Draft

ADVISORY BOARD ON
RADIATION AND WORKER HEALTH

National Institute for Occupational Safety and Health

DRAFT REVIEW OF ORAUT-RPRT-0053:
ANALYSIS OF STRATIFIED COWORKER DATASETS, REV. 1

Contract No. 200-2009-28555
SCA-TR-PR2013-0053, Revision 0

Prepared by

H. Chmelynski
J. Lipsztein
S. Marschke
J. Stiver
S. Cohen & Associates
1608 Spring Hill Road, Suite 400
Vienna, VA   22182

April 2013

Disclaimer

This document is made available in accordance with the unanimous desire of the Advisory Board on Radiation and Worker Health (ABRWH) to maintain all possible openness in its deliberations. However, the ABRWH and its contractor, SC&A, caution the reader that at the time of its release, this report is pre-decisional and has not been reviewed by the Board for factual accuracy or applicability within the requirements of 42 CFR 82. This implies that once reviewed by the ABRWH, the Board's position may differ from the report's conclusions. Thus, the reader should be cautioned that this report is for information only and that premature interpretations regarding its conclusions are unwarranted.

S. COHEN & ASSOCIATES: ***Technical Support for the Advisory Board on Radiation & Worker Health Review of NIOSH Dose Reconstruction Program***	Document No. SCA-TR-PR2013-0053 Effective Date: Draft – April 23, 2013
DRAFT REVIEW OF ORAUT-RPRT-0053: ANALYSIS OF STRATIFIED COWORKER DATASETS, REV. 1	Page 2 of 55
Task Manager: _________________________ Date: ____________ Stephen F. Marschke	Supersedes: N/A
Project Manager: _________________________ Date: ____________ John Stiver, MS, CHP	Reviewers: A. Makhijani, PhD

Record of Revisions

Revision Number	Effective Date	Description of Revision
0 (Draft)	04/23/2013	Initial issue

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOTICE: This report has been reviewed for Privacy Act information and has been cleared for distribution. However, this report is pre-decisional and has not been reviewed by the Advisory Board on Radiation and Worker Health for factual accuracy or applicability within the requirements of 42 CFR 82.

ABBREVIATIONS AND ACRONYMS

AMW	All Monitored Workers
CERCLA	Comprehensive Environmental Response Compensation and Liability Act
CTW	Construction Trades Worker
DQO	Data Quality Objectives
EPA	U.S Environmental Protection Agency
GM	Geometric Mean
GSD	Geometric Standard Deviation
LBGR	Lower Bound of the Gray Region
MARSSIM	Multi-Agency Radiation Survey and Site Investigation Manual
MCPT	Monte Carlo Permutation Test
MDA	Minimum Detectable Amount
MDD	Minimum Detectable Difference
MPM	Maximum Possible Mean
nCi	nanocurie
NIOSH	National Institute for Occupational Safety and Health
non-CTW	Non-Construction Trades Worker
NRC	U.S. Nuclear Regulatory Commission
OPOS	One-Person, One Sample
ORAUT	Oak Ridge Associated Universities Team
OTIB	ORAUT Technical Information Bulletin
PROC	Procedure
ROS	Regression on Order Statistics
RPRT	Report
SC&A	S. Cohen and Associates
SEC	Special Exposure Cohort
SRS	Savannah River Site
UBGR	Upper Bound of the Gray Region
WBC	Whole Body Count
WRS	Wilcoxon Rank Sum test

EXECUTIVE SUMMARY

This report presents SC&A's initial review of ORAUT-RPRT-0053, *Analysis of Stratified Coworker Datasets* (ORAUT 2012d), which describes the National Institute for Occupational Safety and Health's (NIOSH's) proposed methods for reconstructing exposures to unmonitored workers using coworker data. RPRT-0053 addresses three major topics:

(1) Methods for analyzing coworker urinalysis datasets with varying proportions of samples below the detection limit (nondetects), with and without the assumption of lognormality

(2) Reduction of coworker datasets containing individual urinalysis samples to one statistic per worker per time period

(3) Methods for evaluating the stratification of the worker population into two distinct subgroups of workers (strata) using distributions based on the reduced data

The three topics are discussed separately in this report.

RPRT-0053 reviews several statistical methods that are available for analyzing the coworker datasets. A range of methods is included for analysis of datasets with a varying proportion of nondetects, ranging from none to essentially all or most of the available data from monitored workers. The proposed methods for analyzing datasets summarize, and in some cases improve on, methods previously reported in ORAUT-OTIB-0019, *Analysis of Coworker Bioassay Data for Internal Dose Assessment* (ORAUT 2005); ORAUT-PROC-0095, *Generating Summary Statistics for Coworker Bioassay Data* (ORAUT 2006a)*;* ORAUT-OTIB-0075, *Use of Claimant Datasets for Coworker Modeling* (ORAUT 2009a); and ORAUT-RPRT-0044, *Analysis of Bioassay Data with a Significant Fraction of Less-Than Results* (ORAUT 2009b).

The methods presented in RPRT-0053 serve as the theoretical basis for several subsequent reports related to the coworker models at the Savannah River Site (SRS), including ORAUT-RPRT-0055, *A Comparison of Exotic Trivalent Radionuclide Coworker Models at the Savannah River Site* (ORAUT 2012a); ORAUT-RPRT-0056, *A Comparison of Neptunium Coworker Models at the Savannah River Site* (ORAUT 2012b); and ORAUT-RPRT-0058,*A Comparison of Mixed Fission and Activation Product Coworker Models at the Savannah River Site* (ORAUT 2012c). Hence this review also addresses potential issues related to the application of the proposed methods to the SRS coworker models, especially those issues that might be considered Special Exposure Cohort (SEC)-related, in order to review their suitability in practice. We note that the detailed review of the SRS SEC issues is being done separately by SC&A; it is anticipated that that material will be reviewed by the SRS Work Group.

NIOSH proposes a *one person-one sample* (OPOS) statistic be applied by averaging all urinalysis samples collected from a worker in a given time period. This would reduce the urinalysis dataset to a single mean value for each worker in each time period. The set of mean values is then used to fit an exposure distribution in each period. The exposure distribution is used to provide estimates of the geometric mean (GM) and geometric standard deviation (GSD) in that period for use in the coworker model. As the mean value cannot be calculated when there are samples below the detection limit (nondetects) present in the data, NIOSH proposes to use

the *maximum possible mean* (MPM) as an upper bound on the true mean value. The MPM is the mean obtained after first replacing the nondetect samples with the values of their detection limits.

The use of average values does not account for variability of the samples within the time period and the procedure will result in lower values of the GSD used in the coworker model. The OPOS approach represents a significant departure from the previous coworker model methodologies. This change may require re-evaluation of all previous cases with determinations that were based on coworker model estimates. Two analyses were conducted to examine the changes introduced by using the OPOS approach in the development of coworker models. SC&A conducted a simulation analysis to determine the difference between the lognormal models developed under the OPOS approach versus the use of the full set of individual samples from all workers. In a second analysis by SC&A, the SRS tritium database from 1962 through 1964 was examined to compare the parameters of lognormal distributions generated under the OPOS and full coworker models to lognormal distributions estimated for individual workers in each year.

A hypothesis testing procedure is proposed for determining when there are "*significantly different strata*." The hypothesis test procedure compares two worker groups using the Monte Carlo permutation test (MCPT) based on a parametric lognormal model and the nonparametric Peto-Prentice test. The MCPT compares parameters of lognormal distributions fitted to the OPOS data using the regression on order statistics (ROS) method (Helsel 2005). The ROS method permits estimation of lognormal parameters from data containing nondetects.

Two examples are provided in RPRT-0053 to demonstrate the application of these tests. One of these examples with a borderline decision was re-examined using a simulation approach to validate the permutation test results.

The recommended hypothesis tests apply to only two groups, although several methods for multiple comparisons are discussed. Stratified models generally contain more than two strata with the objective of developing estimates for each strata. There is usually no *a priori* requirement that the strata be significantly different, although the resulting estimates may have sufficient precision to determine significant differences *if the sample sizes are sufficiently large*. SC&A has shown in prior work that more than two strata are necessary in at least some cases, so as to ensure that coworker dose estimates are claimant favorable. In one example, we found that SRS construction workers need to be subdivided by job type and area of work (SC&A 2010a, SC&A 2010b). This issue and the possibility of addressing it by multiple comparisons is addressed briefly in the main body of this report.

SC&A has identified the following findings.

Finding No. 1: Due to the dependencies that exist in the ranked data, the R^2 for ROS does not have the usual interpretation. The recommendations in RPRT-0053 for using ROS do not address this concern.

Finding No. 2: In the application of the procedures recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored coworkers would be in question. The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed. The methods proposed in RPRT-0053 for analyzing the coworker datasets require verification that (1) the available coworker data are representative of all groups of workers, and (2) the manner of use of the data is claimant favorable for the specific datasets to which the method is applied. A sound statistical methodology is subject to these two important caveats.

Finding No. 3: The OPOS statistic methodology summarizes a worker's exposure by averaging overall urine samples collected during the specified time period. The use of average values does not account for variability of the samples within the time period, and the procedure will result in lower values of the GSD used in the coworker model.

Finding No. 4: The OPOS method must strictly be applied to comparisons where the sampling protocol was the same. Specifically, when there is evidence that the sampling protocol for one group of workers was different than the protocol used for the other group, the tests do not provide a valid comparison. For example, if the monitoring of one group of workers is incident-driven and the other is not, then the OPOS approach is not appropriate for comparing the two distributions.

Finding No. 5: The methods in RPRT-0053 require a high level of confidence before deciding that the two worker groups are significantly different. The requirement for a high level of confidence in this decision is not claimant favorable when using a null hypothesis of "No Difference." The power of the tests to detect differences given the limited quantity of available data has not been established. The Data Quality Objectives (DQO) process should be used to balance Type 1 and Type 2 decision errors.

Finding No. 6: Given the small number of trades worker (CTW) data points, in many years the tests cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/non-construction trades worker (non-CTW) ratio of GMs. Larger differences have a 95% or better chance of detection. Smaller differences would be in the "gray region" for the test, sometimes detected, sometimes not. Overall, SC&A concludes that the NIOSH method of concluding that there are no significant differences based on the available data would often lead to very claimant-unfavorable results.

Finding No. 7: The statistical tests for comparing two strata require that the samples in each group be independent. If a worker in one group is exposed to radionuclides with long retention in the body and then changes jobs and becomes part of the other group in the same year, the OPOS values are correlated for this worker. This correlation not only violates the assumptions of the tests, but also creates a bias toward a decision of "No Difference" between the two groups.

Finding No. 8: Although one example where a significant difference is found is presented in the report, NIOSH has not provided any measure of the power of the hypothesis test procedure to detect differences within the worker population. This deficiency should be corrected before the test is adopted as an appropriate procedure for coworker models. Conducting the tests at a 90% level of confidence would be claimant favorable.

RECOMMENDATIONS:

(1) NIOSH might consider reversing the null hypothesis for the Peto-Prentice test. NIOSH's implementation of the hypothesis tests to test for differences between CTWs and non-CTWs at SRS uses a null hypothesis that is not claimant favorable, as it places the burden of proof on the CTW claimants to prove a significant difference. The Peto-Prentice test is more generally applicable and may be applied using a claimant-favorable null hypothesis. The groups of workers with suspected high exposures should be considered different in the absence of strong evidence that they are not. This is more likely to result in a claimant-favorable model.

(2) More than two strata would be required to properly characterize the varied worker populations at many sites. Multiple comparisons when there are more than two strata may be possible, but could be complex and suffer from limits imposed by small sample sizes. The analysis may spiral into large numbers of comparisons with inconclusive results.

(3) There is persuasive evidence provided by the analysis of SRS CTWs by job type and by area of work (SC&A 2010a, 2010b) that subgroups of CTWs are not drawn from the same distribution as non-CTWs. When the distributions of CTW subgroups are different from non-CTW, CTW data by job type and area can be used to construct coworker models for their CTW peers. Of course, this requires sufficient data in each job/area category for which a coworker model is to be constructed.

(4) NIOSH has stated that each of the two groups must have a sufficient number of samples; at least 30 samples in each group would be required. If the total number of samples is less than 30, the method is not suitable. Boiling down the number of samples by using OPOS reduces the number of samples and produces greater uncertainty and a larger gray region, making the test less claimant favorable. As an example, the number of samples used in the coworker study for neptunium at SRS was so small in approximately one-half of the years analyzed in RPRT-0056 that the difference between groups would have to exceed a factor of 3 to 4 for a significant difference with these sample sizes. In all remaining years except for 1985, the difference between groups would have to be as large as a factor of 4 to a factor of over 10 before the tests would indicate a significant difference. Due to the low power of these tests, NIOSH findings of no significant differences are due mainly to an inadequate sample size and the use of OPOS values that has further reduced the sample sizes.

(5) In principle, multiple comparisons can be done for more refined groupings, like CTWs by job type with all non-CTWs. But this will run into difficulties in many cases, as we found in prior analyses even for a 10-sample threshold. It will be much more difficult to meet the 30-sample threshold needed for the tests recommended in RPRT-0053, but this is essential for a valid comparison. Moreover, a valid comparison requires that the

30-sample threshold be met for *each o*f the two groups, not just one. RPRT-0053 is not explicit on this point, though it is implied in footnote 6 on page 9. The 30-sample threshold for each group should be made explicit.

1.0 INTRODUCTION

This report presents SC&A's initial review of ORAUT-RPRT-0053, *Analysis of Stratified Coworker Datasets* (ORAUT 2012d), which describes the National Institute for Occupational Safety and Health's (NIOSH's) proposed methods for reconstructing exposures to unmonitored workers using coworker data. RPRT-0053 summarizes and extends the statistical methods for analysis of coworker data that were previously published. The proposed analyses follow the general framework shown in Table 1. The shaded areas in Stage 2 and Stage 3 indicate the main subject areas of the report.

Table 1. Proposed Stages of the Dose Reconstruction Process

Stage 1: Urinalysis Data
Stage 2: OPOS Urinalysis Data
Stage 3: 50th and 84th Percentile Urinalysis Estimates from OPOS Data
Stage 4: 50th and 84th Percentile Intake Rates
Stage 5: Person-specific Intakes and Doses
Stage 6: Probability of Causation

The report addresses three major topics:

(1) Methods for analyzing coworker urinalysis data in Stages 2 and 3 with varying proportions of nondetects, with and without the assumption of lognormality. RPRT-0053 reviews a toolbox of statistical methods appropriate for analyzing the available data in each stratum.

(2) Reduction of the Stage 1 coworker urinalysis dataset containing individual urinalysis samples to a single statistic per worker per time period (Stage 2). NIOSH proposes that a *one person-one sample* (OPOS) statistical approach be applied by averaging all urinalysis samples collected from a worker in a given time period. This would reduce the urinalysis dataset in Stage 1 to a single mean value for each worker in each time period (Stage 2).

(3) Methods for evaluating the stratification of the worker population into subgroups of workers (strata) using distributions based on the reduced data. Several hypothesis testing procedures are proposed for determining when there are "*significantly different strata*." The Peto-Prentice form of the Wilcoxon Rank Sum (WRS) test is recommended for comparing the strata in each time period based on the OPOS statistics (Stage 2). The Monte Carlo permutation test (MCPT) described in ORAUT-RPRT-0049 (ORAUT 2010a) is also recommended for comparing the strata in each time period based on the parameters obtained by fitting lognormal distributions to the OPOS statistics (Stage 3).

The three topics are addressed separately in the following sections of the report.

1.1 RELATED ORAUT PUBLICATIONS

1.1.1 Source Documents

RPRT-0053 includes a toolbox of statistical methods that are available for analyzing the data in each strata. A range of methods is included for analysis of datasets, with the proportion of nondetects ranging from none to essentially all or most of the available data from monitored workers. The range of proposed methods for analyzing datasets summarizes methods previously reported in:

- ORAUT-OTIB-0019, *Analysis of Coworker Bioassay Data for Internal Dose Assessment* (ORAUT 2005)
- ORAUT-PROC-0095, *Generating Summary Statistics for Coworker Bioassay Data* (ORAUT 2006a)
- ORAUT-OTIB-0075, *Use of Claimant Datasets for Coworker Modeling* (ORAUT 2009a)
- ORAUT-RPRT-0044, *Analysis of Bioassay Data with a Significant Fraction of Less-Than Results* (ORAUT 2009b)
- ORAUT-RPRT-0049, *Discussion of Tritium Coworker Models at the Savannah River Site – Part 1* (ORAUT 2010a)

1.1.2 Dependencies

Several recent Oak Ridge Associated Universities Team (ORAUT) documents are based on the analytical methods proposed in RPRT-0053. These documents include:

- ORAUT-RPRT-0055, *A Comparison of Exotic Trivalent Radionuclide Coworker Models at the Savannah River Site*, (ORAUT 2012a)
- ORAUT-RPRT-0056, *A Comparison of Neptunium Coworker Models at the Savannah River Site* (ORAUT 2012b)
- ORAUT-RPRT-0058, *A Comparison of Mixed Fission and Activation Product Coworker Models at the Savannah River Site (*ORAUT 2012c)

As the titles indicate, these applications address comparisons of coworker models for the CTWs and non-CTWs at the Savannah River Site (SRS). The reports use the Monte Carlo permutation test methodology recommended in ORAUT-RPRT-0053 and/or the Peto-Prentice test for these comparisons. In RPRT-0055, trivalent radionuclides bioassay data are first reduced using the OPOS method. Similar applications of the RPRT-0053 methodology at SRS are made in ORAUT-RPRT-0056 and ORAUT-RPRT-0058. All of these applications discussed the comparison of coworker models at SRS. Our review includes several examples of the RPRT-0053 methodology drawn from these derivative publications.

2.0 STATISTICAL METHODS FOR ANALYZING COWORKER URINALYSIS DATASETS

Several statistical procedures are recommended in RPRT-0053 to estimate parameters for use in the coworker model. The coworker model is used in the dose reconstruction process to estimate dose to an individual in periods with unmonitored or undocumented exposure. RPRT-0053 provides a toolbox of statistical methods for analyzing the available coworker data and estimating the 50th and 84th percentiles of exposure. Three approaches for fitting probability distributions to bioassay data are discussed. The three approaches cover a range of methods appropriate for analysis of datasets with a number of nondetects ranging from none or few to essentially all or most of the available data from monitored workers. RPRT-0053 summarizes and extends the statistical methods for analysis of coworker data previously published in ORAUT-OTIB-0019, *Analysis of Coworker Bioassay Data for Internal Dose Assessment* (ORAUT 2005); ORAUT-PROC-0095, *Generating Summary Statistics for Coworker Bioassay Data* (ORAUT 2006a); ORAUT-OTIB-0075, *Use of Claimant Datasets for Coworker Modeling* (ORAUT 2009a); ORAUT-RPRT-0044, *Analysis of Bioassay Data with a Significant Fraction of Less-Than Results* (ORAUT 2009b); and ORAUT-RPRT-0049, *Discussion of Tritium Coworker Models at the Savannah River Site – Part 1* (ORAUT 2010a).

RPRT-0053 recommends the OPOS statistic approach to derive a single value that represents the exposure to each worker in each time period. The OPOS methodology summarizes a worker's exposure by averaging the concentration in all urine samples collected during the specified time period. The following guidelines are provided in the text and footnotes on page 9 of RPRT-0053 for the application of the OPOS methodology:

> *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.*
>
> [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 (*[EPA 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.*

Time periods may be as long as 3 years, but more often as short as 1 year. Yearly OPOS estimates are calculated for workers at SRS for exotic trivalent radionuclides in RPRT-0055, for neptunium in RPRT-0056, and for mixed fission and activation products in RPRT-0058, with few exceptions. Problems associated with the use of the OPOS statistic to represent annual worker exposure are discussed in more detail in Section 3. The questions of minimum sample size and EPA's advice on the sample size issue noted in footnote 6 in the passage above are addressed further in Section 4.

The regression on order statistics (ROS) method for estimating the geometric mean (GM) and the geometric standard deviation (GSD) of a lognormal distribution was described in ORAUT-PROC-0095, *Generating Summary Statistics for Coworker Bioassay Data* (ORAUT 2006a). The deficiencies of ORAUT-PROC-0095 when there is a large proportion of nondetects were noted by SC&A in *Findings from 3rd set of Procedures* (SC&A 2007). In *NIOSH Responses to Selected Findings from 3rd set of Procedures* (ORAUT 2010b), NIOSH explained that a new report (i.e., ORAUT-RPRT-0044) has been developed to better address the issue of censored data. RPRT-0044 describes in more detail the methods recommended in RPRT-0053 applicable to datasets with a large proportion of nondetects. SC&A identified the following relevant findings during the review of RPRT-0044 (SC&A 2010c).

> ***Finding No. 1:*** *The statistical methods proposed in ORAUT-RPRT-0044 are based on sound statistical methodologies, and the material is well presented. The proposed methods are an improvement over the regression methods proposed in ORAUT-PROC-0095 when essentially all or most of the data are less-than results, the limit of detection was the same for all samples in the dataset, and the samples above the limit of detection are randomly spread across workers, job types, and work areas.*
>
> *The work location and work assignments of the workers with positive results are not considered in the NIOSH approach. Before these methods are used in a coworker model, further analysis of the positive results is required. In particular, identification of the workers, work areas, and processes accounting for the positive results in the datasets is required to reveal possible patterns that may explain the occurrence of positive results.*
>
> *NIOSH does not offer any consideration relating to the pattern or time distribution of the positive results. For example, the positive results could be present x times per year, during defined periods of time, or during a specific campaign. It is possible that the same subgroup of workers accounted for most of the positive readings year after year.*
>
> ***Finding No. 2:*** *ORAUT-RPRT-0044 does not address the representativeness of the dataset for workers in all work areas and job types. No individual worker analysis was performed, as the report concentrates only on analysis of a collection of analytical results. (…)*
>
> ***Finding No. 4:*** *The methods proposed in RPRT-0044 for datasets with essentially all or most of the less-than results are based on samples obtained from all workers, regardless of job type or location. No attempt was made to determine the work areas, processes, or job types of workers with positive results. This approach is not claimant favorable for construction workers for three reasons:*
>
>> *(1) In many cases, construction workers were not monitored as frequently as non-construction workers, hence the dataset may not*

> > *be representative of the distribution of construction worker exposures.*
> >
> > *(2) Because constructions workers were sampled less frequently, a higher percentage of these workers will require use of the coworker model.*
> >
> > *(3) The positive samples may come from very few workers or restricted time periods, which may not be representative of the worker population.*
>
> *The work assignments of the workers with samples in the upper tail of the mixture distribution may have an unexpectedly high number of construction workers when compared with their degree of monitoring. The work assignments of workers with samples in the two populations should be inspected and categorized by job type to look for such disparities.*

A part of the present report reviews whether NIOSH has addressed these concerns in RPRT-0053.

2.1 REGRESSION ON ORDER STATISTICS (ROS)

The ROS method has been used routinely by NIOSH as a convenient way to estimate parameters of a lognormal distribution when there are nondetects in the dataset. The ROS method is incorporated in RPRT-0053 by reference to prior publications. On page 9 of RPRT-0053, NIOSH explains that the recommended ROS procedure is the same as that recommended in PROC-0095, with the following exception:

> *ORAUT-PROC-0095 uses Hazen plotting points (...), 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 the data in the dataset). Here, Helsel-Cohn plotting points are used (...), which are suitable for single and multiple left-censoring... With the exception of the plotting points, the ROS used here is the same as the ROS used in ORAUT-PROC-0095.*

The ROS method is based on least squares regression. The regression model for a dataset of size n is written as:

$$\ln(y_i) = mx_i + b$$

where the values y_i, $i = 1, \ldots, r$ denote the $r \le n$ values above the limit of detection (i.e., not including the “less than” values). The symbols b and m denote the y-intercept and slope of the regression line, respectively, and *ln* denotes the natural logarithm.

Each observation is assigned a corresponding normal score depending on its plotting position: $x_i = \Phi^{-1}\left[(R_i - 0.5)/n\right]$ for $i = 1, \ldots, r$. Here, $\Phi^{-1}[\cdot]$ denotes the inverse of the cumulative normal distribution function. The argument used for this function is a function of the rank R_i, which is the rank of y_i when all n data values are included in the ranking and the (n-r) values below the minimum detectable amount (MDA) are assigned the lowest (n-r) ranks. In this example, Hazen plotting position is the percentile midpoint; ROS with other plotting positions has the same form of model with slightly different values for the x_i. The regression estimates are based on the entire set of data above the detection limit. If the data fit a lognormal distribution where the logarithms have mean μ and standard deviation σ, then a scatter plot of the r points $[x_i , \ln(y_i)]$ will lie on a straight line with slope $m = \sigma$ and a y-intercept $b = \mu$. The regression method may be used not only to verify that the data follow a lognormal distribution, but also to provide estimates for the parameters of the lognormal distribution when there are values in the dataset below the detection limit.

A determination of the goodness-of-fit of the lognormal distribution is based on regression R^2 although, due to the dependencies that exist in the regression estimates derived from ranked data, the R^2 does not have the usual interpretation. RPRT-0053 states on page 8:

> *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 (…) 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 (…).*

Although NIOSH has an apparent concern that the assumptions of linear regression apply, the data values in the ROS scatter plot are not independent observations. If $x_i \leq x_j$, then it is known with certainty that $y_i \leq y_j$. This dependence among the observations violates the usual assumption of conditional independence of the y values in the regression, given the set of x values. In addition, the y_j are autocorrelated and heteroscedastic, as noted long ago in Looney and Gulledge (1985):

> (…) *we propose the use of the correlation coefficient in constructing a goodness-of-fit test statistic from a plot based on any particular plotting position. Since* ***the*** ***$Y_{(i)}$ are highly correlated and heteroscedastic, the usual distributional results for the correlation coefficient do not apply****. Instead, empirical sampling methods must be used to determine the null distribution of the test statistic.*

The heteroscedasticity of the order statistics derived from a set of samples is well known (David and Nagaraja 2003). Clearly, the sample minimum and maximum have the greatest sampling variance. Moreover, when the order statistics are used in ROS, these extreme values on the far left and far right of the distribution have the greatest influence on the lognormal parameter estimates. The GSD derived from the ROS slope estimate is particularly sensitive to the influence of heteroscedasticity in the order statistics. The recommendations in RPRT-0053 for using ROS do not address these serious deviations from the standard linear regression model assumptions when interpreting the R^2 values as a measure of goodness of fit of the ROS probability plot.

Finding No. 1: Due to the dependencies that exist in the ranked data, the R^2 for ROS does not have the usual interpretation. The recommendations in RPRT-0053 for using ROS do not address this concern.

When there is a large proportion of nondetects and/or the ROS method generates a GSD estimate over 6, NIOSH suggests the use of the maximum likelihood methods discussed in ORAUT-RPRT-0044, *Analysis of Bioassay Data with a Significant Fraction of Less-Than Results* (ORAUT 2009b). A second method recommended in RPRT-0053 for datasets with a large proportion of nondetects (>85%) is the binomial fit method, also discussed in RPRT-0044. The binomial fit method is not recommended by NIOSH in RPRT-0053 for use in stratifying datasets, as detailed in Section 2.3 of this report.

2.2 EFFECTIVE FIT

The maximum likelihood method, which is also called "*effective fit,*" attempts to fit a mixture of two distributions. The effective fit method assumes the distribution of results is not a single lognormal distribution, but a mixture of two distinct distributions, one normal and the other lognormal. In this model, there are two populations of samples; most samples have no measureable level of analyte in the urine, but a small fraction of the samples do. In this mixed model, the former group of samples with less-than results is assigned a normal distribution representing "background" exposures, and the latter group of samples is assigned a lognormal distribution of exposures.

Maximum likelihood techniques are used to estimate the parameters of the mixed model. If a dataset contains urine results for which most of the workers do not have analyte in their urine but a small fraction of the workers do, then the methods presented in RPRT-0053 are an improvement over the PROC-0095. However, NIOSH does not offer any consideration relating to the pattern or time distribution of the positive results. It is necessary to know if the positive results occur every year, and if those results are related to a particular procedure. For example, the positive results could be present *x* times per year, during defined periods of time, or during a specific campaign.

In the application of the procedure recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. Some workers were not monitored; otherwise there would be no need for a coworker model. The underlying assumption appears to be that the workers with the most exposure potential were monitored, but we have seen in a number of cases that this was not necessarily true. If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored workers would be in question.

Finding No. 2: In the application of the procedures recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. If the unmonitored workers are not from a population that had the highest exposure potential, the applicability of a coworker model derived from monitored coworkers would be in question. The matter of the relative exposure potential of the monitored workers needs to be demonstrated rather than assumed. The methods proposed in RPRT-0053 for analyzing the coworker datasets require verification that (i) the available coworker data are representative of all groups of workers, and (ii) the manner of use of the data is claimant favorable for the specific datasets to which the method is applied. A sound statistical methodology is subject to these two important caveats.

2.3 BINOMIAL FIT

The binomial fit method is recommended in RPRT-0053 if the coworker dataset has a very large proportion of nondetects, up to and including 100%. This method for estimating the GM and GSD was presented previously in ORAUT-RPRT-0044 (ORAUT 2009b). In this situation, there are insufficient data to permit a comparison of distributions. This method will not be used in comparing worker groups, as NIOSH recommends the following:

> *Datasets that are modeled using the binomial fit are considered to not contain enough information to decide if strata are different, and it is recommended that such datasets not be stratified.*

3.0 ONE STATISTIC PER WORKER PER TIME PERIOD

Prior to RPRT-0053, coworker models were constructed from the full set of samples for the workers of interest, as indicated by NIOSH in the Introduction:

> *Coworker models are typically constructed using data from all monitored workers by fitting a lognormal probability distribution to the data (…) 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 when a complete dataset was not available (…)*

RPRT-0053 recommends a new approach to estimating coworker models. NIOSH proposes that an OPOS approach be applied by averaging all urinalysis samples collected from a worker in a given time period. NIOSH also proposes that the sample statistic selected for the OPOS statistic is the mean of the individual sample values (average value). As the mean value cannot be calculated when there are samples below the detection limit (nondetects) present in the data, NIOSH proposes to use the *maximum possible mean* (MPM) as an upper bound on the true mean value. The MPM is the mean obtained after first replacing the nondetect samples with the values of their detection limits.

3.1 OUTLINE OF OPOS PROCEDURE

The basic steps of the OPOS approach are as follows.

- For each individual and for each bioassay sample determine the group classification (Group A, Group B, or unknown).
- Determine the OPOS statistic (usually the MPM) for each individual and time period.
- Determine the GM and GSD urinary excretion rates using the ROS, effective fit, or binomial methods as appropriate for each strata for each time period.
- Compare the strata using the MCPT at Stage 3 and/or Peto-Prentice test at Stage 2 to determine if there is a statistically significant difference between the two strata.

This procedure would reduce the original urinalysis sample dataset to a single mean value for each worker in each time period. The set of mean values is then used to fit an exposure distribution for the coworker model in each period. The exposure distribution is used to provide estimates of the GM and GSD in that period for the coworker model. The following discussion of the OPOS approach is recorded in the notes from the Subcommittee on Procedures Review meeting on November 1, 2012 (Meeting Transcript 2012, pp. 118 to 120):

> *MEMBER ZIEMER: It would change the distribution.*
>
> *MR. HINNEFELD: Nominally probably.*
>
> *CHAIR MUNN: Yes.*

> *MR. HINNEFELD: Yes, it would change the distribution of doses because you drop out some of the high-end stuff. So no one has yet jumped in to correct me, so maybe I got it right.*

In this section of the report, we examine how the distribution will change if the recommended OPOS procedure is adopted. The OPOS exposure distributions are compared with the exposure distributions derived from the full set of samples. The OPOS methodology summarizes a worker's exposure by averaging overall urine samples collected during the specified time period. In the case of nondetect samples, the detection limit is used for the nondetects in the calculation of the average to yield the MPM. The use of average values does not account for variability of the samples within the time period, and the procedure will result in lower values of the GSD used in the coworker model compared with previous procedures. A GSD must be assigned for the missing dose to a worker in each year, and that GSD should reflect the variability in that worker's exposure during the year. The OPOS GSD measures the variability of average annual dose across workers, and ignores variability for an individual worker within the year.

Finding No. 3: The OPOS statistic methodology summarizes a worker's exposure by averaging overall urine samples collected during the specified time period. The use of average values does not account for variability of the samples within the time period and the procedure will result in lower values of the GSD used in the coworker model.

The OPOS methodology does not examine the temporal pattern of individual exposures for longer than one time period. NIOSH's strategy for bounding internal dose based on bioassay has been to assume a claimant-favorable chronic intake throughout the year. That is, the bounding assumptions presumably account for the variability in intake regimens. It remains to be determined whether those assumptions sufficiently bound intakes based on the OPOS methodology. It will be necessary to develop a list of radionuclides for which the procedure reflects the temporal patterns of exposure and subsequent organ doses.

When comparing two populations using a statistical test for differences, it is important that the data are collected following the same protocol for both groups of workers. In the specific case of CTW versus non-CTW comparisons in RPRT-0056, NIOSH has said that sampling was incident-related for CTWs and routine for non-CTWs, so the OPOS method does not appear appropriate for comparing the two distributions.

The valid use of OPOS values in two-group comparisons requires that the sampling protocols used for each group be similar. There are many ways that the sampling protocols may differ, and NIOSH has not addressed the extent of this problem. For example:

(1) Do some workers have OPOS results for a short period of time and not for the whole year?

(2) Do some workers have OPOS results derived from a large number of samples during the period of time and other workers from the same population have one or only a few samples? Should a lognormal distribution be formed by combining OPOS and individual samples into a single distribution?

(3) Does the lognormal distribution from one population consist of OPOS results for all workers and the other lognormal distribution from the second population have a mixture of OPOS results and individual samples, or consist almost entirely of individual samples? Should the two lognormal distributions be compared?

The answers to these questions are important because the use of OPOS values introduces complications in the subsequent coworker model analyses that rely on these values. OPOS values are not measurements, but are statistics derived from a set of measurements. The OPOS values are averages of a varying number of samples, with a different number for each worker. Since it is an average, each OPOS value has an uncertainty associated with the calculated value. Mathematically, the sampling variance of an average of *n* independent samples $\overline{X} = \sum_{i=1}^{n} X_i / n$ is $Var(\overline{X}) = Var(X)/n$. The sampling variance of the average increases with the variance in the individual worker's samples during the time period covered by the OPOS value, and the sampling variance of the OPOS average is smaller for workers with a larger number of samples during the period. Use of the MPM in place of a simple average for the OPOS value when there are nondetects introduces an additional level of uncertainty in the OPOS values that is not addressed in the equation above.

A difference in the number of samples available for the workers in each group implies a difference in the uncertainty for the OPOS values for each group. In general, more samples are available for the onsite workers who are part of an ongoing monitoring program. Due to the larger number of samples, the OPOS values for the onsite workers may be measured with greater precision than is available for other groups of workers.

Since there is uncertainty in the OPOS statistics, and this uncertainty varies from worker to worker and from one group of workers to another, all subsequent analyses based on OPOS values are conducted using heteroscedastic data. Finding 1 in Section 2.1 indicates that the ROS method conducted on individual samples ignores the heteroscedastic nature of the order statistics derived from the sample values. If the order statistics are derived from OPOS values, this introduces a second problem unique to the use of OPOS values in ROS: values that are being ranked may not come from the same distribution unless the monitoring protocol is the same for all members of the group.

The uncertainty in the OPOS averages affects the hypothesis tests used for comparing two groups of workers. Since the MCPT is conducted using ROS on OPOS values to estimate the lognormal parameters for the test, this test does not address the issue of heteroscedasticity in the OPOS values. The uncertainty in the OPOS averages also affects the nonparametric two-sample tests used for comparing two groups of workers. The assumptions underlying the tests are violated if the nonparametric tests are applied using data with different variances in each group. The WRS test and the generalized WRS tests, including the Peto-Prentice test, are based on an assumption that the only difference between the two groups is a difference in the location of the distributions (Conover 1980, p. 217). This means that the shapes and variances of the two distributions should be approximately the same. If the OPOS values for one group of workers are derived from a relatively small number of samples per worker while the OPOS values from the other group are derived from a larger number of samples per worker, the assumption of equal variance for both groups is suspect. Attachment B to RPRT-0053 reviews the discussion of the generalized WRS tests found in Helsel (2005). The literature search does not include reference to the Brunner-Munzel test (Brunner and Munzel 2000). This nonparametric test is another generalization of the WRS test designed for comparisons of populations with different variances.

Given the problems introduced by the use of OPOS when there are different sampling protocols for each group, SC&A recommends that:

(1) OPOS values should not be combined into a single lognormal distribution when the sampling protocols for subsets of workers in the group differ

(2) Distributions of OPOS values can be compared only when the sampling protocols are the same for both groups.

Finding No. 4: The OPOS method must strictly be applied to comparisons where the sampling protocol was the same. Specifically, when there is evidence that the sampling protocol for one group of workers was different than the protocol used for the other group, the tests do not provide a valid comparison. For example, if the monitoring of one group of workers is incident-driven and the other is not, then the OPOS approach is not appropriate for comparing the two distributions.

The OPOS approach proposed in RPRT-0053 represents a significant departure from the previous coworker model methodologies. This report contains results of two analyses of the effects of the OPOS approach in the development of a coworker model. First, a simulation is conducted to determine the difference between the lognormal models developed under the OPOS approach versus the use of the full set of individual samples from all workers. In the second analysis, the SRS tritium database from 1962 through 1964 is examined to compare the parameters of lognormal distributions generated under the OPOS and full models to lognormal distributions estimated for SRS individual workers in each year.

3.2 SIMULATION TO COMPARE OPOS AND FULL COWORKER MODELS

The simulation analysis indicates that the OPOS approach results in underestimation of the range of variability across workers reflected in estimates of the GSD and 95th percentile, which are biased low relative to the original samples. As shown in Table 2, the magnitude of this bias increases as the number of samples per worker used in the MPM increases. A positive bias in the GM estimate is also noted. This complementary bias acts to keep expected values at approximately the same level. In Figure 1, the combined effect of the downward bias in the GSD and upward bias in the GM generates a downward bias in the estimates of the 95th percentile. For a 5-sample MPM, the bias may be as large as minus 30%. When a larger number of samples is averaged, the bias would be larger in magnitude.

Table 2. Bias of OPOS Parameter Estimates versus Individual Samples

	Bias (%)		
Number of Samples in MPM	**GSD (Slope)**	**GM (Intercept)**	**95th Percentile**
1	0	0	0
2	-23	32	-13
3	-30	38	-24
4	-33	46	-24
5	-35	45	-29

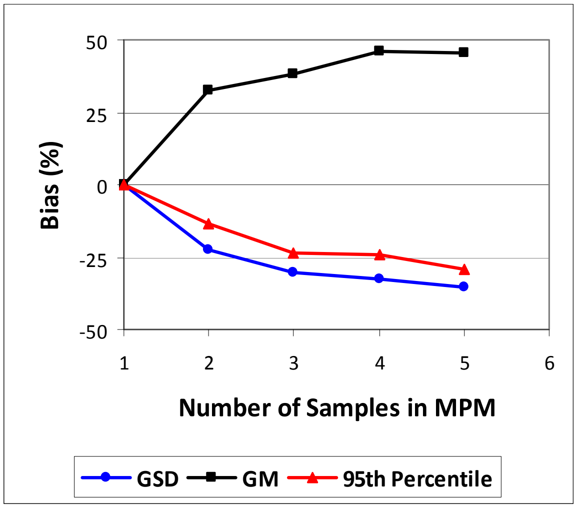

Figure 1. Bias of Simulated OPOS Estimates for Geometric Mean, Geometric Standard Deviation, and 95th Percentile vs. Individual Samples

The large differences between the OPOS GSD and 95th percentile and the full data model GSD and 95th percentile observed in this initial simulation provided the impetus for a second analysis of the OPOS method using actual data. In this analysis, the tritium bioassay data at SRS are used to compare OPOS and full models with individual worker distributions.

3.3 OPOS AND FULL MODEL COWORKER MODELS COMPARISON WITH INDIVIDUAL WORKER EXPOSURES

To evaluate the impacts of changing to an OPOS methodology, tritium bioassay data from workers at SRS from 1962 to 1964 are used to construct annual lognormal exposure distributions for the OPOS and Full models. The tritium dataset used for this analysis is robust in the sense that it contains many bioassays per worker in each year, enabling the construction of individual lognormal exposure distributions for each worker. The dataset contains a reasonably large number of detects per worker, and no very extreme values. However, the data may include workers with many closely spaced, high sample values collected after incidents.

In this analysis, the OPOS and Full model coworker distributions are compared with the exposure distributions estimated for individual workers. The unscreened 1962–1964 SRS tritium dataset contains approximately 50 workers, many with a relatively large number of samples per worker in each year. Many samples are nondetects, so the ROS method was used to estimate a lognormal distribution for each worker. An example of a worker distribution for a worker with 46 bioassays in 1962 is shown in Figure 2. The ROS lognormal estimates were developed in each year for all workers with more than 3 detects in their samples.

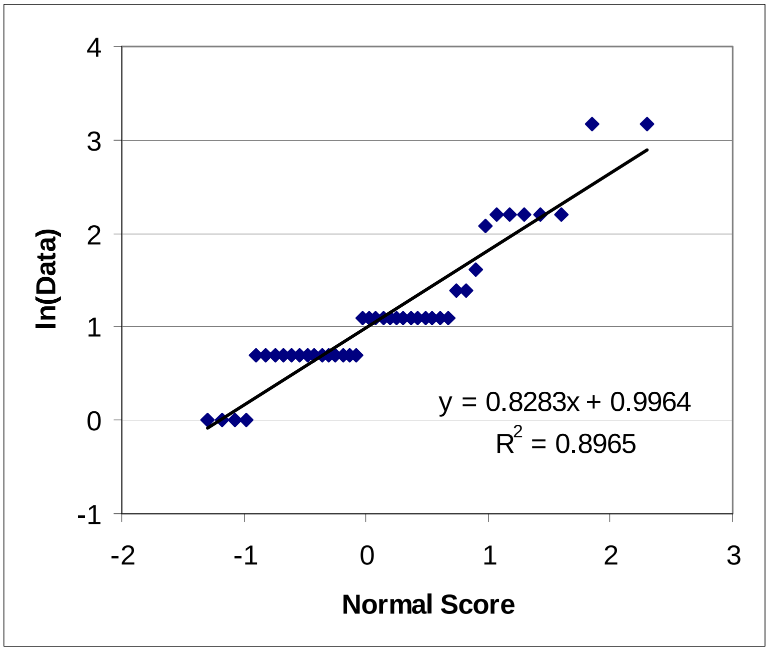

Figure 2. ROS Plot for Worker 2, 1962

Table 3 shows a comparison of the lognormal distributions generated by the two models for the 1962 through 1964 time period. The table shows the GM and GSD estimated using each model, and the mean, standard deviation, and 95th percentile of each model. The highlighted region of each table shows the ratio of the GSD, standard deviation, and 95th percentile for each model. The ratio is obtained by dividing the OPOS model estimate by the Full model estimate.

The GM and GSD of the worker lognormal distributions are plotted in Figures 3, 4, and 5. The plots also show points for the lognormal distributions for the OPOS and Full models. As expected based on the previous analysis, the OPOS model has a GSD that is substantially lower than the GSD of the Full model, and a higher GM. The OPOS estimates shown in Table 4 for the standard deviation and the 95th percentile are also substantially lower than the Full model estimates. The OPOS estimates for these three measures of the spread of the coworker distribution are from 20% to 44% lower than the estimate generated by the Full model.

Table 3. Comparison of OPOS and Full Coworker Models for SRS Tritium, 1962 to 1964

	1962			1963			1964		
	OPOS	**Full**	**Ratio**	**OPOS**	**Full**	**Ratio**	**OPOS**	**Full**	**Ratio**
GM	1.59	0.95	1.67	1.71	0.84	2.04	2.56	2.15	1.19
GSD	2.03	3.31	0.61	1.99	3.52	0.56	1.80	2.62	0.69
Mean	2.03	1.95	1.04	2.16	1.85	1.17	3.04	3.43	0.89
Standard Deviation	1.58	2.42	0.66	1.66	2.47	0.67	2.19	3.31	0.66
95th Percentile	5.06	6.81	0.74	5.28	6.64	0.80	6.73	10.53	0.64

The Full model tends to have a GSD near (or below) the upper 10% of workers with the highest GSDs. The OPOS model tends to have a GSD near (or below) the median worker. In all three figures, the OPOS GSD is exceeded by the GSD of more than one-half of the workers. It is likely that the upper 10% in Figures 3 through 5 includes workers with many closely spaced, high sample values collected after "incidents." If that were true, then the Full model would be bounding for the other 90% of workers, while the OPOS model regresses toward the median worker.

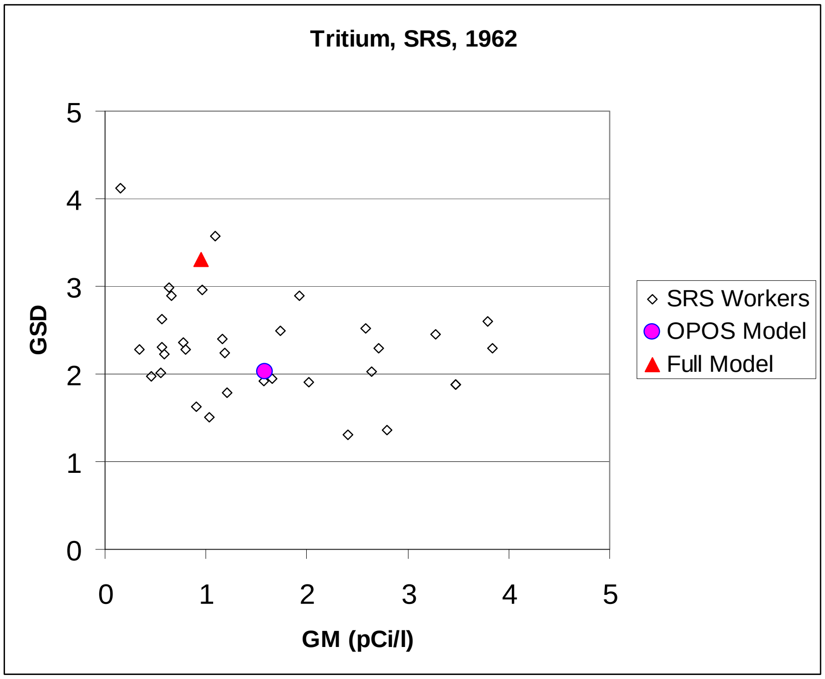

Figure 3. Scatter Plot of GSD vs. GM by Worker with OPOS and Full Models, 1962

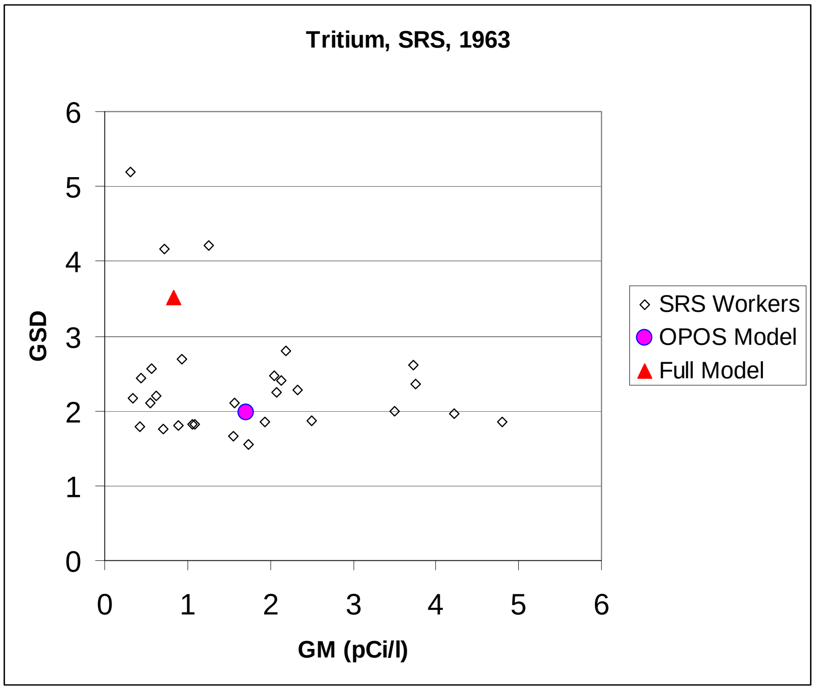

Figure 4. **Scatter Plot of GSD vs. GM by Worker with OPOS and Full Models, 1963**

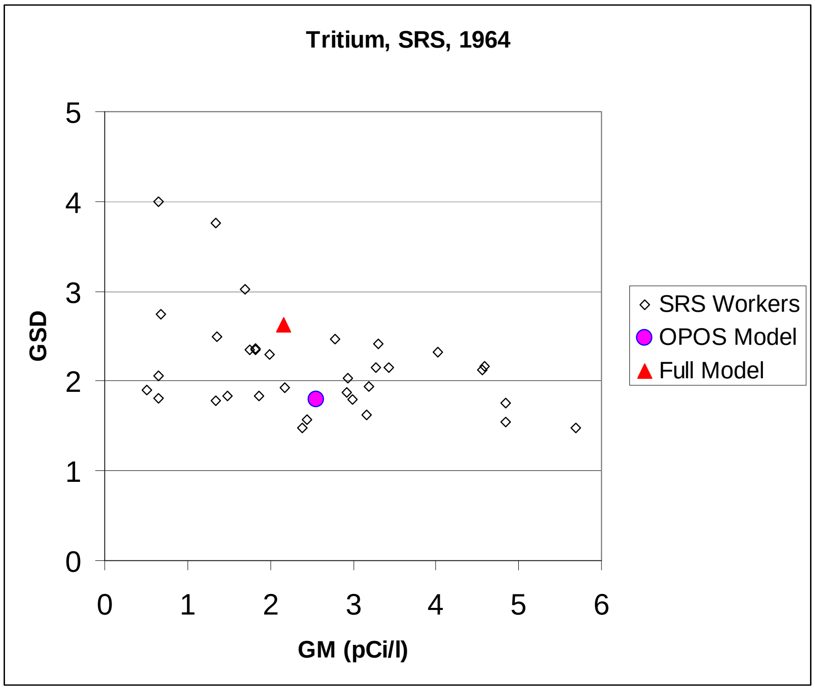

Figure 5. **Scatter Plot of GSD vs. GM by Worker with OPOS and Full Models, 1964**

4.0 STRATIFICATION OF THE COWORKER POPULATION

A hypothesis testing procedure is proposed for determining when there are "*significantly different strata*." The hypothesis test procedure compares the two strata using an MCPT and the nonparametric Peto-Prentice test. In the analysis of previously collected data, it is necessary to determine if the sample size was sufficient. NIOSH has made no effort to determine sample sizes that allow for sufficient power to detect differences.

More than two strata would be required to characterize properly the varied worker populations at many sites, including SRS. Multiple comparisons when there are more than two strata may be possible, but could be complex and suffer from limits imposed by small sample sizes. The analysis may spiral into large numbers of comparisons with inconclusive results.

4.1 HYPOTHESIS TESTING

In RPRT-0053, NIOSH recommends the MCPT for comparing two groups of workers. The permutation test (Noreen 1989) is designed to test for a statistically significant difference between lognormal distributions for the two groups. An incorrect variation of the MCPT was described in ORAUT-RPRT-0049, *Discussion of Tritium Coworker Models at the Savannah River Site – Part 1* (ORAUT 2010a). In that report, NIOSH compared the distribution of one group of workers to the entire population of workers to test for a significant difference, violating the independence of the two samples. Historically, ORAUT-RPRT-0049 was initiated by ORAUT in response to SC&A comments on ORAUT-OTIB-0075 (SC&A 2010a, SCA 2010b), which concerned the use of stratified coworker models for CTWs at SRS. The SC&A comments pertained not only to the differences in tritium exposure of the CTWs as a group versus non-CTWs, but also to the varying tritium exposures within the various construction trades. SC&A 2010a (Section 4) also looked at other radionuclides at SRS and came to a similar conclusion—taking work area and job type into account is critical for the development of a claimant-favorable coworker model.

NIOSH proposes that strong evidence ($\alpha = 0.05$ or a 95% level of confidence) is necessary before any differences between groups of workers should be considered in the coworker model. In hypothesis testing, the demand for a high degree of confidence in a decision (α or Type 1 error) is usually balanced by a requirement for adequate power (β) to ensure the test has a capability of detecting differences thought to be of importance. Although there is a general discussion of power in the literature review included in Attachment B of RPRT-0053, NIOSH has not provided any measure of the power of the MCPT to detect differences given the sample sizes and variability encountered in the available datasets. One example where a significant difference was found is presented in the report. This deficiency should be corrected before the MCPT is adopted as an appropriate testing procedure.

In RPRT-0053, NIOSH proposes to use nonparametric hypothesis tests to determine if there are differences between groups of workers. NIOSH bases the sample size for these tests suggestions contained in the *Draft ProUCL Technical Guide* (EPA 2010). This general advice is contained within the following passages.

> *The two-sample hypothesis testing approach is used when many (e.g., exceeding 8 to 10) site, as well as background, observations are available. For better and more accurate results with higher statistical power, the availability of more observations (e.g., exceeding 10-15) from each of the two populations is desirable, perhaps based upon an appropriate DQO process, as described in an EPA guidance document (2006).* (EPA 2010, Sect 1.3)

And later:

> *As mentioned before, every effort should be made to collect as many samples as determined using DQO processes as described in EPA documents (2006).* [EPA 2010, Sect 1.6.2]

In addition to the general advice of 10–15 samples, the *Draft ProUCL Technical Guide* contains further advice to use the DQO process. Appendix B, Section B1.3.2, of the same document (EPA 2010) contains detailed instructions for determining the required sample size based on data variability and DQO parameters. Instructions for 1-sided and 2-sided tests are provided. NIOSH has made no effort to determine sample sizes that allow for sufficient power to detect differences given the available sample sizes and variability.

It is well known that the application of hypothesis tests may result in two types of decision errors; false rejection of the null hypothesis (Type 1 error) and false acceptance (Type 2 error). NIOSH considers only Type 1 errors in RPRT-0053. The hypothesis testing framework recommended in the multi-agency document MARSSIM (EPA 2000) provides a basis for determining the necessary sample size for controlling decision errors of both types. The *Draft ProUCL Technical Guide* used by NIOSH was developed to implement the recommendations contained in MARSSIM; the companion document for Superfund sites *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites* (EPA 2002); and other similar EPA guidance based on the DQO process. Although the *Draft ProUCL Technical Guide* may contain useful suggestions, it should be noted that the document has been in draft form for several years now and has not yet received final approval from EPA for publication. MARSSIM and the CERCLA guidance have received approval following extensive peer review processes. Several discussions in the *Draft ProUCL Technical Guide* refer to a “revised” version of the 2002 CERCLA guidance document, but the revised version could not be located in an EPA search and final approval status is unknown.

The statistical methods developed for MARSSIM and Superfund are particularly useful in this discussion. The statistical problems associated with clean-up and decommissioning of NRC-licensed and Superfund facilities require comparison of data from two independent populations, samples from the decommissioned site (Group A), and those from nearby reference areas (Group B), to decide if the site concentrations exceed those in the reference areas. In the coworker model Group A are, say, the CTWs and Group B all other workers. Under constant revision, MARSSIM is the product of a multi-agency effort to provide detailed guidance for using decision theory for comparing the two populations.

MARSSIM recognizes that the use of hypothesis tests for decision-making requires consideration of both types of decision errors; Type 1 and Type 2. MARSSIM also recognizes that there is a trade-off between Type 1 and Type 2 error rates. When designing data collection for a hypothesis test, the required sample size is determined by specifying both the Type 1 error rate α and the Type 2 error rate β for the decision. Lower error rates require larger sample sizes. A realistic assessment of the sample sizes required to achieve pre-specified error rates depends on knowledge of the underlying variability in the two datasets. RPRT-0053 makes no use of this information.

The following EPA guidance for setting Type 1 and Type 2 error rates during the DQO process has survived the test of time. Formalized first in the CRCLA guidance for Superfund sites (EPA 2002, Section 3.2), the guidance is repeated almost verbatim in Chapter 6 of the *Draft ProUCL Technical Guide* (EPA 2010).

> *The selection of appropriate levels for decision errors and the resulting number of samples is a critical component of the DQO process that should concern all stakeholders. Because there is an inherent tradeoff between the probabilities of committing Type I or Type II error, a simultaneous reduction in both types of errors can only occur by increasing the number of samples. If the probability of committing a false positive error is reduced by increasing the level of confidence associated with the test (in other words, by decreasing α), the probability of committing a false negative is increased because the power of the test is reduced (increasing β).*
>
> *Typically, the following values for error probabilities are selected as the minimum recommended performance measures (EPA, 1990 and EPA, 2002).*
>
> > *· For the Background Test Form 1, the confidence level should be at least 80% (α = 0.20) and the power should be at least 90% (β = 0.10).*

In MARSSIM terminology, Background Test Form 1 is a 2-sample test with a null hypothesis that there is no difference between the site and background reference areas. These recommendations indicate that the power of the test should be of greater importance than the confidence level when the test has a null hypothesis that there is no difference between the populations. To facilitate this goal, the confidence level may be relaxed to 80% so the test can achieve greater power to detect differences between the two populations.

Finding No. 5: The methods in RPRT-0053 require a high level of confidence before deciding that the two worker groups are significantly different. The requirement for a high level of confidence in this decision is not claimant favorable when using a null hypothesis of "No Difference." The power of the tests to detect differences given the limited quantity of available data has not been established. The Data Quality Objectives (DQO) process should be used to balance Type 1 and Type 2 decision errors.

RPRT-0053 discusses hypothesis testing as though there were only two possible outcomes of the test. When the sample sizes are fixed by circumstance, there are, in fact, three possible outcomes, not two. The three outcomes are:

(1) Accept the Null hypothesis of No Difference
(2) Reject the Null hypothesis of No Difference
(3) No conclusion can be reached from these data

This 3-way list characterizes the "win," "lose," or "tie" nature of the decision-making under uncertainty. The process is best described in terms of the gray region for the test. The gray region is related to item #3 in the list above.

4.1.1 The Gray Region for a Statistical Test

Figure 6 shows an example of the gray region for a test of the No Difference null hypothesis. In this figure, the probability that the test will reject the null hypothesis is plotted on the vertical axis. The difference between the two populations is plotted on the horizontal axis expressed as a multiplier. Below the gray region, the difference between the two distributions is sufficiently small that the test almost always will accept the null hypothesis of No Difference. Above the gray region, the difference between the two distributions is sufficiently large that the test almost always will reject the null hypothesis of No Difference. Other features shown in Figure 6 will be discussed in the following sections.

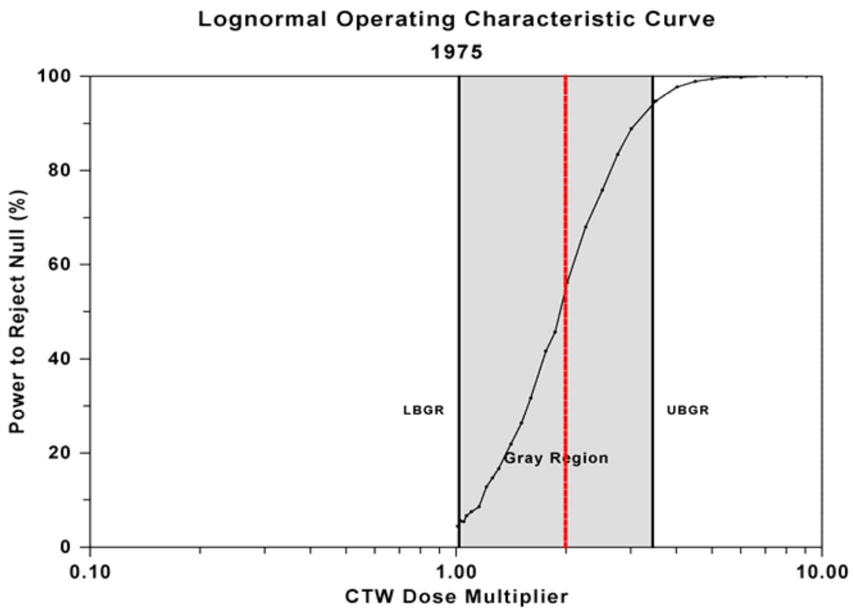

Figure 6. Simulated Test Performance Plot and Gray Region for 2-Sample t-Test for Difference between Coworker Model Lognormal Distributions for CTW and non-CTW Neptunium Whole-Body Counts in 1975

The question of the statistical significance of any observed difference depends on the sample sizes and underlying variability of the populations. With small sample sizes and high variability, it is unlikely that the observed difference will be found significant unless the difference is very large. Alternatively, when the sample sizes are large and the variability is low, the same difference may be statistically significant. If a test concludes that there is no significant difference, this should not be taken as evidence that there is no difference, but rather that the data are insufficient to decide if there is a difference. The statistical significance issue is not actually a question of whether there is a difference in the two datasets, but whether the quantity of data is sufficient to resolve the difference given the variability in the data. The power of resolution of the test, given the quantity and variability of the available data, is measured by the width of the gray region.

Using the somewhat cumbersome MARSSIM terminology, the gray region shown in Figure 6 is bounded on the left by the Lower Bound of the Gray Region (LBGR), and on the right by the Upper Bound of the Gray Region (UBGR). The distance between the LBGR and the UBGR is the width of the gray region. The gray region depicts a range of differences Δ that is too small to resolve with the available number of samples due to the level of variability. The width of the gray region defines the minimum detectable difference (MDD) for the test under the given conditions. Differences smaller than the MDD cannot be used to derive a conclusion about whether the samples are from different populations or not, given the available sample sizes. The pitfalls of using a statistical test to resolve differences smaller than the MDD are well known. They bear repeating here for the record.

The width of the gray region depends primarily on two factors, (1) the available number of samples and (2) the variance of the sample values. The gray region is a useful planning tool for determining the required sample size for a statistical test, since the width of the gray region for the test is reduced as sample size is increased. If the null hypothesis of No Difference is used, the width of the gray region determines the power of the test to detect a difference and reject the null hypothesis. When the number of samples is determined in advance of data collection, the variability is fixed by circumstance, but the number of samples may be adjusted to achieve acceptable test performance given the anticipated level of variability. Instructions for doing this step of the data collection planning stage are contained in MARSSIM.

In retrospective analysis of data, the gray region is also a useful tool for evaluating the performance of a test applied with sample sizes that are fixed and cannot be increased. When both the number of samples and the sample variability are fixed by circumstance, the width of the gray region is also fixed. In this case, it is necessary to determine if there is sufficient power in the available data to detect differences of the size of interest. One tool recommended in MARSSIM for analyzing the power of a hypothesis test is the test performance plot. This curve and its use in decision-making in retrospective analyses are discussed in the following section.

4.1.2 Test Performance Plots

In this section, we use test performance plots to explore how large the differences between the two groups could be for NIOSH to still conclude that there is no significant difference. The discussion includes examples based on the application of the RPRT-0053 hypothesis testing methodology to distinguish between the CTW and other onsite, non-CTW neptunium exposure distributions in RPRT-0056.

Finding No. 6: For many years, given the small number of CTW data points, the tests cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/non-CTW ratio of geometric means. Larger differences have a 95% or better chance of detection. Smaller differences would be in the "gray region" for the test, sometimes detected, sometimes not. Overall, SC&A concludes that the NIOSH method of concluding that there are no significant differences would often lead to very claimant-unfavorable results.

The test performance plot of a test for differences between two lognormal distributions is another feature shown in Figure 6. This test performance plot is for a 1-sided, 2-sample t-test[1] applied to test for differences between two lognormal distributions. The intersection points where the test performance curve crosses the left and right boundaries of the gray region define the test performance parameters α and β for the test. The Type 1 error rate for the test (α) is measured by the value of the test performance curve at the left edge of the gray region where Δ is equal to the LBGR. NIOSH has set the value of α at 5% by the decision to use a hypothesis test at the 95% level of confidence. This ensures that the Type 1 percentage error rate α is controlled at a maximum of 5% when there are, in fact, no differences between groups. This means that the null hypothesis will be falsely rejected in less than 5% of applications of the test procedure.

A higher Type 1 error rate will result in a lower Type 2 error rate and vice versa. If the Type 1 error rate is controlled at no more than 5%, it is likely that the Type 2 error rate will be large, unless an adequate number of samples are available in each group to provide sufficient power to detect differences. In the presence of highly variable data and many nondetects, a small sample size may result in unacceptably high Type 2 error rates.

The Type 2 error rate for the test (β) is measured by 100 minus the value of the test performance curve when Δ is at the UBGR at the right edge of the gray region. The value of β is of greater interest in this analysis as it defines the power of the test to reject the null hypothesis. For a given sample size and test level α, the power of the test depends on the variability in the data. Due to this dependency, it is not possible to define a single sample size that will be adequate for all problems. NIOSH does not report the Type 2 error rate when the tests recommended in RPRT-0053 are applied to the neptunium CTW/non-CTW comparison in RPRT-0056.

The sample sizes used by NIOSH in the preparation of RPRT-0056 are quite small. As few as 6 detected values were available for CTW modeling in some years. RPRT-0056 compares the two groups of workers using annual datasets containing values for as few as 11 CTWs in the 1977 neptunium coworker model comparison. Three OPOS values in this year are nondetects, leaving only 8 values to use in the regression. In 1985, 1986, and 1988, only 13 CTW values are available for the comparison, and in each year, there are 6 or 7 nondetects, representing about one-half of the values. These sample sizes are quite small and the gray region is expected to be very large in these years.

In all years with small sample sizes, NIOSH finds no significant evidence for rejecting the null hypothesis at a 95% level of confidence. NIOSH concludes from these tests that there is no

[1] In RPRT-0056, the hypothesis tests are conducted using 2-sided tests at a 95% level of confidence. In this discussion, only 1-sided tests are considered. The use of 1-sided tests is more relevant and claimant favorable. The choice of 1-sided versus 2-sided tests is discussed in Section 4.2.1.

statistically significant difference between the two groups of workers. In truth, the lack of rejection in these years is due to exceedingly small sample sizes for CTWs. The test result provides no statistical evidence one way or the other concerning the difference between the two groups. In many years, the test outcomes are in the gray region.

4.1.3 Simulated Test Performance Plots

Due to the concerns raised by the small sample sizes used in RPRT-0056, the power of the recommended tests was evaluated by studying an example application of the proposed hypothesis testing methodology. A simulation analysis was conducted to evaluate the performance of the test procedures when applied to neptunium at SRS. Test performance was measured by examining the test performance plot for the test. Two types of tests were considered in this example; parametric and nonparametric. The first simulation is conducted using the t-test to compare the ROS-estimated lognormal distributions for CTW and non-CTW neptunium WBC data contained in Appendix A, Tables A-18 through A-51 of RPRT-0056.

In the first example, samples of the appropriate sizes are drawn from the CTW and non-CTW distributions and compared using a two-sample t-test on the logarithms of simulated values from the two distributions. This simulation is based on ideal conditions—data truly are from a lognormal distribution, there are no nondetects, and all samples are positive, permitting calculation of the logarithms of the samples. The logarithms of the samples have normal distributions. The two-sample t-test is expected to perform well under these ideal conditions. The power of the two-sample t-test under these conditions represents an upper bound on the power of tests for comparing the two groups under less ideal conditions. Any test applied under more realistic conditions with data containing outliers and nondetects would be expected to have lower power than is observed in this parametric simulation. The second simulation is based on nonparametric tests.

The test performance plot summarizes test performance by calculating the probability of rejecting the null hypothesis as a function of Δ, the difference in the locations of the two distributions. With lognormal-like data, a suitable definition of the difference in location is $\Delta = log[L(f_2)/L(f_1)]$ or a multiplicative factor $L(f_2)= e^{\Delta} L(f_1)$. Here, the location of a distribution L[f] is defined as the expected value, the median, or any upper percentile of the distribution. Given two lognormal distributions with the same GSD, all three definitions of location yield the same ratio and hence the same $\Delta=log(GM_2/GM_1)$. The quantity e^{Δ} is a multiplier that determines the location of the CTW distribution relative to the non-CTW distribution in multiplicative terms.

The lognormal distributions used in the first simulation are the distributions estimated for the comparison of CTW and non-CTW whole-body counts (WBCs) for 1975 shown in Figures A-22 and A-39 in Attachment A of RPRT-0056, respectively. The lognormal distribution for the non-CTW group has a GM of 0.43 nCi with a GSD of 4.8. This distribution was estimated using 64 OPOS values, of which 25 are nondetects, leaving 39 measured values (39 workers, 1 OPOS value per worker) for the ROS calculations. The lognormal distribution for the CTW group has a GM of 0.86 nCi with a GSD of 4.5. The CTW distribution was estimated using 21 CTW OPOS values, of which 4 are nondetects, leaving 17 measured values (17 workers, 1 OPOS value per worker) for the ROS calculations. These two distributions have approximately the same GSD, another ideal condition for the two-sample t-test. The CTW GM is 2 times higher than the CTW GM. Since the GSDs are about the same, the ratio of the mean values and the ratios of the 84th and 95th percentiles have similar values. Although the distributions differ by a factor of 2, the results in RPRT-0053 using the Peto-Prentice test to compare these two groups in 1975 show no significant difference at the 95% confidence level.

The test performance plot for the two-sample t-test used to test for differences between the two lognormal distributions is shown in Figure 6. The horizontal axis in Figure 6 labeled the "CTW Dose Multiplier" measures the difference in location of the two distributions using the location multiplier e^{Δ} . The CTW multiplier would be of particular interest if an "adjustment factor" approach were used to ensure claimant favorability for CTW claimants. The vertical bar in Figure 6 is drawn at the actual CTW multiple of 1.99 in 1975.

The vertical axis is the power of the test, i.e., the probability that the test will reject the null hypothesis, expressed in percentage terms. The test performance curve plotted as an ascending curve in the figure defines the power of the test at each value of the CTW dose multiplier e^{Δ}. The test performance curve rises from α at the LBGR to (100-β) at the UBGR.

In sample design problems, the usual question is how large of a sample is required to detect differences that are as large as or larger than the size of the gray region. In this case, the sample variability and the desired Type 1 and Type 2 error rates α and β are fixed in advance. The width of the gray region is adjusted by determining the sample size required to detect the MDD of interest. At the selected sample size, the test performance curve will reach the desired power of (100-β) when the CTW dose multiplier e^{Δ} is at the UBGR. In Figure 6, the β for the test is set at 5%.

In retrospective analysis of power, the sample sizes and the variability are known and the Type 1 error rate α is specified by selecting the confidence level used for the test. In this case, the power (100-β) and the width of the gray region are unknown. When the Type 1 error rate is controlled at a maximum of α at the LBGR, it is claimant favorable to require an equivalently stringent maximum Type 2 error rate of β at the right edge of the gray region when the difference Δ is as high as or higher than the UBGR. Then the test performance curve is used to determine the UBGR, which is equal to the value of Δ where the test performance curve equals (100-β). Since the β for the test is set at 5%, this ensures that there will be a high chance of rejecting the null hypothesis if the difference Δ is at or above the UBGR.

The test performance curve is estimated by simulation and used to determine the UBGR for the test, i.e., the magnitude of difference that can be detected reliably with an error rate of less than β given the available sample sizes. The parametric simulation is summarized by asking the question: *How much higher would the CTW lognormal distribution have to be for the test to have sufficient power to reject the null hypothesis of No Difference?* When the sample size in one or both groups is small, the difference between the distributions must be large before the null hypothesis will be rejected.

The simulation was conducted by scaling the CTW lognormal distribution by e^{Δ}, while leaving the non-CTW distribution unchanged, then using the t-test to test for a significant difference between the two distributions at each selected value of Δ. The simulation conducted 3,000 iterations of the test at each value of Δ, and then calculated the percentage of iterations where the null hypothesis was rejected by the test. This percentage is an estimate of the test performance when the difference is Δ. The width of the gray region is defined by the value of Δ when the test performance curve reaches the desired power of (100-β).

Table 4 shows the results of the analysis of differences between the CTW and non-CTW neptunium WBC OPOS datasets for all years with coworker model lognormal distributions reported in RPRT-0056. The two-sample t-test and the WRS test were used to test for differences between the two distributions. The table shows the number of OPOS values in each group in each year, the number of samples above the detection limit (detects) used to fit the lognormal models, and the observed ratio of the GMs. The right-hand column shows the upper bound of the gray region for the tests for the given sample sizes. In most years, the WRS test has slightly less power (and thus larger UBGRs) than the t-test. This is to be expected given the normality of these ideal datasets. The smallest upper bounds occur in the earliest years when the sample sizes were large. In the years 1961 and 1962, there are 47 and 40 CTW samples, respectively. Here the sample sizes are sufficiently large that the tests reliably detect differences larger than a factor of 1.5 (i.e., differences of 50% or larger). In 1963, the number of CTW OPOS values is 28. This smaller sample size results in an increase of the UBGR from 1.5 to a factor of 2. In this year, differences larger than 100% can be reliably detected. Differences smaller than 100% would not be reliably detected.

NIOSH did not do a hypothesis test in RPRT-0056 for the first 3 years shown in Table 4. In 1962 and 1963, the observed ratio of GMs significantly exceeds the upper bound of the gray region indicating a significant difference in these 2 years that was not analyzed by NIOSH.

Later years in Table 4 have much smaller sample sizes and much higher values for the UBGR. In all years after 1963, the difference in GMs would have to exceed the UBGR to indicate a significant difference. In approximately one-half of the years after 1963, the difference would have to exceed a factor of 3 to 4 for a significant difference with these sample sizes. In all remaining years, except the anomaly in 1985, the difference would have to be as large as a factor of 4 to a factor of over 10 before the WRS test will indicate a significant difference. The UBGRs shown in Table 4 are for tests with lognormal data with no nondetects or extreme outliers. The UBGRs for the nonparametric Peto-Prentice test with nondetects are expected to be higher than the bounds shown in Table 4. Factors as large 10 may be required to show a significant difference.

Table 4. Sample Sizes, Ratio of Geometric Means, and Upper Bound of the Gray Region (UBGR) for Tests of the No Difference Hypothesis

	Number of Samples		Number of Detects		Ratio	Upper Bound of Gray Region (UBGR)	
Year	**non-CTW**	**CTW**	**non-CTW**	**CTW**	**$GM_{CTW}/GM_{non\text{-}CTW}$**	**t-Test**	**WRS Test**
1961	252	57	47	6	1.0	1.5	1.5
1962	734	175	40	8	4.9	1.5	1.5
1963	362	82	28	11	3.2	2.0	2.1
1974	58	10	54	9	1.0	3.5	3.8
1975	64	21	39	17	2.0	3.4	3.8
1977	43	11	27	8	0.6	8.4	10.2
1978	73	19	49	10	0.6	3.9	4.6
1979	55	12	39	7	2.7	4.9	4.6
1980	87	19	30	6	1.1	3.8	4.3
1981	99	23	36	7	1.6	4.4	4.5
1983	82	24	23	12	1.1	3.2	3.4
1984	92	25	31	9	1.0	3.8	4.1
1985	62	13	23	6	0.2	29.0	57.4
1986	65	13	22	7	2.7	5.5	5.6
1987	81	15	51	7	3.4	6.4	8.6
1988	77	13	31	7	3.7	5.0	5.0
1989	69	17	31	9	0.3	7.9	11.0

One improvement that should be noted; the MCPT approach proposed in RPRT-0053 is based on samples from two mutually exclusive populations of workers. In RPRT-0049, the MCPT procedure compared coworker samples for one group of workers with samples drawn from the set of all workers. The current report properly compares the parameters of the lognormal distributions estimated separately for each group of workers.

There are problems with implementation of the hypothesis testing strategy. Examples to illustrate these problems are found in the derivative publications that implement the RPRT-0053 methodology. For example, on page 8 of RPRT-0056, NIOSH states:

> ***CTWs are potentially subject to different bioassay practices than other workers. CTWs, many of whom are contractors, commonly submit bioassay samples after suspected uptakes and at the completion of jobs. This is in contrast to other workers, especially those employed directly by the prime contractor, who are more likely to be on a routine bioassay program in addition to submitting bioassay samples after suspected uptakes.*** *A post-job bioassay is more likely to be soon after an uptake, either suspected or unidentified, than is a routine bioassay sample and thus more likely to have a larger result. This potential difference in how the strata are monitored for intakes would result in higher results for CTWs compared to the other strata.* [Emphasis added.]

Essentially the same comment appears on page 9 of RPRT-0058. Although the two groups are defined to be mutually exclusive, the bioassay program for the two groups should be the same. This is in direct contrast to the above quote, which clearly indicates the CTW group had more results from accidents and the non-CTW group had more routine results. If a worker had a suspected intake, it is likely that his/her monitoring would be geared to detect exposures from the particular radionuclide that caused the potential exposure. Hence, the entire dose reconstruction of a CTW would be different from a non-CTW if this sampling protocol was followed, since the CTW samples should not be construed as routine.

The final sentence in the quote above suggests that the CTW bioassay results should be considered as bounding. If this were true, the mixture of bounding results for CTWs and routine results for non-CTWs in the MCPT is questionable, as it raises the question of exactly what the parameters estimated in the 10,000 permutations represent. Each permutation contains a varying proportion of bounding and routine assay results.

A second concern with the hypothesis test strategy is that cases may arise when both groups contain the same worker. For example, in the derivative report RPRT-0056 (p. 12), NIOSH states the following [essentially the same passage appears in RPRT-0058 (page 12)]:

> *Because it was possible for a worker to change jobs during the course of a single evaluated period, it is possible that a worker would have some samples identified as nonCTW and others as CTW in the same period. Therefore, one person might have as many as four different OPOS results, one each for the AMW, CTW, nonCTW, and nonCTW+unk strata.*

When the radionuclide is long-lived, the OPOS values generated in each group for that worker will be strongly correlated. For example, Np has a half-life of 7.8 E8 days and is Type M. If the worker changes jobs during the time period, and if some of his/her bioassay results are for the CTW group and other results are for the non-CTW group, there will be an influence of Np exposures in one job on the results from the other job. This happens because Np will stay in the body for a long period after the first intake.

Table 5 shows the retention of Np-237 in the body after a single intake of the nuclide as a function of time after the intake. Over the course of a year, the body burden is reduced by less than half. The bioassay results for this worker are not independent. For example, if a worker in the CTW group was exposed then changes jobs and becomes part of the non-CTW group in the same year, the non-CTW results for this worker will carry the influence of the previous exposure as a CTW. This means that the bioassays recorded for this worker in each group will be strongly correlated, as will be the MPMs calculated from these assays.

The statistical tests for comparing the CTW strata and non-CTW strata require that the samples in each group be independent and they are not in this case. This correlation not only violates the assumptions of the tests, but also creates a bias toward a decision of “No Difference” between the two groups.

Table 5. Retention of Np-237 in the Body after a Single Intake

Time(days)	Whole Body
10	7.15E-02
20	6.63E-02
30	6.24E-02
40	5.92E-02
100	4.95E-02
180	4.46E-02
300	4.10E-02

Finding No. 7: The statistical tests for comparing two strata require that the samples in each group be independent. If a worker in one group is exposed to radionuclides with long retention in the body, then changes jobs and becomes part of the other group in the same year, the OPOS values are correlated for this worker. This correlation not only violates the assumptions of the tests, but also creates a bias toward a decision of "No Difference" between the two groups.

4.2 DESCRIPTION OF THE MONTE CARLO PERMUTATION TEST

An example of the MCPT procedure is shown in Figures A-2 through A-4 of RPRT-0053. Figures A-2 and A-3 show ROS plots for two groups of workers with two different job classifications and their estimated lognormal distribution parameters. There are 219 workers in Group A and 113 in Group B. Approximately 40% of the values in each group are nondetects.

The ROS plots of the ordered data approximately follow the regression line in both figures except in the extreme upper tails, where the data points fall below the regression line, indicating that the lognormal distribution may be an appropriate and claimant-favorable probability model for these datasets. As shown in Table 6, the GM is 0.66 for Group A and 0.93 for Group B, with a GSD of 4.3 for Group A and 3.8 for Group B. Based on these lognormal distributions, Group A has a mean of 1.94 and a standard deviation of 5.3, while Group B has a mean of 2.23 and a standard deviation of 4.9. The 95th percentiles are 7.4 and 8.2 for Groups A and B, respectively. In this example, Group A has a larger GSD and standard deviation, but a smaller GM, mean, and 95th percentile. The largest percentage difference is for the GM, but all parameters are within a range of ±30%. This is a first indication that it will be difficult to distinguish significant differences between the two distributions.

Table 6. Comparison of Lognormal Distributions for Group A and Group B

Lognormal Parameter	Group A	Group B	Difference (%)
GM	0.66	0.93	-29.0
GSD	4.33	3.75	15.4
Mean	1.94	2.23	-13.3
Standard Deviation	5.33	4.86	9.5
95th Percentile	7.37	8.19	-10.1

A Monte Carlo simulation was conducted to examine the differences between these two lognormal distributions when used in the coworker model. Although the presence of nondetects is an important consideration when estimating the lognormal distributions for each group, the application of the fitted distributions within the coworker model would not include nondetects.

Hence, the simulation compares the fitted lognormal distributions with and without nondetects in the simulated sample values. While simulation with nondetects would be useful for examining the estimation process, simulation without nondetects is of greater interest, as it addresses the question of whether there will be differences in the predicted values when the estimated lognormal distributions are used to simulate worker exposures.

In the simulation, a random sample of the appropriate size was drawn from each of the two lognormal distributions: $x_1,..., x_{n_A}$ for Group A; and $y_1,..., y_{n_B}$ for Group B. Here, n_A represents the number of workers in Group A and n_B the number in Group B. The ROS method was used then to estimate the GM and GSD of the two samples using the ROS procedures described in RPRT-0053. Although use of the ROS procedure is necessary when there are nondetects present in the sample, it may also be used with no nondetects. In this case, the estimates of the slope and intercept of the regression line are uncorrelated, since the sum of the independent variable values in the regression is zero.

The experiment was then repeated for 1,000 iterations to simulate the range of uncertainty in the lognormal parameter estimates for the given sample sizes. The results are shown in Figure 7, which contains scatter plots of the estimated GM and GSD for each group in each iteration.

With no nondetects in the samples, the uncertainty distributions of the lognormal parameters appear to be different, but distributions of the GM and GSD clearly do overlap to a small degree. When a hypothesis test is applied to test whether there is a significant difference between the two distributions, the test results will be determined by the level of confidence required of the test. Setting a high level of confidence would result in the conclusion that, for the given sample sizes, no significant difference is observed, while a lower level of confidence would indicate a significant difference between the groups is observed. Since there is always a trade-off between Type 1 and Type 2 errors when conducting a hypothesis test with fixed sample sizes, setting a higher level of confidence would reduce the power of the test to detect a significant difference, while setting a lower level of confidence would increase the power of the test to detect a significant difference.

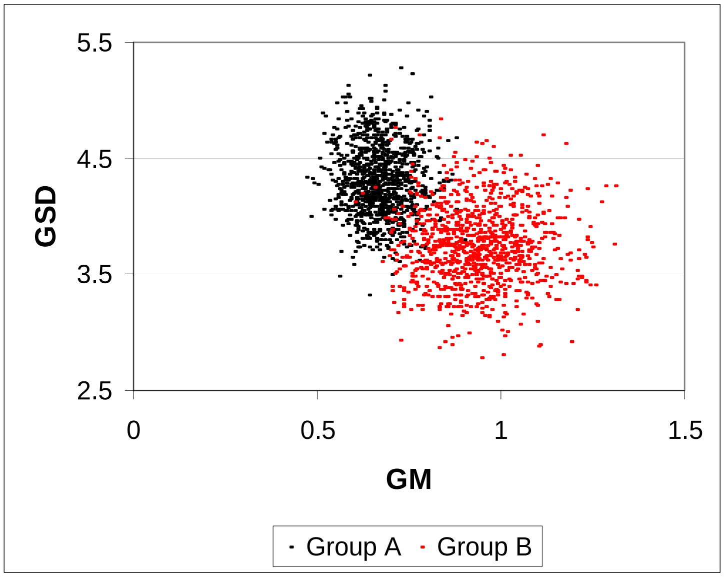

Figure 7. Scatter Plot of ROS Estimates of GM and GSD for Samples of Size n_A and n_B from the Lognormal Distributions for Group A and Group B

(1,000 iterations)

Claimant favorability is always of concern when setting the standards for the level of significance. NIOSH has proposed that strong evidence is necessary before any differences between groups of workers should be considered in the coworker model. In the examples in Sections 5.1 and 5.2 of RPRT-0053, and in subsequent applications to neptunium (ORAUT 2012b), mixed fission and activation products (ORAUT 2012c), and exotic trivalent radionuclides (ORAUT 2012a) at SRS, NIOSH conducts the statistical tests for a significant difference at the $\alpha = 0.05$ probability level requiring a 95% level of confidence. A higher level of confidence makes it more difficult to decide if there are differences between the two groups. A 90% level of confidence for the MCPT would be more claimant-favorable. The issue of confidence levels and power are further addressed in Sections 4.1 and 4.2.

A high level of confidence in the test for differences requires a high standard of proof before accepting the conclusion that there are differences within the worker population. Conversely, a high standard reduces the power to detect differences of any magnitude. NIOSH has not provided any measure of the power of the MCPT to detect differences, although one example where a significant difference is found is presented in the report. This deficiency should be corrected before the MCPT is adopted as an appropriate testing procedure. To provide more information to decision makers, NIOSH also should report the p-level for each test, rather than only providing the test outcome for a 95% level of confidence. This would permit discussion of exactly what level of certainty is required before admitting to differences.

Finding No. 8: Although one example where a significant difference is found is presented in the report, NIOSH has not provided any measure of the power of the hypothesis test procedure to detect differences within the worker population. This deficiency should be corrected before the test is adopted as an appropriate procedure for coworker models. Conducting the tests at a 90% level of confidence would be claimant favorable.

The distribution of ROS estimates of the GM and GSD for each group are shown in Figures 8 and 9. In both cases, the distributions of the parameters overlap, although the distributions of the GMs have better separation.

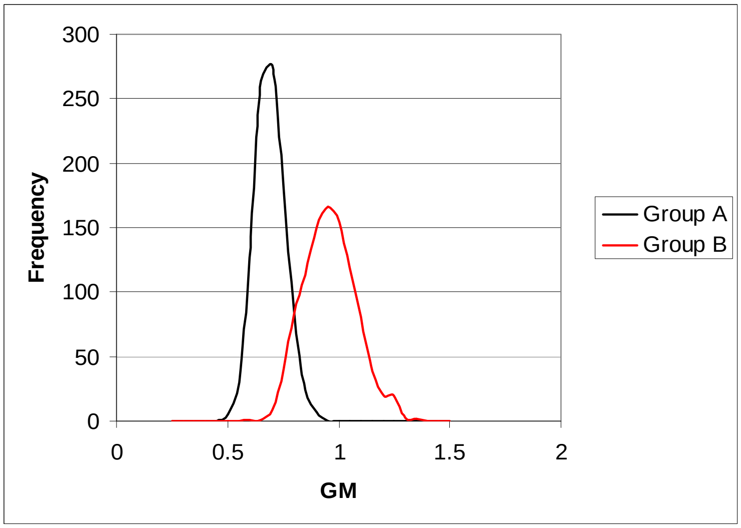

Figure 8. Distribution of ROS Estimates of GM for Samples of Size n_A and n_B from the Lognormal Distributions for Group A and Group B

(1,000 iterations)

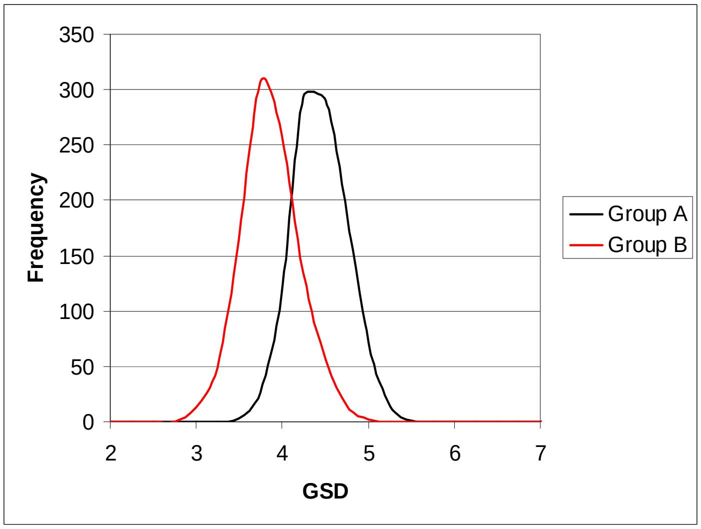

Figure 9. Distribution of ROS Estimates of GSD for Samples of Size n_A and n_B from the Lognormal Distributions for Group A and Group B

(1,000 iterations)

When there are no nondetects present, there are many ways to determine if the parameter distributions shown in Figures 7, 8, and 9 are different. Figure 10 shows a scatter plot of the differences (Group A minus Group B) of the simulated GM and GSD values in Figure 7. The difference in each iteration in the estimated GM of the two samples is plotted on the horizontal axis and the difference in the GSD on the vertical axis. The actual value of the difference in the GM and GSD of the lognormal distributions is shown as a triangle in this figure and the origin of the plot is shown as a circle. The uncertainty distribution of the parameter estimates is centered around the true value and clearly includes the origin, indicating that the observed differences are not significantly different from the origin at (0,0). The simulation is found to agree with NIOSH's conclusion that there is no significant difference between the two groups.

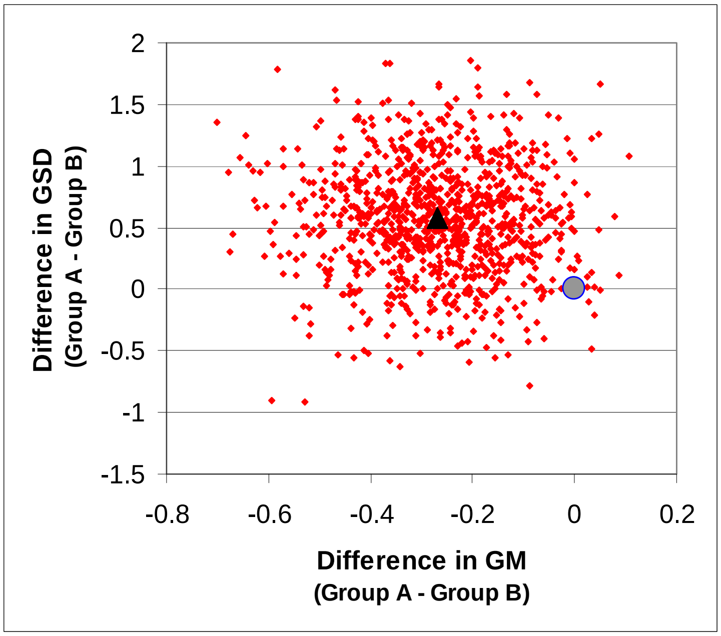

Figure 10. Distribution of ROS Estimates for Differences in GM and GSD for Samples of Size n_A and n_B from the Lognormal Distributions for Group A and Group B

(1,000 iterations)

NIOSH has selected to use a different approach to test for a significant difference, one that is not based on simulation. The NIOSH approach is data-driven, based on repeated permutations of the original data values (including the nondetects). Two examples are provided demonstrating the application of the test when two groups of workers are compared. One of these examples with a borderline decision was examined in this review using a simulation approach to illustrate and validate the permutation test results.

The MCPT is applied to test for a significant difference in the ROS lognormal parameters estimated for the two groups. In this permutation test, the combined original samples from the two groups are arranged in 10,000 permutations filling the two samples without regard to job classification. The distribution of parameter differences from the repeated permutations forms the null distribution, i.e., the distribution of parameter values that would result from samples of the given sizes drawn without knowledge of the job classification. The actual difference in parameters is then compared with the null distribution to see if the observed difference using the job classification information is significant. The null hypothesis for this test is that the distribution of the OPOS bioassay data is the same in Group A and Group B, and the alternate hypothesis is that the distribution of data is not the same. The null hypothesis is rejected at level α as the observed difference lies outside of an elliptical region containing 100(1-α)% of the permutations.

To continue with the Group A/B example above, the MCPT was applied first with no nondetects in either sample, then with nondetects in both samples. In the previous simulation, all samples were selected randomly from the two lognormal distributions. In permutation sampling, the samples are obtained by repeatedly drawing the samples from one set of urinalysis samples, which includes both Group A and Group B workers.

Figures 11 and 12 show the results of the permutation test with no nondetects and with nondetects, respectively. The difference in the estimated GM of the two samples is plotted on the horizontal axis, and the difference in the GSD on the vertical axis. The triangular point plotted in each figure represents the difference between the parameters of the two lognormal distributions shown in Table 6. The shape of the uncertainty distribution changes to an elliptical shape when nondetects are added. This is an indication of the correlation between the parameter estimates due to the unbalanced design when there are censored data. The presence of a large proportion of nondetects makes it more difficult to find any significant difference between groups.

In the MCPT, the distribution of the estimates of parameters from the permutation samples is centered at the origin (under the null hypothesis), while the observed difference lies near the edge of the null region. As expected, the difference between the two groups has only borderline significance both with and without nondetects.

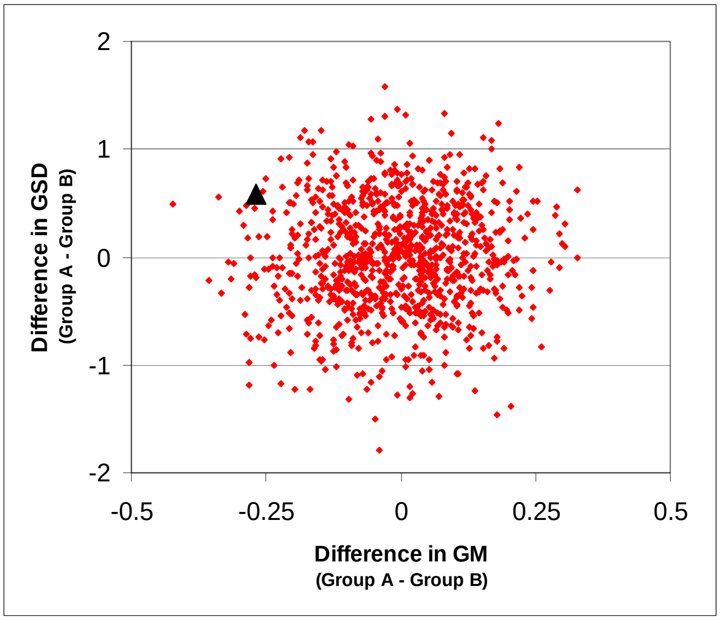

Figure 11. Permutation Test for a Significant Difference in ROS Estimates of GM and GSD with No Nondetects

(1,000 permutations)

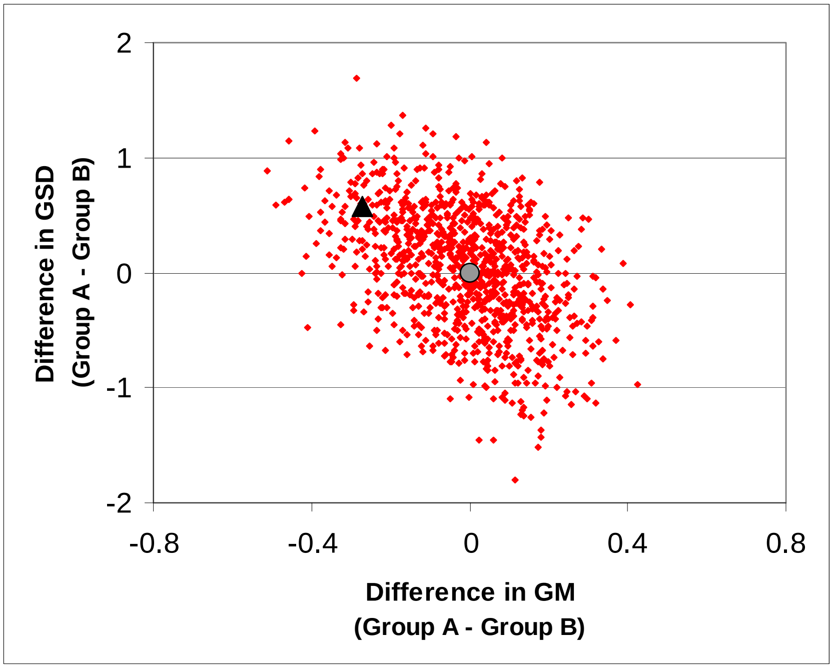

Figure 12. Permutation Test for a Significant Difference in ROS Estimates of GM and GSD with Approximately 40% Nondetects

(1,000 permutations)

If there are many nondetects in the datasets, the MCPT may fail when one or more of the permutations results in a sample containing only nondetects. The ROS method does not apply in this case, and other methods must be applied.

4.3 OTHER NONPARAMETRIC METHODS

In addition to the use of the MCPT for comparing the two groups by fitting lognormal distributions, RPRT-0053 discusses the use of more general hypothesis tests for comparing the groups. These nonparametric methods are applicable to a broader range of distributions than the lognormal distribution and will give results with only a small number of values above the detection limit. Again, the goal of the test is to determine if there is a significant difference between the two groups.

Many nonparametric methods have been developed for comparing two independent samples. Nonparametric methods do not require the assumption that data follow the lognormal distribution. Many of these methods are based on the relative ranks of the two datasets when samples from both groups are ranked in a single list. The two-sample rank tests determine if one group tends to have higher ranks than the other. One of the most commonly used rank tests is the WRS test. This test is based on a comparison of the sum of the ranks of the workers in each group. RPRT-0053 considers several nonparametric tests based on ranks, which are generalizations of the WRS test for use with nondetects.

4.3.1 Peto-Prentice Test

When there are nondetects present in either dataset or both, the ranks for the nondetects are assigned scores. There are many generalizations of the WRS test that differ in the manner these scores are assigned. Generalized WRS tests reviewed in RPRT-0053 include the Gehan test, the generalized Wilcoxon-Gehan test, and the Peto-Prentice test. These three generalized WRS tests are identical to the WRS test when there are no nondetects in the data.

The Gehan Test is the most commonly used generalized WRS test for comparing two datasets containing nondetects with the same detection level. Attachment B to RPRT-0053 reviews the discussion of the generalized WRS tests found in Helsel (2005). After review of the available generalized WRS tests, RPRT-0053 recommends the Peto-Prentice test as most powerful if there are nondetects with varying levels of detection in the dataset. This conclusion agrees with the evaluations by Helsel (2005, p. 150) and Latta (1981). The literature search does not include reference to the Brunner-Munzel test (Brunner and Munzel 2000). This nonparametric test is another generalization of the WRS test designed for comparisons of populations with different variances.

The Peto-Prentice test typically leads to the same conclusion as the MCPT when both are applicable. The Peto-Prentice test is recommended by NIOSH for use when the MCPT is not applicable. The null hypothesis proposed by NIOSH for this test is that the distribution of the OPOS bioassay data is the same in Groups A and B. The alternative hypothesis is that the distributions are not the same.

As noted above, the MCPT test fails when one or more of the permutations contains all nondetects. When the data contain a large proportion of nondetects, the MCPT is likely to fail and the Peto-Prentice test must be used. In the application of the RPRT-0053 methodology to neptunium (ORAUT 2012b), mixed fission and activation products (ORAUT 2012c), and exotic trivalent radionuclides (ORAUT 2012a) at SRS, only the Peto-Prentice test was used due the high proportion of nondetects in these datasets.

The Peto-Prentice test has several advantages over the MCPT:

(1) The Peto-Prentice test is a nonparametric test and does not require the assumption of a lognormal distribution necessary in the parametric models underlying the MCPT

(2) The Peto-Prentice test may be used in cases when the MPCT is not applicable due to a high proportion of nondetects

(3) The Peto-Prentice test may be applied to test for several different hypotheses, rather than only a test for a significant difference

The Peto-Prentice test is a generalization of the WRS test which is a test for the location of one distribution relative to the other. Tests of location may be applied using three different forms of the hypothesis test, which differ in terms of the null hypothesis (H_0). Three hypothesis test forms may be tested using the Peto-Prentice statistic z:

(A) H_0: The distribution of the bioassay data is the same for CTW and non-CTW
vs. H_A: the distribution of data is not the same for CTW and non-CTW

(B) The distribution of the bioassay data non-CTW is higher than for CTW
vs. H_A: the distribution of data for CTW is higher than for non-CTW

(C) H_0: The distribution of the bioassay data for CTW is higher than for non-CTW
vs. H_A: the distribution of data for non-CTW is higher than for CTW

Although the three test forms may appear similar, in practice, there are large differences between the three test forms in terms of claimant favorability. The differences arise because the null hypothesis is assumed true until the data provide sufficient evidence to reject the null hypothesis. If the sample size for one or both groups is too small, the test would not have sufficient power to reject the null hypothesis. Test form A is a 2-sided test. With a 2-sided test, the null hypothesis of "No Difference" is rejected if the CTW data are either significantly higher or lower than the non-CTW data. If the sample size is too small, the test may have insufficient power to reject the null hypothesis of No Difference. Using this test form, the null hypothesis is accepted due only to a lack of evidence in the data that proves the CTW are different. This is not claimant favorable, as it places the burden of proof on the claimants despite the known lack of sufficient data to provide such proof.

Test forms B and C are both 1-sided tests. In test form B, the null hypothesis is that the non-CTW data are higher than the CTW data. In this form of a 1-sided test, the null hypothesis is rejected if the CTW data are significantly higher than the non-CTW data. If the sample size is too small, the test may have insufficient power to reject the null hypothesis that the non-CTW distribution is at least as high as the CTW distribution. As with test form A, the null hypothesis may be accepted due only to a lack of evidence in the data to prove the CTWs are different from non-CTWs. Test form B is also not claimant favorable, as it places an unreasonable burden of proof on the claimant to show that the CTW data are higher than the non-CTW data despite the known lack of sufficient data. The 1-sided test form B is more relevant than the 2-sided test form A. Unlike test form A, test form B at least provides a clear answer as to whether the CTW are higher than the non-CTW data, which is the issue in question.

Test form C is also a 1-sided test. In test form C, the null hypothesis is that the CTW data are higher than the non-CTW data. In this form of a 1-sided test, the null hypothesis is rejected if the non-CTW data are significantly higher than the CTW data. If the sample size is too small, the test may have insufficient power to reject the null hypothesis that the CTW distribution is higher than the non-CTW distribution. Of the three test forms, only test form C is claimant favorable when the sample sizes are too small to provide clear evidence. Unless there is significant statistical evidence to the contrary, the null hypothesis that the CTW samples are higher than non-CTW should be accepted as claimant favorable.

In RPRT-0053, NIOSH applies test form A for both the MCPT and the Peto-Prentice test. The MCPT is not a test of location and it may be difficult to use test forms B and C for the MCPT. The Peto-Prentice test is a test of location and may be used with all three hypothesis test forms.

RECOMMENDATION:

NIOSH might consider reversing the null hypothesis for the Peto-Prentice test. The hypothesis tests applied by NIOSH to test for differences between CTWs and non-CTWs use a null hypothesis that is not claimant favorable, as it would place the burden of proof on the CTW claimants to prove a significant difference. The Peto-Prentice test is more generally applicable and may be applied using a claimant-favorable null hypothesis. The groups of workers with high suspected exposures should be considered different in the absence of strong evidence that they are not. This is more likely to result in a claimant-favorable model. Moreover, there is persuasive evidence provided by the analysis of SRS CTWs by job type and by area of work (SC&A 2010a, 2010b) that subgroups of CTWs are not drawn from the same distribution as non-CTWs. When the distributions of CTW subgroups are different from non-CTWs, CTW data by job type and area can be used to construct coworker models for their CTW peers. Of course, this requires sufficient data in each job/area category for which a coworker model is to be constructed.

5.0 REVIEW CHECKLIST

Table 7. Procedure Review Outline/Checklist

Document No.: ORAUT-RPRT-0053 Rev. 01	**Effective Date: 7/26/2012**
Document Title: Analysis of Stratified Coworker Datasets	
Reviewer: Harry J. Chmelynski	

No.	Description of Objective	Rating 1-5*	Comments
1.0	**Determine the degree to which the procedure supports a process that is expeditious and timely for dose reconstruction.**		
1.1	Is the procedure written in a style that is clear and unambiguous?	4	
1.2	Is the procedure written in a manner that presents the data in a logical sequence?	4	
1.3	Is the procedure complete in terms of required data?	2	Data are reduced to one statistic per worker per time period.
1.4	Is the procedure consistent with all other procedures that are part of the hierarchy of procedures employed by NIOSH for dose reconstruction?	4	
1.5	Is the procedure sufficiently prescriptive in order to minimize the need for subjective decisions and data interpretation?	3	
2.0	**Determine whether the procedure provides adequate guidance to be efficient in instances where a more detailed approach to dose reconstruction would not affect the outcome.**		
2.1	Does the procedure provide adequate guidance for identifying a potentially high probability of causation as part of an initial dose evaluation of a claim?	N/A	
2.2	Conversely, for claims with suspected cumulative low doses, does the procedure provide clear guidance in defining worst-case assumptions?	3	Worst-case analysis is not addressed. Recommended approach is "one size fits all."

No.	Description of Objective	Rating 1-5*	Comments
3.0	**Assess the extent to which the procedure accounts for all potential exposures and ensures that resultant doses are complete and based on adequate data in instances where the POC is not evidently clear.**		
3.1	Assess quality of data sought via **interview**:	----	
3.1.1	Is scope of information sufficiently comprehensive?	N/A	
3.1.2	Is the interview process sufficiently flexible to permit unforeseen lines of inquiry?	N/A	
3.1.3	Does the interview process demonstrate objectivity and is free of bias?	N/A	
3.1.4	Is the interview process sensitive to the claimant?	N/A	
3.1.5	Does the interview process protect information as required under the Privacy Act?	N/A	
3.2	Assess whether the procedure adequately addresses generic as well as **site-specific data** pertaining to:	----	
3.2.1	Personal dosimeters (e.g., film, TLD, PICs)	N/A	Site-specific data are not addressed.
3.2.2	In vivo/In vitro bioassays	N/A	
3.2.3	Missing dosimetry data	N/A	
3.2.4	Unmonitored periods of exposure	N/A	
4.0	**Assess procedure for providing a consistent approach to dose reconstruction regardless of claimants' exposures by time and employment locations.**		
4.1	Does the procedure support a prescriptive approach to dose reconstruction?	3	Specific methods are recommended for determining if exposures differ by job type or work area.
4.2	Does the procedure adhere to the hierarchical process as defined in 42 CFR 82.2?	4	

No.	Description of Objective	Rating 1-5*	Comments
5.0	**Evaluate procedure with regard to fairness and giving the benefit of the doubt to the claimant.**		
5.1	Is the procedure claimant favorable in instances of missing data?	2	RPRT-0053 recommends a one person - one sample (OPOS) approach to derive a single value for each worker by averaging over all urine samples collected in each time period. Time periods may be as long as 3 years. This procedure ignores variation within each individual worker's bioassay samples in the period and may result in underestimation of the geometric standard deviation (GSD) used in the coworker model.
5.2	Is the procedure claimant favorable in instances of unknown parameters affecting dose estimates?	2	Dose estimation requires a large number of parameters other than urine concentration estimates. The recommended procedures are not radionuclide specific, hence ignore these differences.
5.3	Is the procedure claimant favorable in instances where claimant was not monitored?	2	Requirement of 95% confidence before deciding that there are differences in the worker population is not claimant favorable.
6.0	**Evaluate procedure for its ability to adequately account for the uncertainty of dose estimates.**		
6.1	Does the procedure provide adequate guidance for selecting the types of probability distributions (i.e., normal, lognormal)?	2	MCPT procedure is based on the lognormal distribution. The nonparametric Peto-Prentice test is recommended for cases when MCPT cannot be run.
6.2	Does the procedure give appropriate guidance in the use of random sampling in developing a final distribution?	2	Procedure may be used to derive a single coworker distribution for all workers, regardless of work location or job type. Differences between worker groups are ignored in this approach.
7.0	**Assess procedure for striking a balance between the need for technical precision and process efficiency.**		
7.1	Does the procedure require levels of detail that can reasonably be accounted for by the dose reconstructor?	4	
7.2	Does the procedure avoid levels of detail that have only limited significance to the final dose estimate and its POC?	N/A	
7.3	Does the procedure employ scientifically valid protocols for reconstructing doses?	2	See Sections 3 and 4.

* Rating System of 1 through 5 correspond to the following: 1=No (Never), 2=Infrequently, 3=Sometimes, 4=Frequently, 5=Yes (Always). N/A indicates not applicable

6.0 CONCLUSIONS

In conclusion, the statistical methods proposed in ORAUT-RPRT-0053 are based on sound statistical methodologies, and the material is well presented. Several previously published statistical procedures for estimating the GM and GSD of coworker urinalysis datasets are reviewed.

In the application of the procedures recommended in RPRT-0053, the issue of completeness of the available coworker data has not been addressed. The methods proposed in RPRT-0053 for analyzing the coworker datasets require verification that (1) the available coworker data are representative of all groups of workers, and (2) the manner of use of the data is claimant favorable for the specific datasets to which the method is applied. A sound statistical methodology is subject to these two important caveats.

Due to the dependencies that exist in the ranked data, the R^2 for ROS does not have the usual interpretation. The recommendations in RPRT-0053 for using ROS do not address this concern.

The use of average values (OPOS) does not account for variability of the samples within the time period, and the procedure may result in a lower GSD used in the coworker models. This procedure may not be appropriate for certain radionuclides with long half-lives or long retention times.

The statistical tests for comparing the strata require that the samples be collected using the same or a similar protocol. However, there is evidence presented by NIOSH that the sampling protocol for CTWs at SRS was different than the protocol used for non-CTWs.

More than two strata would be required to properly characterize the varied worker populations at many sites, including SRS. Multiple comparisons when there are more than two strata may be possible, but could be complex and suffer from limits imposed by small sample sizes.

A high level of confidence is required before deciding that the two worker groups are significantly different. The requirement for a high level of confidence in this decision is not claimant favorable when using a null hypothesis of "No Difference."

The statistical tests for comparing strata require that the samples in each group be independent. If a worker in one group is exposed to radionuclides with long retention in the body, then changes jobs and becomes part of the other group in the same year, the OPOS values are correlated for this worker.

NIOSH has not provided any measure of the power of the hypothesis tests to detect differences within the worker population. This deficiency should be corrected before the tests are adopted as an appropriate testing procedure.

Given the small number of CTW data points at SRS, in many years the tests cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/non-CTW ratio of GMs. Overall, SC&A concludes that the NIOSH method of concluding that there are no significant differences based on the available data would often lead to very claimant-unfavorable results.

The MCPT applied by NIOSH to test for differences between CTW and non-CTW uses a null hypothesis that is not claimant favorable, as it places the burden of proof on the claimants to prove a significant difference. The Peto-Prentice test is more generally applicable and may be applied using a claimant-favorable null hypothesis.

REFERENCES

42 CFR 82. 2002. *Methods for Conducting Dose Reconstruction Under the Energy Employees Occupational Illness Compensation Program Act of 2000*, U.S. Code of Federal Regulations, Title 42, Chapter I, Subchapter G, Part 82, Subpart A, §82.2.

Brunner, E., and U. Munzel, 2000. "The Nonparametric' Behrens Fisher Problem: Asymptotic Theory and a Small-Sample Approximation, *Biometrical Journal*, Vol. 42, pp. 17–25.

Conover, W.J., 1980. *Practical Nonparametric Statistics, 2rd Ed.*, John Wiley & Sons, New York, New York.

David, H.A. and H.N. Nagaraja, 2003. *Order Statistics, 3rd Ed.*, John Wiley & Sons, Hoboken, New Jersey.

EPA 1990. *Guidance for Data Usability in Risk Assessment: Interim Final*, October 1990. EPA 540-G-90-008, PB91-921208. U.S. Environmental Protection Agency, Washington, DC.

EPA 2000. U.S. Environmental Protection Agency, U.S. Nuclear Regulatory Commission (NRC), et al., *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM). Revision 1.* EPA 402-R-97-016. Available at http://www.epa.gov/radiation/marssim/

EPA 2002. *Guidance for Comparing Background and Chemical Concentrations in Soil for CERCLA Sites*. EPA 540-R-01-003-OSWER 9285.7-41. U.S. Environmental Protection Agency.

EPA 2006. *Data Quality Assessment: Statistical Methods for Practitioners*, EPA QA/G-9S. EPA/240/B-06/003. Office of Environmental Information, Washington, DC.

EPA 2010. *ProUCL Version 4.1 Technical Guide (Draft) – Statistical Software for Environmental Applications for Data Sets with and without Nondetect Observations*. EPA/600/R-07/041. U.S. Environmental Protection Agency, Technology Support Center, Atlanta, Georgia.

Helsel, D.R., 2005. *Nondetects and Data Analysis*, Wiley Interscience, New York, New York.

Latta, R.B., 1981. "A Monte Carlo study of some two-sample rank tests with censored data," *Journal of the American Statistical Association*, Vol. 76, No. 375, pp. 713–719.

Looney, S.W., and T.R. Gulledge, 1985. "Use of the Correlation Coefficient with Normal Probability Plot," *American Statistician*, Vol. 39, No. 1, pp. 75–79.

Meeting Transcript 2012. Minutes of the November 1, 2012, meeting of the Subcommittee on Procedures Review, Advisory Board on Radiation and Worker Health. 2012.

Noreen, E.W., 1989. *Computer Intensive Methods for Testing Hypotheses*, John Wiley and Sons, New York, New York.

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

ORAUT 2006a. *Generating Summary Statistics for Coworker Bioassay Data,* ORAUT-PROC-0095, Oak Ridge Associated Universities Team, Cincinnati, Ohio. June 5, 2006.

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

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

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

ORAUT 2010b. *NIOSH Responses to Selected Findings from 3rd set of Procedures*, Oak Ridge Associated Universities Team, Cincinnati, Ohio.

ORAUT 2012a. *A Comparison of Exotic Trivalent Radionuclide Coworker Models at the Savannah River Site*, ORAUT-RPRT-0055, Rev. 00, Oak Ridge Associated Universities Team, Cincinnati, Ohio. July 2012.

ORAUT 2012b. *A Comparison of Neptunium Coworker Models at the Savannah River Site*, ORAUT-RPRT-0056, Rev. 00, Oak Ridge Associated Universities Team, Cincinnati, Ohio. August 20, 2012.

ORAUT 2012c. *A Comparison of Mixed Fission and Activation Product Coworker Models at the Savannah River Site*, ORAUT-RPRT-0058, Rev. 00, Oak Ridge Associated Universities Team, Cincinnati, Ohio. August 20, 2012.

ORAUT 2012d. *Analysis of Stratified Coworker Dataset*, Rev. 1. ORAUT-RPRT-0053, Oak Ridge Associated Universities Team, Cincinnati, Ohio. July 16, 2012.

SC&A 2007. *Findings from 3rd Set of Procedures*, S. Cohen & Associates, Vienna, Virginia.

SC&A 2010a. *Review of ORAUT-0075: Use of Claimant Datasets for Coworker Modeling for Construction Workers at Savannah River Site*, S. Cohen & Associates, Vienna, Virginia. January 2010.

SC&A 2010b. *Comparison of Claimant Tritium Samples from Construction Trade Workers and Non-Construction Workers at Savannah River Site*. S. Cohen & Associates, Vienna, Virginia. November 2010.

SC&A 2010c. *Draft Review of ORAUT-RPRT-0044: Analysis of Bioassay Data with a Significant Fraction of Less-Than Results*, SCA-TR-PR2010-0009*,* S. Cohen & Associates, Vienna, Virginia*.* November 2010.