15–27.

Letón, E., and P. Zuluaga, 2008, "Unbalanced Groups in Nonparametric Survival Tests," *Statistics and Econometrics*, Statistics and Econometrics Series 15, Working Paper 08-52, Universidad Carlos III de Madrid, Madrid, Spain, October. [SRDB Ref ID: 113465]

Lohr, S. L., 2010, *Sampling: Design and Analysis*, Brooks/Cole, Boston, Massachusetts.

Millard S. P., and S. J. Deverel, 1988, "Nonparametric Statistical Methods for Comparing Two Sites Based on Data With Multiple Nondetect Limits," *Water Resources Research*, volume 24, number 12, pp. 2087–2098.

Monette, G., 1990, "Geometry of Multiple Regression and Interactive 3-D Graphics," Chapter 5 in *Modern Methods of Data Analysis*, edited by J. Fox and J. S. Long, Sage Publications, Newbury Park, California.

Noreen, E. W., 1989, *Computer Intensive Methods for Testing Hypotheses: An Introduction*, John Wiley & Sons, New York, New York.

ORAUT (Oak Ridge Associated Universities Team), 2005, *Analysis of Coworker Bioassay Data for Internal Dose Assessment*, ORAUT-OTIB-0019, Rev. 01, Oak Ridge, Tennessee, October 7.

ORAUT (Oak Ridge Associated Universities Team), 2006, *Generating Summary Statistics for Coworker Bioassay Data*, ORAUT-PROC-0095, Rev. 00, Oak Ridge, Tennessee, June 5.

ORAUT (Oak Ridge Associated Universities Team), 2007, *Internal Dose Reconstruction*, ORAUT-OTIB-0060, Rev. 00, Oak Ridge, Tennessee, February 6.

ORAUT (Oak Ridge Associated Universities Team), 2009a, *Use of Claimant Datasets for Coworker Modeling*, ORAUT-OTIB-0075, Rev. 00, Oak Ridge, Tennessee, May 25.

ORAUT (Oak Ridge Associated Universities Team), 2009b, *Analysis of Bioassay Data with a Significant Fraction of Less-Than Results*, ORAUT-RPRT-0044, Rev. 00, Oak Ridge, Tennessee, August 7.

ORAUT (Oak Ridge Associated Universities Team), 2010, *Discussion of Tritium Coworker Models at the Savannah River Site – Part 1*, ORAUT-RPRT-0049, Rev. 00, Oak Ridge Tennessee, November 23.

Peto, R., and J. Peto, 1972, "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society*, Series A, volume 135, number 2, pp. 185–207.

Prentice, R. L., 1978, "Linear Rank Tests with Right-Censored Data," *Biometrika*, volume 65, number 1, pp. 167–179.

Prentice, R. L., and P. Marek, 1979, "A Qualitative Discrepancy Between Censored Data Rank Tests," *Biometrics*, volume 35, number 4, pp. 861–867.

Rice, J. A., 2007, *Mathematical Statistics and Data Analysis*, Duxbury Press, Belmont, California.

Rousseeuw, P. J., I. Ruts, and J. W. Tukey, 1999, "The Bagplot: A Bivariate Boxplot," *American Statistician*, volume 53, number 4, pp. 382–387.

SAS (SAS Institute), 2011, *SAS/STAT User's Guide, Version 9*, Cary, North Carolina, http://support.sas.com/documentation.

Schochet, P. Z., 2008, *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations*, NCEE 2008-4018, Institute of Education Sciences, Washington, D.C., May. [SRDB Ref ID: 112834]

Singh, A., N. Armbya, and A. K. Singh, 2010, *ProUCL Version 4.1.00 Technical Guide (Draft)*, EPA/600/R-07/041, U.S. Environmental Protection Agency, Office of Research and Development, Washington, D.C., May. [SRDB Ref ID: 113464]

Wilk, M. B., and R. Gnanadesikan, 1968, "Probability Plotting Methods for the Analysis of Data," *Biometrika*, volume 55, number 1, pp. 1–17.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 1 of 13

The results in this attachment were calculated using the R code and data in LaBone (2012).

In Example 1, a group of *N* = 332 workers were monitored for a given radionuclide in urine, which yields 332 OPOS MPM statistics. There were n = 196 uncensored OPOS statistics. The ROS fit of the lognormal model to the data is shown in Figure A-1 (the *empirical quantiles* on the *y*-axis are the ordered OPOS results). The 50th percentile used in the coworker model is the GM and the 84th percentile is the product of the GM and GSD.

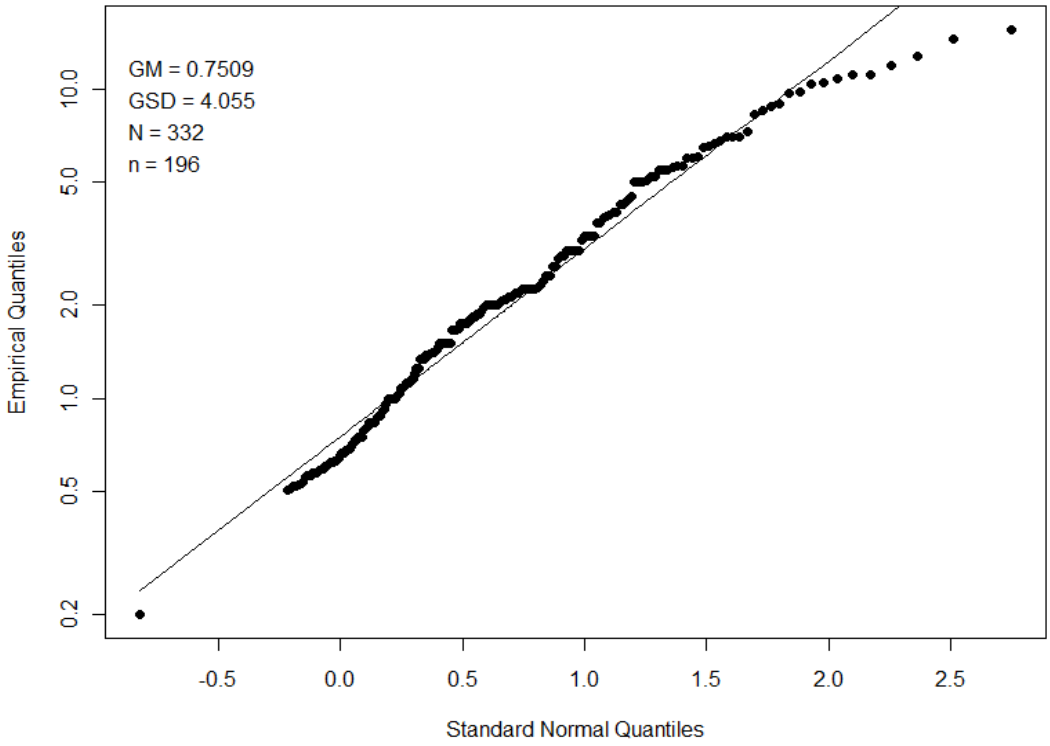

Figure A-1. ROS fit of lognormal distribution to the OPOS data for all monitored workers.

The data were stratified into two groups based on job descriptions. The lognormal plots for the strata are shown in Figures A-2 and A-3. At first glance, the GM and GSD for the two fits do not appear to be significantly different. This is confirmed by the Monte Carlo permutation tests in Figure A-4 (bivariate normal ellipse) and Figure A-5 (nonparametric bagplot). The EDF plots for the two strata are shown in Figure A-6 along with the results of the Peto-Prentice Test (p = 0.17), which indicates there is no reason to believe these curves are different (i.e., the strata are not different). Thus, in this case the coworker model shown in Figure A-1 would be appropriate to use because the strata are not significantly different.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 2 of 13

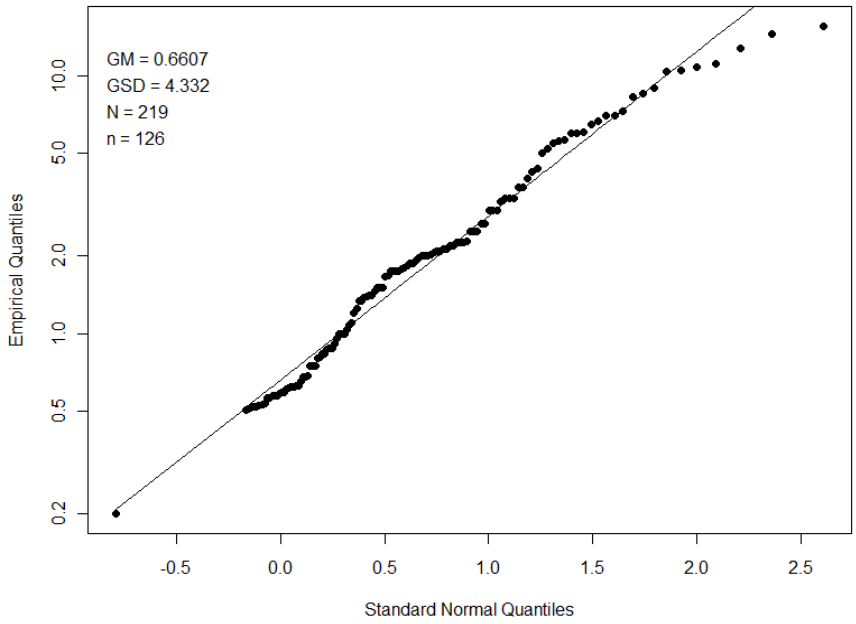

Figure A-2. Coworker model for Stratum A based on job title.

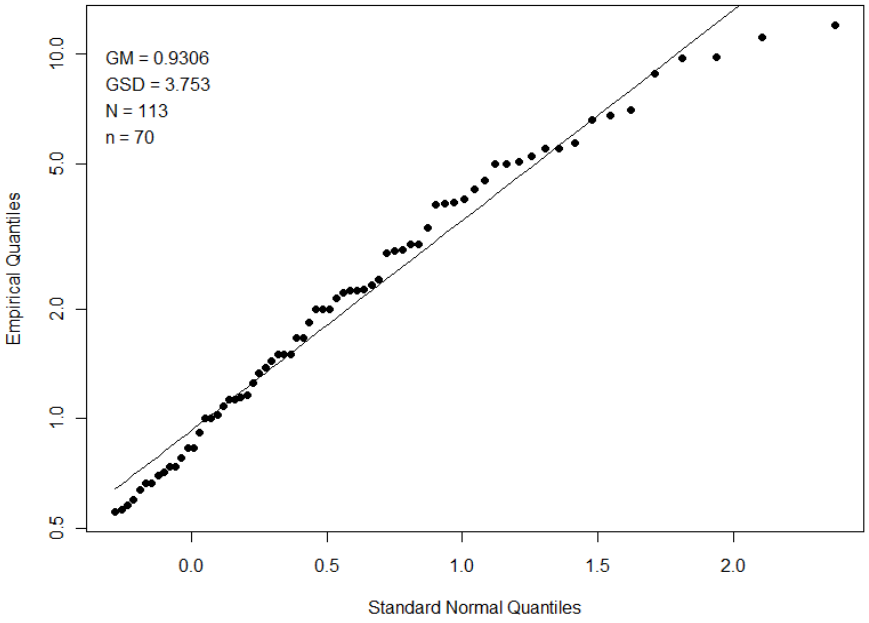

Figure A-3. Coworker model for Stratum B based on job title.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 3 of 13

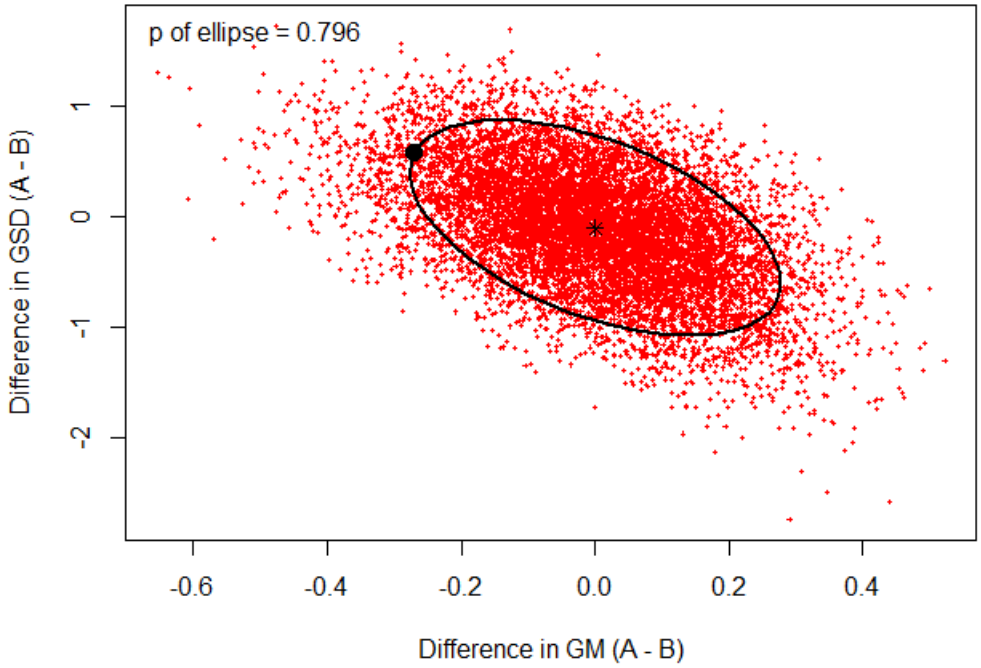

Figure A-4. Monte Carlo permutation test of coworker models for Stratum A and Stratum B (job title) using a bivariate normal probability ellipse. The large dot on the ellipse corresponds to the observed differences in GM and GSD. The p-value of the test is 1 - 0.796 = 0.204, which means the strata are not significantly different at the $\alpha = 0.05$ confidence level.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 4 of 13

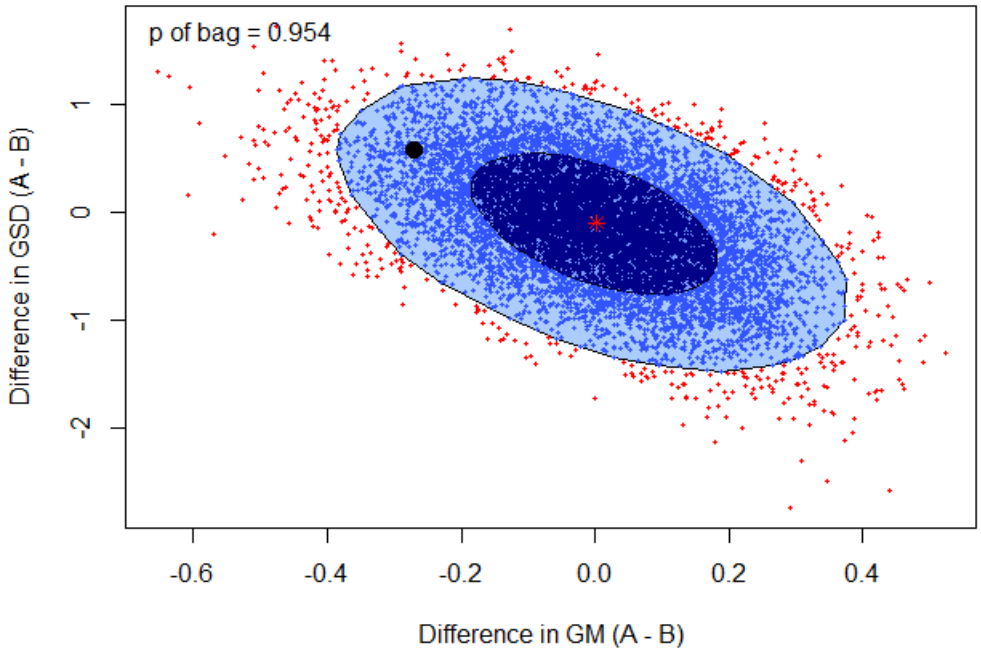

Figure A-5. Monte Carlo permutation test of coworker models for Stratum A and Stratum B (job title) using a nonparametric bagplot probability polygon. The outer polygon encloses ~95% of the points and the inner polygon encloses ~50% of the points. The strata are not significantly different at the α = 0.05 confidence level. The large dot inside the outer polygon corresponds to the observed differences in GM and GSD.

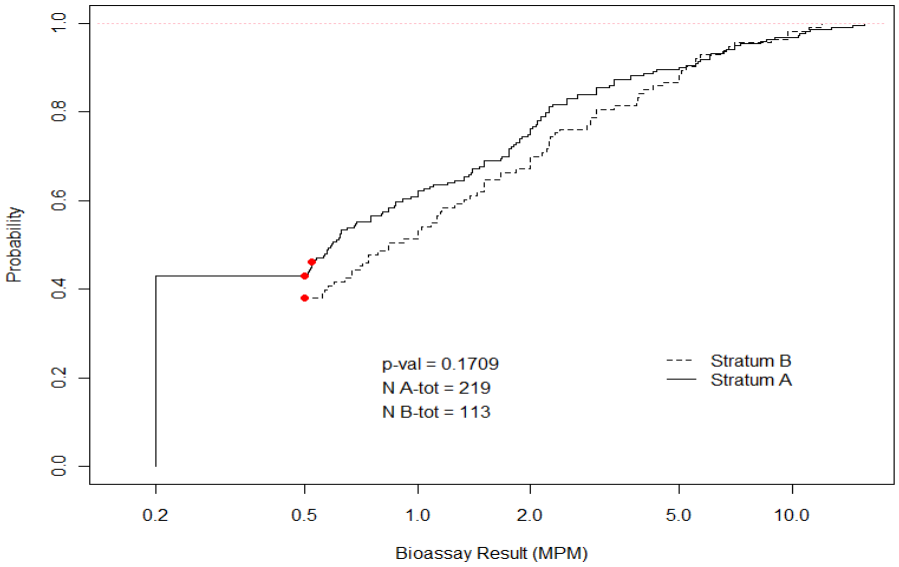

Figure A-6. EDF plots and results of Peto-Prentice Test on OPOS statistics for Stratum A and Stratum B (job title). The strata are not significantly different at the α = 0.05 confidence level.

Next, the same data were stratified into two groups based on work location. The lognormal plots for the strata are shown in Figures A-7 and A-8. At first glance, the GM for the two fits appears to be significantly different.

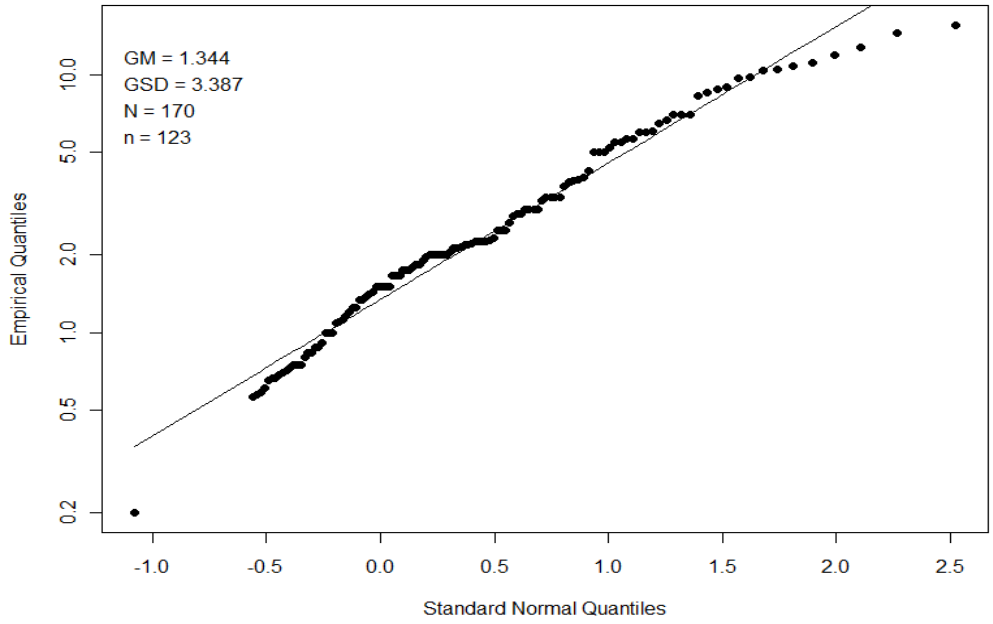

Figure A-7. Coworker model for Stratum A based on work area.

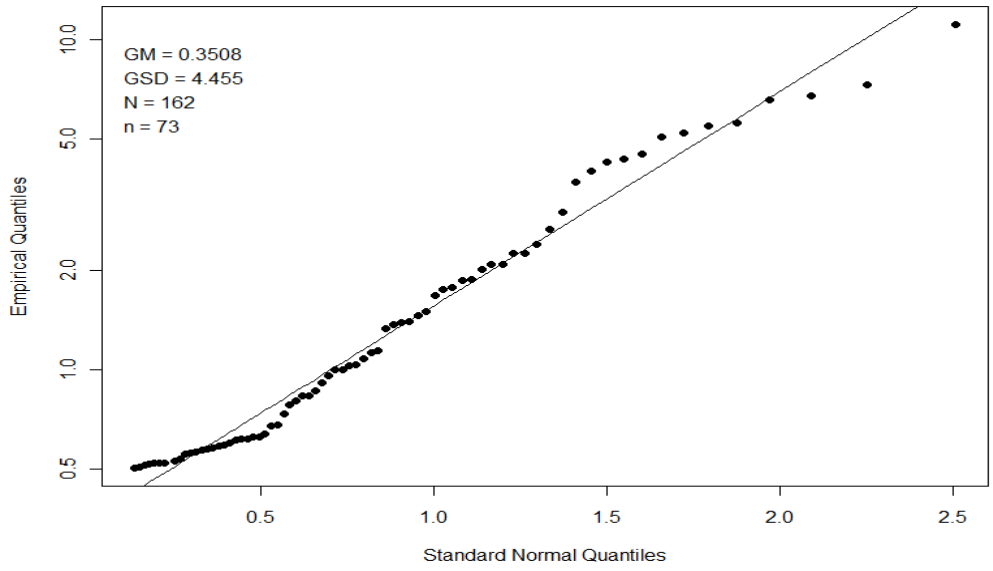

Figure A-8. Coworker model for Stratum B based on work area.

This is confirmed by the Monte Carlo permutation tests in Figure A-9 (bivariate normal ellipse) and Figure A-10 (nonparametric bagplot).

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 6 of 13

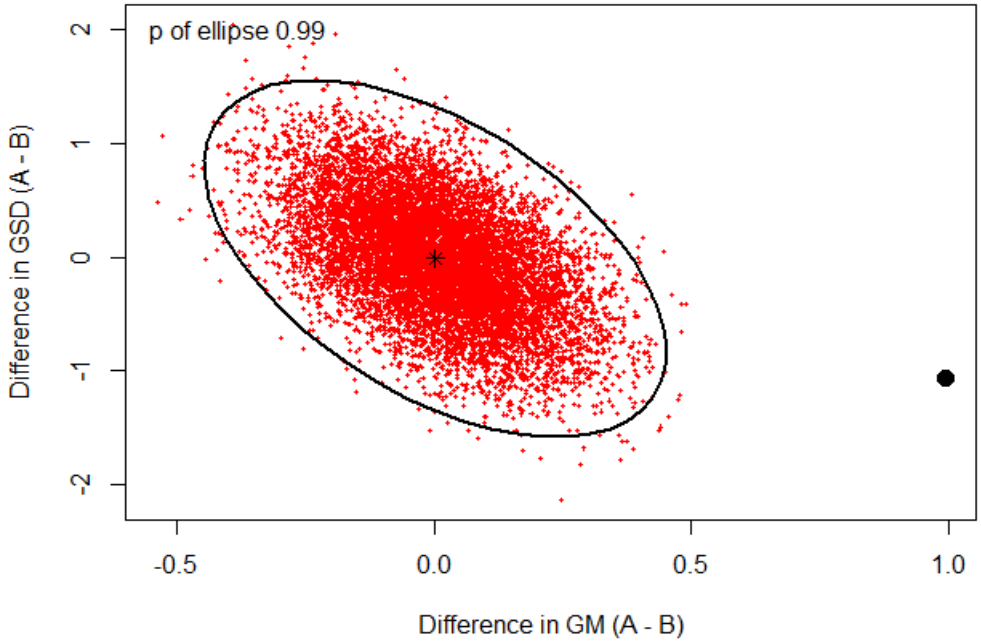

Figure A-9. Monte Carlo permutation test of coworker models for Stratum A and Stratum B (work area) using a bivariate normal probability ellipse. The large dot in the lower right corner corresponds to the observed differences in GM and GSD. The strata are significantly different at the $\alpha = 0.05$ confidence level.

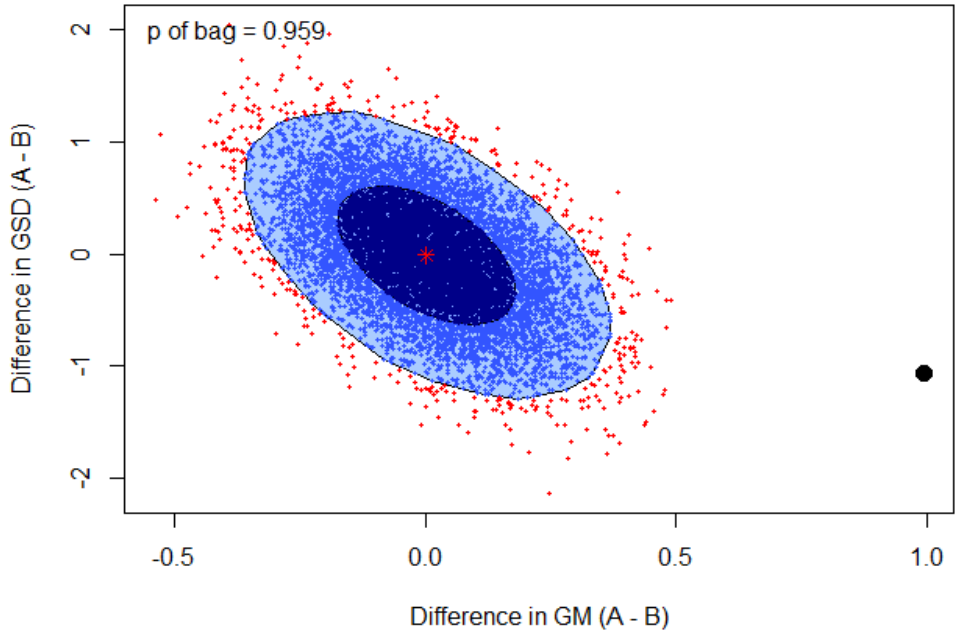

Figure A-10. Monte Carlo permutation test of coworker models for Stratum A and Stratum B (work area) using nonparametric bagplot probability polygon. The outer polygon encloses ~95% of the points whereas the inner polygon encloses ~50% of the points. The strata are significantly different at the $\alpha = 0.05$ confidence level. The large dot in the lower right corner corresponds to the observed differences in GM and GSD.

The GSDs of the two fits are not significantly different, whereas the GMs are. The survival curves for the two strata are shown in Figure A-11 along with the results of the Peto-Prentice Test (p = 0), which indicates the EDF plots are significantly different. Thus, in this case the coworker model shown in Figure A-7 would be considered for workers in Stratum A and the coworker model shown in Figure A-8 would be considered for workers in Stratum B because the strata are significantly different. The final decision on whether to use the stratified coworker models would be based on the practical significance of the difference between the strata.

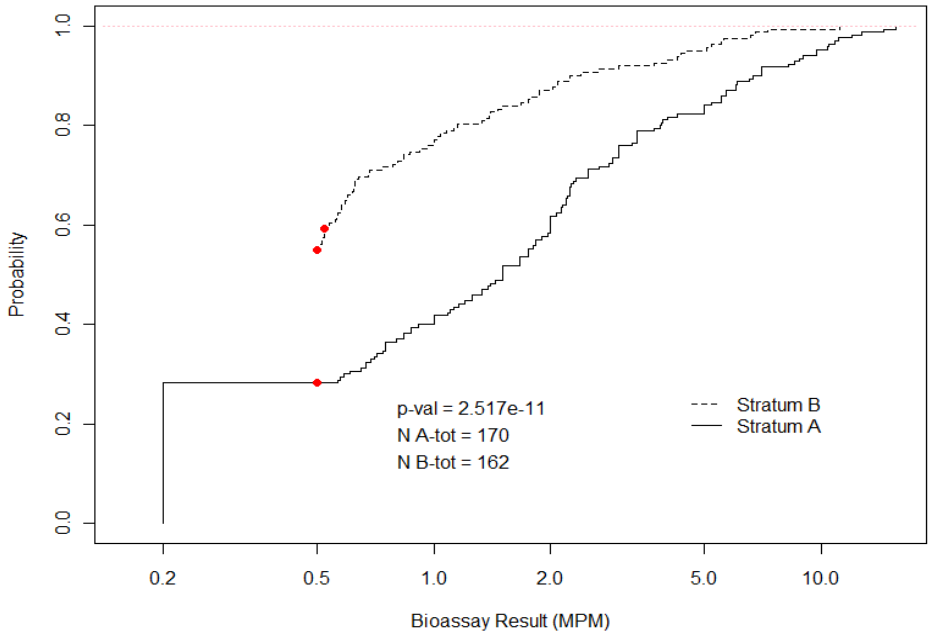

Figure A-11. EDF plots and results of Peto-Prentice Test on OPOS statistics for Stratum A and Stratum B (work area). The strata are significantly different at the α = 0.05 confidence level.

In Example 2, a group of *N* = 1651 workers were monitored for a radionuclide in urine by gross alpha counting, which yielded 1,651 OPOS MPM statistics. There were *n* = 111 uncensored OPOS statistics. The ROS fit of the lognormal model to the data is shown in Figure A-12 (the *empirical quantiles* on the *y*-axis are the ordered OPOS results). The 50th percentile used in the coworker model is the GM and the 84th percentile is the product of the GM and GSD. The GSD = 9.088 is large and ~93% of the data are censored, which suggests this is a bimodal distribution (i.e., a mixture of two distributions). The fits of lognormal distributions to Stratum A and Stratum B, which are based on job titles, are shown in Figures A-13 and A-14. At first glance, the GM and GSD of the coworker models based on the strata do not appear to be significantly different. The EDF plots for the two strata are shown in Figure A-15 along with the results of the Peto-Prentice Test (p = 0.093), which indicates the survival curves are not significantly different. Thus, the coworker model in Figure A-12 is considered to be appropriate for all workers in the population of monitored workers. However, the standard ROS methods might not be appropriate because the distribution of the data appears to be bimodal. This issue is addressed by performing an effective fit to the data, which is discussed next.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 9 of 13

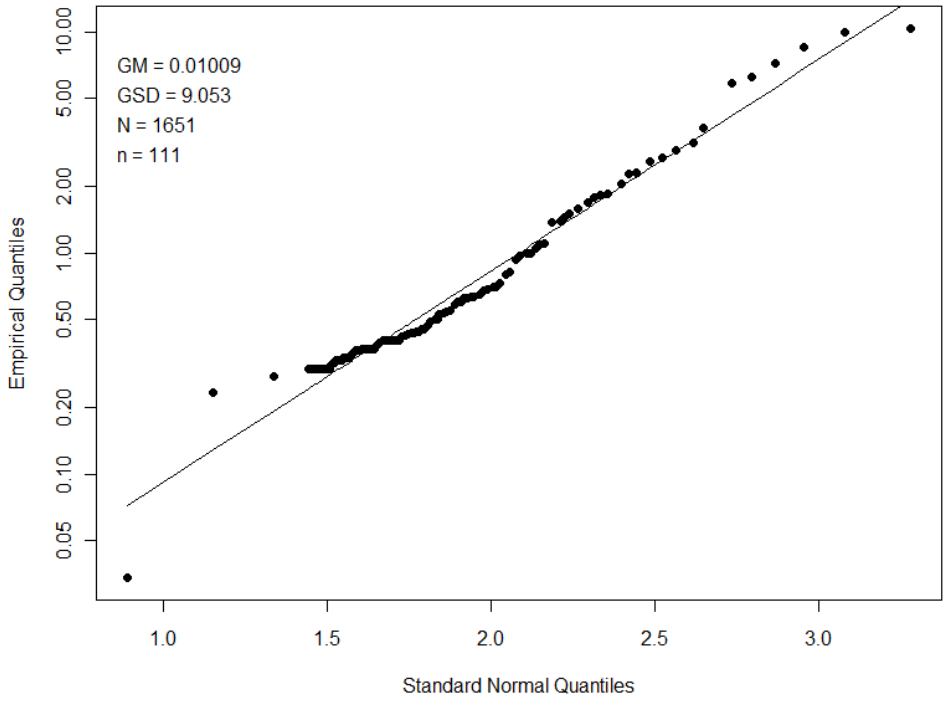

Figure A-12. ROS fit of lognormal distribution to the OPOS data for all monitored workers.

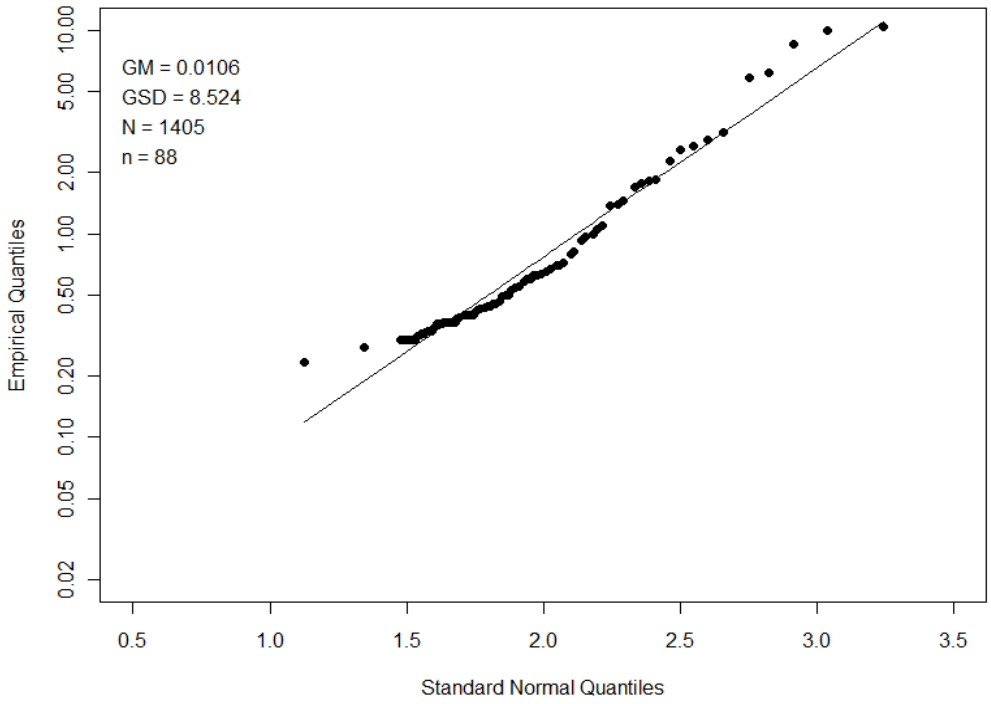

Figure A-13. Coworker model for Stratum A based on job title.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 10 of 13

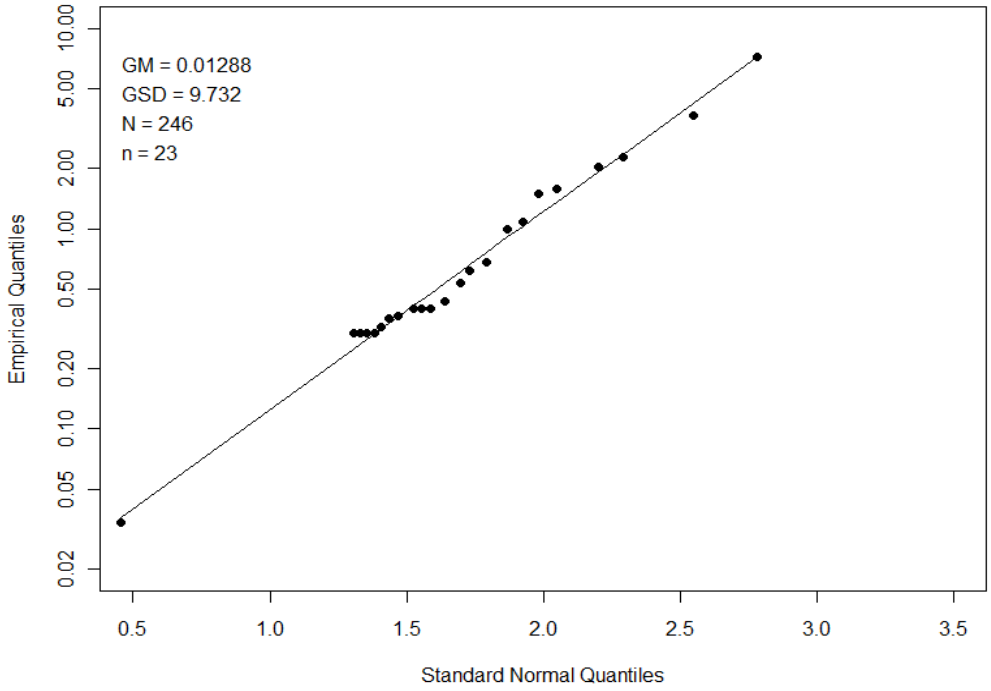

Figure A-14. Coworker model for Stratum B based on job title.

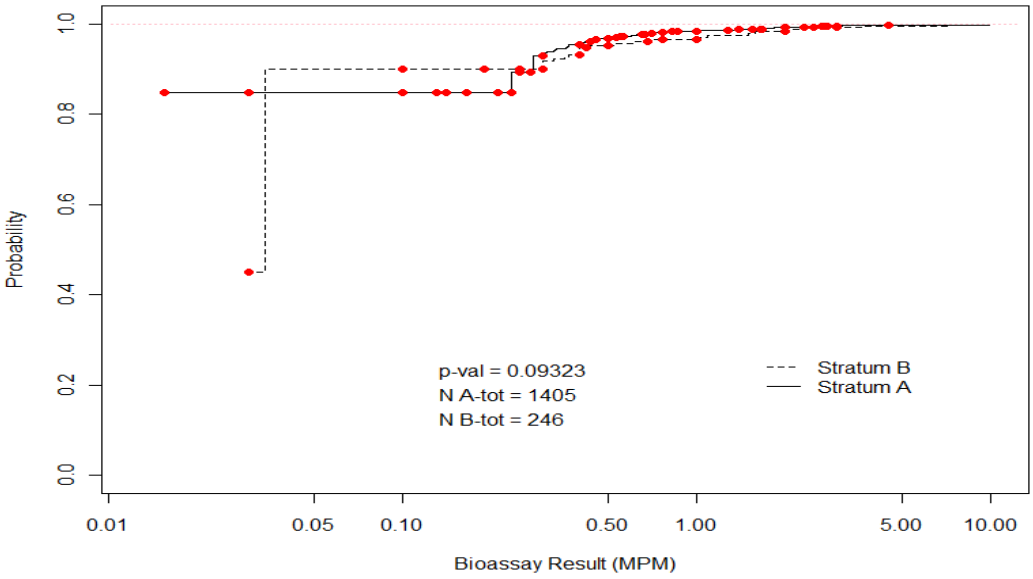

Figure A-15. EDF plot and results of Peto-Prentice Test on OPOS statistics for Stratum A and Stratum B (based on job title). The strata are not significantly different at the α = 0.05 confidence level. The points are the values of the censoring levels applied to the data.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 11 of 13

An effective fit to the data was performed by first using the maximum likelihood methods described in ORAUT-RPRT-0044 (ORAUT 2009b) to model the data from all monitored workers, yielding a mixture[12] of distributions where 98.1% of the population has one lognormal distribution (with GM = 0.05 and GSD = 3) and 1.9% of the population has another lognormal distribution (with GM = 1.79 and GSD = 2.65). A probability-probability (P-P) plot[13] of the fit is shown in Figure A-16, which indicates that the model fits the data well.

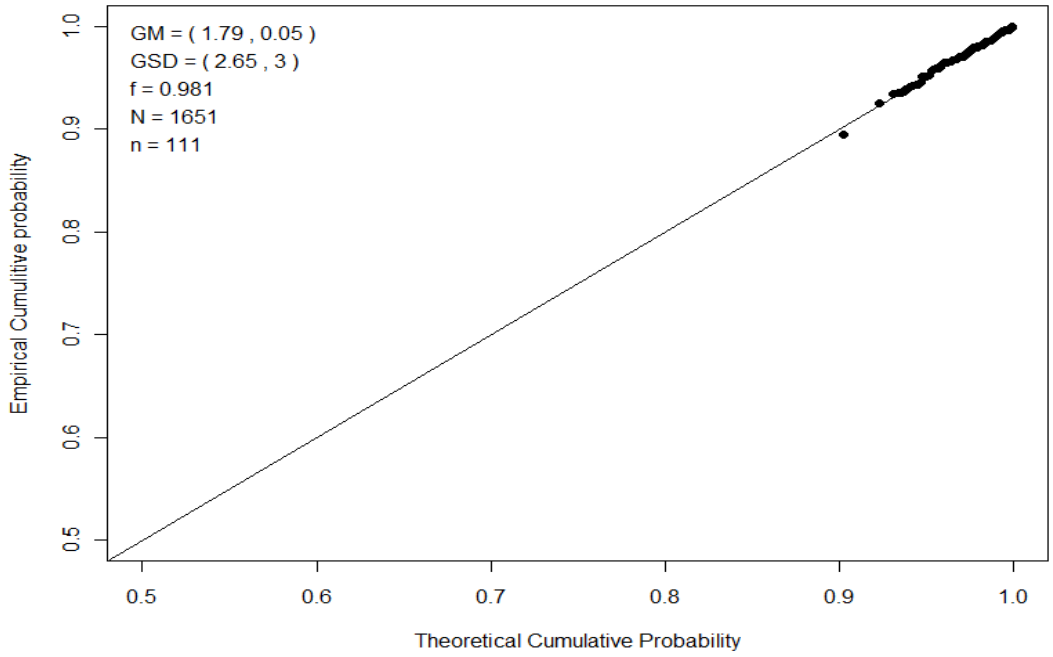

Figure A-16. P-P plot where the points are the uncensored data. The GM and GSD each have two values, one for each distribution in the mixture. For example, the first distribution has a GM = 1.79 and a GSD = 2.65. If the model fits the data, the data should fall on a straight line through the origin with a slope of 1.

The average probabilities of the censored data were imputed using this bimodal model (see P-P plot in Figure A-17) and then these probabilities were converted[14] to quantiles (bioassay results). The imputed quantiles (bioassay results) were then combined with the uncensored quantiles and a single lognormal distribution was fit to them. This effective fit is shown in Figure A-18.

12 A lognormal-lognormal model is fit to the data in this example. The code (LaBone 2012) includes an example of fitting a normal-lognormal model to the data.

13 In a P-P plot the empirical cumulative probability of the data is plotted on the *y*-axis and the theoretical probability of the data (as calculated with the bimodal lognormal distribution) is plotted on the *x*-axis. See Wilk and Gnanadisikan (1968) for a more detailed discussion of P-P plots.

14 This is basically the same process used by Helsel (2005, p. 68) in the ROS fitting technique.

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 12 of 13

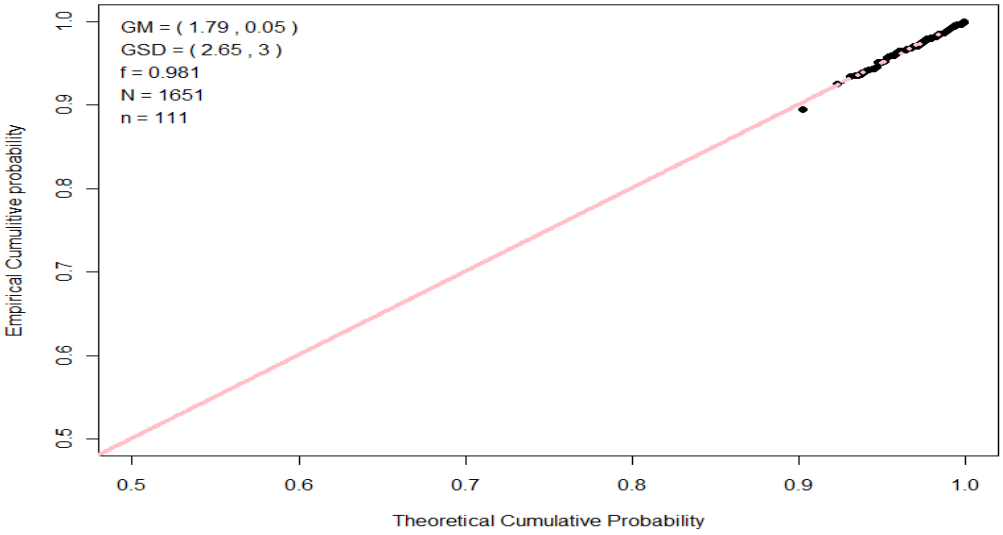

Figure A-17. P-P plot where the dark points are the uncensored data and the smaller light points are the imputed values for the censored data. The GM and GSD each have two values, one for each distribution in the mixture.

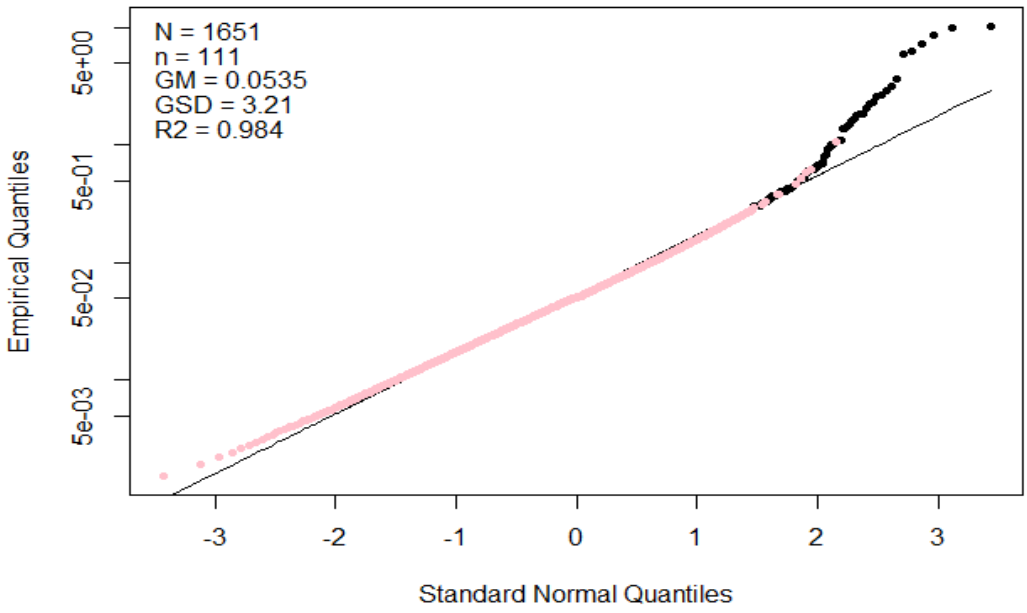

Figure A-18. Effective fit to the uncensored data (dark points) and imputed values of the censored data (light points).

ATTACHMENT A
EXAMPLES OF FITTING METHODS
Page 13 of 13

In Example 3, a group of N = 28 workers were monitored for a radionuclide in urine by gross alpha counting, which yielded a total of 28 OPOS MPM statistics. There were n = 1 uncensored OPOS statistics and k = 27 censored statistics. The censoring level is 2. As discussed in ORAUT-RPRT-0044 (ORAUT 2009b), a single censoring level is required to perform the binomial fit. The probability pCL associated with the censoring level is estimated to be

$$pCL = \frac{27}{28} = 0.96429\,, \tag{A-1}$$

i.e., the CL = 2 is at ~96th percentile. Assuming a GSD = 1.55 based on a normally distributed analytical background with mean of 0, the GM of the distribution is estimated using the following relationship:

$$GM = CL \times GSD^{-qnorm(pCL)}\,, \tag{A-2}$$

thus

$$GM = 2(1.55)^{-1.802743} = 0.908 \tag{A-3}$$

$$GM \times GSD = 1.407\,. \tag{A-4}$$

These points are depicted in Figure A-19. As discussed in the body of this report, it is not feasible to stratify this dataset because of its size. Even if it could be stratified, a comparison of the strata is not considered because of the high degree of censoring.

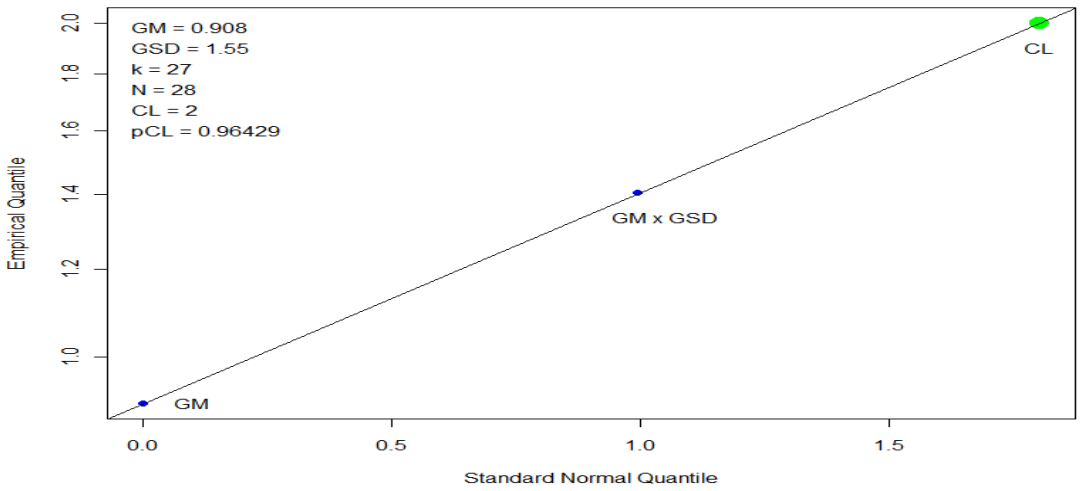

Figure A-19. ORAUT-RPRT-0044 binomial fit to highly censored urine data (ORAUT 2009b).

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA

The Peto-Prentice Test was selected to use for comparing strata at Step 2 in Figure 2-1 (the OPOS data). Reasons for selecting this test are discussed in this attachment.

One method to compare two groups is by using a nonparametric approach. Nonparametric methods do not require an assumption that data follow a specific distribution (like lognormal or normal distributions), and they use no information on the shape of the distribution in conducting the tests. In the nonparametric tests considered here, the data are ranked, providing information on the relative positions of each observation. Tests determine if one group generally has more frequent high or low values. For censored data, positions are represented by scores, which are ranks of data adjusted for the missing information because some values are censored. Tests based on scores (i.e., score tests) are used to determine if the distribution functions differ among groups of censored data. Score tests are designed to handle data censored at multiple detection limits, using the information in detected values between detection limits in addition to the information in the proportion of values below each detection limit. Some of these tests are designed to handle ties in the censored data, in the uncensored data, or even between censored and uncensored data.

Some score tests are direct extensions of the Wilcoxon Rank-Sum Test, including the Gehan Test, the Generalized Wilcoxon-Gehan Test, and the Peto-Prentice Test. If uncensored data are used for these Wilcoxon-type tests, the results are identical to the Wilcoxon Rank-Sum Test (and Mann-Whitney U Test). Another family of score tests used for censored data, such as the Logrank Test, the Peto-Peto Test, the Tarone-Ware family of tests, and the Fleming- Harrington family of tests were originally developed in the survival analysis field for right-censored data, based on the Kaplan-Meier (KM) nonparametric methods, but were easily adapted and implemented in the field of environmental monitoring. To apply the score tests to a left-censored dataset, the data values must be flipped first by subtracting each value from a constant chosen to be larger than the maximum value in the dataset; once the data are converted from the left-censored format to the right-censored format, all score tests from the survival analysis field can be applied directly to this right-censored dataset.

The tests from the survival analysis field were designed to compare the survival curves in the two groups. The null and alternative hypotheses being tested in this case are the following:

$$\begin{aligned} H_0 &: S_1(t) = S_2(t), \\ H_1 &: S_1(t) \neq S_2(t), \end{aligned} \tag{B-1}$$

where

$S_1(t) = 1 - F_1(t) = P(X > t), S_2(t) = 1 - F_2(t) = P(Y > t)$ denote the survival functions for the two groups,

and

$F_1(t) = P(X \leq t), F_2(t) = P(Y \leq t)$ denote the cumulative distribution functions (CDFs) for the two groups (X and Y represent the random variables associated with the observations in each of the two groups).

For the left-censored data, the CDFs for the two groups are compared, with the null and alternative hypotheses tested as follows:

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA

$$H_0 : F_1(t) = F_2(t), \quad H_1 : F_1(t) \neq F_2(t), \tag{B-2}$$

which is equivalent to stating that the distribution functions in the two populations are equal at all values greater than the minimum observed value (Klein and Moeschberger 2003, p. 206).

The most common score test used for comparing two groups with censored data is the Gehan Test, introduced by Gehan (1965) as a generalization of the Wilcoxon Rank-Sum Test, when censored data are present. In the Gehan Test, the observations from the two groups are combined and then ranked to construct the test statistic. If we denote by x_i, $i = 1,\ldots,m$ the values from the first sample, by y_j, $j = 1,\ldots,n$ the values from the second sample, and by v_k, k=1,…,m + n the values from the combined sample (both groups), the Gehan score for an observation in the combined sample is the number of observations in the combined sample that are definitely greater than the observation in question, minus the number of observations definitely smaller than this observation (Helsel 2005, p. 145; Higgins 2004, p. 233; Desu and Raghavarao 2004, p. 114):

$$h_k = (\#\text{ of } v_l \text{ that are definitely} > v_k) - (\#\text{ of } v_l \text{ that are definitely} < v_k). \tag{B-3}$$

The sum of all Gehan scores in the combined sample is equal to zero. The Gehan Test for comparing the two groups with censored values is a permutation test applied to the Gehan scores. If no observations are censored, the test is equivalent to the Wilcoxon Rank-Sum Test and the Mann-Whitney U Test.

An alternative method to perform the Gehan Test is to compute a test statistic W, as the sum of all U_{ij} scores obtained from the $m \times n$ matrix U with all the possible pair comparisons of values from the two samples. The $m \times n$ matrix U of scores is defined as (Helsel 2005, p. 143):

$$U_{ij} = \begin{cases} -1, \text{ if } x_i > y_j \text{ } (y_j \text{ may be a nondetect}) \\ 1, \text{ if } x_i < y_j \text{ } (x_i \text{ may be a nondetect}) \\ 0, \text{ if } x_i = y_j, \text{ or for indeterminate comparisons.} \end{cases} \tag{B-4}$$

and the test statistic is then computed as:

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} U_{ij}. \tag{B-5}$$

If h_k is the Gehan score attached to the k^{th} observation in the combined sample (as defined above), the variance of W, also called the *permutation* variance, is computed as:

$$Var(W) = \frac{mn \sum_{k=1}^{m+n} h_k^2}{(m+n)(m+n-1)}. \tag{B-6}$$

The standardized statistic

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA
Page 3 of 7

$$Z_W = \frac{W}{\sqrt{Var(W)}} \quad \text{(B-7)}$$

is compared with the corresponding value from the standard normal distribution to determine the statistical significance.

The Gehan Test is widely applied and implemented in most software packages, including SAS, R, StatXact, NCSS, and the ProUCL package (Singh, Armbya, and Singh 2010). The Gehan Test is widely used and has a high power when the data are lognormally distributed with equal variance in both groups but with different means. However, there are some situations in which it might not be the most appropriate test to use. Prentice and Marek (1979) showed that the Gehan Test has little power to detect differences between the groups when censoring rates are high because the test statistic is dominated by a small number of large observations.

There are two ways to define the tests that are used for the two group comparisons with censored data. One way is to define them as score tests, which assign scores to each ordered observation in the combined sample; the second way is to define them as weighted tests, which are based on weights that are assigned in the context of a contingency table, for each uncensored observation. There is an equivalence between the score tests and weighted tests (Desu and Raghavarao 2004, p. 118; Leton and Zuluaga 2001, pp. 597–599), with the test statistic obtained from a score test, being the same as the test statistic from a weighted test. The difference between the score tests and the weighted tests is that they use slightly different variances and, as a consequence, the reported z-scores and the associated p-values can be slightly different. The score tests use the *permutation variance*, which is appropriate when censoring patterns are the same in the two groups. If the censoring mechanisms in the two groups are different, the permutation variance is not appropriate, and the *hypergeometric variance* (also called *conditional permutation variance*) should be used, because it is valid in all the cases (Leton and Zuluaga 2001, p. 594; Lawless 1982, pp. 419–425). The hypergeometric variance is always smaller than the permutational variance, so the Z-value of the weighted test will be greater than the value of the score test (Leton and Zuluaga 2001, p. 594). The weighted tests tend to underestimate the variance, producing tests with higher power (anticonservative), so if one can assume independence between the censoring and groups, the tests to be used should be the score tests (Leton and Zuluaga 2002, p. 24). In general, it appears that unless censoring patterns differ somewhat between the two groups, or if the sample sizes are very small, the two variances should be in close agreement (Lawless 1982, p. 422). Sometimes, a third type of variance, the *asymptotic variance* (Brown 1984; Millard and Deverel 1998, p. 2092; Latta 1981, p. 714), is used for these types of tests. The asymptotic variance is equal to either the permutation variance or the hypergeometric variance under certain conditions (Brown 1984); for example, if most of the uncensored observations occur very shortly after their censoring times, the hypergeometric variance and the asymptotic variance are very close together (Brown 1984, p. 71).

To introduce a general definition for the weighted tests, denote by t_j, $j = 1,\ldots,k$ the uncensored observations in the combined (both groups) right-censored sample, and let m_{j1} (m_{j2}) be the multiplicity in the first (second) group, r_{j1} (r_{j2}) be the number of observations in the first (second) group that are greater than or equal to t_j, and $m_j = m_{j1} + m_{j2}$, and $r_j = r_{j1} + r_{j2}$. The statistic for the weighted test is computed as (Desu and Raghavarao 2004, pp. 118–119):

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA
Page 4 of 7

$$U = \sum_{j=1}^{k} w(r_j)\left[m_{j2} - r_{j2}\frac{m_j}{r_j} \right], \tag{B-8}$$

where *w(.)* is a weight function, which differs from test to test. The variance of the weighted test, also called the *hypergeometric variance*, is computed as:

$$Var(U) = \sum_{j=1}^{k} w^2(r_j)\frac{r_{j1}r_{j2}}{r_j^2(r_j - 1)} m_j(r_j - m_j), \tag{B-9}$$

and the standardized statistic is computed as:

$$Z_U = \frac{U}{\sqrt{Var(U)}}. \tag{B-10}$$

One of the main differences between all the tests mentioned above is how they weigh the observations in the combined sample; for example, the Logrank Test assigns equal weights to all the observations:

$$w(r_j) = 1, \tag{B-11}$$

the Gehan Test places very heavy weight on large observations:

$$w(r_j) = r_j, \tag{B-12}$$

the Tarone-Ware Test places heavy weight on large observations, but the weights are smaller in comparison with the weights used by the Gehan Test:

$$w(r_j) = \sqrt{r_j}, \tag{B-13}$$

while the Peto-Peto and Peto-Prentice Tests assign just a little bit more weight on large observations than on small observations (Klein and Moeschberger 2003, p. 211).

In general, the various weightings should provide similar results, and will usually lead to the same decision on whether the null hypothesis is rejected. Although most of the software packages will provide by default the results of multiple tests, only one test chosen *a priori* should be used in deciding the significance between the two groups. The choice of which test to use should be based on the statistical power of each test, the best test being the one with the most power (i.e., the test that is more likely to reject a false null hypothesis).

There is a lot of discussion in the literature on the topic of choosing the most powerful test, but there is not a clear winner in the sense that it has the most power in all possible scenarios (including situations with different sample sizes between the two groups, different censoring proportions in each of the groups, different censoring mechanisms in the groups, etc.). Helsel (2005, p. 150) mentions

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA

that the Peto-Prentice Test with asymptotic variance is most powerful when underlying data are lognormal. The Peto-Prentice Test is a modification of the Gehan Test, based on improvements suggested by Peto and Peto (1972) and Prentice (1978). Helsel's recommendation is based primarily on papers by Latta (1981) and Millard and Deverel (1988).

Latta (1981, p. 719) gives specific recommendations for two different scenarios:

- Situations with heavy censoring and samples sizes that are far apart, or censoring mechanisms that differ greatly; in this case the Peto-Prentice Test with asymptotic variance is the one recommended, being the most powerful;

- Situations with heavy censoring, approximately equal sample sizes, and censoring mechanisms that are similar; in this case the Gehan Test and the Peto-Prentice Test are the most powerful, when the data follow a lognormal distribution.

At the end of his article, Latta (1981, p. 719) concluded that the test with the overall best performance is the Peto-Prentice Test with the asymptotic variance estimate.

Millard and Deverel (1988, pp. 2095-2096) showed that the Peto-Prentice Test with the asymptotic variance is the most powerful (or almost the most powerful among the tests analyzed) in two different scenarios:

- When sample sizes and censoring proportions in the two groups are the same;

- When the sample size in the first group is smaller than the sample size in the second group, and the censoring proportion in the first group is larger than the censoring proportion in the second group.

Millard and Deverel (1988, p. 2097) also concluded that the Peto-Prentice Test with the asymptotic variance might be the most powerful test for cases in which sample sizes and censoring mechanisms do not differ greatly between the two groups.

Based on these recommendations (Latta 1981; Millard and Deverel 1988; Helsel 2005), it was decided to use in our analysis the Peto-Prentice Test with the asymptotic variance because it exhibits the most power if the underlying data are lognormal and in situations involving different sample sizes and censoring proportions between the two groups.

The Peto-Prentice score test is described in detail by Helsel (2005, p. 146). Scores for the Peto-Prentice Test are a weighted version of the scores used in the Gehan Test, adjusting the *U* scores from the Gehan Test by the survival function at each observation to create a new score. The survival function for an uncensored observation t_j is computed according to the KM Product-Limit estimator (Helsel 2005, p. 65; Desu and Raghavarao 2004, p. 32; Leton and Zuluaga 2001, p. 594):

$$S(t_j) = \prod_{i=1}^{j} \frac{r_i - m_i}{r_i}. \qquad \text{(B-14)}$$

The scores for the Peto-Prentice Test are computed as follows (Helsel 2005, p. 146):

$$U_i = S(t_i) + S(t_{i-1}) - 1 \text{ for all uncensored observations } t_i, \quad \text{(B-15)}$$

$$U_j = S(t_{i-1}) - 1 \text{ for all censored observations } t_j^* \text{ (where } t_{i-1} \le t_j^*), \quad \text{(B-16)}$$

where

$S(t_i)$ = the value of the survival function for the uncensored observation t_i,
$S(t_{i-1})$ = the value of the survival function for the previous uncensored observation t_{i-1}, and
$S(t_0)$ = 1

The sum of scores for all the observations in the combined sample is equal to zero. The *W* statistic for the Peto-Prentice Test is the sum of scores for the second group:

$$W = \sum_{i=1}^{n} U_i \,. \quad \text{(B-17)}$$

The permutation variance of *W* is computed as:

$$Var(W) = \frac{mn\sum_{i=1}^{m+n} U_i^2}{(m+n)(m+n-1)}, \quad \text{(B-18)}$$

the standardized statistic of the Peto-Prentice Test is computed as:

$$Z_W = \frac{W}{\sqrt{Var(W)}}, \quad \text{(B-19)}$$

and the associated two-sided p-value is computed as:

$$P = \text{Prob}(|Z| > Z_W), \quad \text{(B-20)}$$

where $Z \sim N(0,1)$. Some software packages will report, rather than the Z_W statistic, a chi-square statistic, which is approximately equal to the square of the Z_W statistic, and the p-values obtained from the two approaches should be about the same (Helsel 2005, p. 147).

The Peto-Prentice weight test is obtained using the general approach described above (see equations B-8 to B-9), and with the weight based on the survival time for the previous uncensored observation in the combined sample, as a way to ensure that these weights are known just before the time at which the comparison is to be made (Leton and Zuluaga 2001, p. 600; Klein and Moeschberger 2003, p. 208):

$$w(r_j) = S(t_{j-1}) = \prod_{i=1}^{j-1} \frac{r_i - m_i}{r_i} \,. \quad \text{(B-21)}$$

In case of no censoring, the Peto-Prentice statistic reduces to one that is equivalent to the Wilcoxon Rank-Sum (or Mann-Whitney U Test) statistic (Millard and Deverel 1988, p. 2091). In practice, there is little difference between the p-value reported by the Peto-Prentice score test (using the permutation variance), the p-value reported by the Peto-Prentice weight test (using the hypergeometric variance),

ATTACHMENT B
DISCUSSION OF THE TWO-GROUPS COMPARISON TESTS FOR CENSORED DATA

or the Peto-Prentice Test with the asymptotic variance, so they should all produce the same significance result.

Most of the commercial software packages (SAS, R, S-Plus, SPSS, Stata, BMDP) implement the weight versions of the two-group comparison tests (Leton and Zuluaga 2001, 2008). The Peto-Prentice Test weight test is implemented in several software packages, including the R package NADA (Helsel and Lee 2010), as well as in PROC LIFETEST in the SAS System software (SAS 2011), as the Flemming-Harrington Test with $\rho_1 = 1$ and $\rho_2 = 0$.