COWORKER DOSE MODELING

Arjun Makhijani, SC&A
Harry Chmelynski, SC&A

Advisory Board on Radiation and Worker Health

Meeting of the SEC Issues Work Group
Cincinnati, OH
September 26, 2013

I. REVIEW

- ORAUT-RPRT-0053 discusses methods for developing coworker models. Statistical tests are recommended to decide if a single coworker model is appropriate for all workers at a site, or if separate coworker models are necessary for different sub-groups of workers (strata).
- Several ORAUT documents are based on the analytical methods proposed in ORAUT-RPRT-0053. These documents compare coworker models for construction trades workers (CTWs) and non-construction trades workers (NCWs) at the Savannah River Site (SRS).
  - ORAUT-RPRT-0055, *A Comparison of Exotic Trivalent Radionuclide Coworker Models at the Savannah River Site*, July 2012
  - ORAUT-RPRT-0056, *A Comparison of Neptunium Coworker Models at the Savannah River Site,* August 2012
  - ORAUT-RPRT-0058, *A Comparison of Mixed Fission and Activation Product Coworker Models at the Savannah River Site*, September 2012.

NIOSH submitted a *Response to SC&A Comments on ORAUT-RPRT-0053 in* August 2013 (ORAUT 2013).

FINDING 1: USE OF R^2 FOR GOODNESS OF FIT IN ROS

- 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.

NIOSH response (p 18): R^2 is not mentioned anywhere in the text of RPRT-0053 as a goodness of fit criteria. However, the R^2 statistic appears in some ROS plots. We think this was done at the request of someone in ORAUT, perhaps to be consistent with previous practice (i.e., PROC-95). ***R^2 was not used by the statisticians to evaluate fits in ROS plots*** *so we don't think this topic warrants a "finding" and does not need to be addressed in RPRT-0053.*

The applicability of the R^2 statistic in the evaluation of cumulative probability plots was previously raised by SC&A in their reviews of PROC-0095 and OTIB-0019. All findings related to this issue were resolved and closed in 2007 under the OTIB-0019 review. The closure language can be found in the Board's review system. [emphasis added]

SC&A Note**:**

What is used to determine goodness of fit of the lognormal distributions?

FINDING 2: REPRESENTATIVENESS AND COMPLETENESS

- The completeness and representativeness of the data available for coworker model are not addressed in ORAUT-RPRT-0053. If the unmonitored workers are from a different population, the applicability of a coworker model derived from monitored coworkers would be in question.

<u>NIOSH response (p 18)</u>: In the development of coworker models we assume that either unmonitored individuals are members of the monitored population who were not monitored completely at random, or unmonitored individuals were unmonitored because they had no potential for exposure to radioactive materials. In the first case we have the right model and in the second a conservative model. One can also theorize that these assumptions are wrong and that perhaps unmonitored workers were highly exposed and intentionally not monitored because of this. This fundamental and largely unstated difference in assumptions probably needs to be discussed and eventually resolved.

- Characteristics of monitored and unmonitored populations should be the same. **The relative exposure potential of the monitored versus unmonitored workers needs to be demonstrated rather than assumed.**

<u>NIOSH response (p 18)</u>: The reason why individuals were unmonitored will, in general, always be largely an assumption that cannot be "proved" to the satisfaction of everyone. The validity of this assumption is basically a function of the maturity of the radiation protection program in place at a given facility and the level of documentation available.

FINDING 2: REPRESENTATIVENESS AND COMPLETENESS (continued)

- The methods proposed in ORAUT-RPRT-0053 for analyzing the coworker datasets require verification that:
  (1) The available coworker data are representative of all groups of workers
  (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.
- **To this end, it is necessary to examine subgroups of CTWs. Data for the coworker model must be representative of these groups, and there must be sufficient data for pairwise comparisons with other monitored workers.**

<u>NIOSH response (p 5)</u>: An implicit assumption in any statistical analysis, including those in RPRT-0053, is that the data being analyzed are representative of the population in question (e.g., are "complete"). In our opinion, the issue of data completeness is not within the scope of RPRT-0053 and should not be identified as a "finding."

…in principle, no coworker model(s) can be "claimant favorable" to all strata in the model at the same time.

FINDING 3: VARIABILITY ESTIMATES USING OPOS

- The OPOS statistic methodology summarizes a worker’s exposure by averaging over all 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.

NIOSH response (p 6): In the presence of data dominance and dependent data (see Comment 9), the GM and GSD calculated with individual bioassay measurements do not have familiar statistical properties and are therefore not useful measures of central tendency and variance of the data. The OPOS statistic was adopted in an effort to deal with these major issues. We feel that the use of the OPOS statistic better achieves the goal of accurately estimating the intake rates and ultimately the dose to workers than does the use of individual bioassay results. Thus, it is not relevant whether or not the OPOS statistics have a higher or lower GSD than the individual data.

FINDING 4: “SAMPLING PROTOCOL” ISSUE

- Strata comparisons are valid only when the sampling protocols were the same.

<u>NIOSH response (p 7)</u>: In this comment the statistical term "sampling protocol" is incorrectly used as being synonymous with the term "internal dosimetry monitoring program." There is no statistical requirement that all workers be on the same monitoring program in order to use the data to develop a coworker model, as long as the monitoring programs adequately characterize all significant intakes.

- If internal dosimetry monitoring programs differ, valid statistical comparisons cannot be made.

<u>NIOSH response (p 7)</u>: ... we feel that it is appropriate (for example) to compare intakes calculated from "special" and "task-related" bioassay performed in one group to intakes calculated from "special", "task-related", and "confirmatory bioassay" in another group.

FINDING 4: “SAMPLING PROTOCOL” ISSUE (continued)

- Reason for SC&A concern: NIOSH stated in RPRT-0056 and RPRT-0058:

> *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.*

FINDING 5: LEVEL OF CONFIDENCE

- The test procedures recommended in RPRT-0053 require a high level of confidence before deciding that two worker groups are significantly different.
- This is not always claimant favorable for the most highly exposed groups of workers, since there is a trade-off between high confidence and power to detect differences given a fixed sample size. A high level of confidence reduces the power of the test to detect differences.
- Conducting the 2-sided test of the “No Difference” hypothesis at a 90% level of confidence would result in lower Type 2 error rates and be more claimant favorable to the more highly exposed group.

NIOSH response (p 10): If conducting tests at an $\alpha = 0.1$ significance level (90% confidence level) would be "claimant favorable" as claimed in this comment, one might conclude that conducting the tests at a 50% confidence level would be even more "claimant favorable." Where does it end? The answer to that question is that the significance level chosen for a null hypothesis test is ultimately a judgment based primarily on the conventions established in a particular scientific field. More specifically, a significance level of $\alpha = 0.05$ (95% confidence level) appears to be the standard significance level used in the most areas of science.

FINDING 6: SMALL SAMPLE SIZES

- **Power of the statistical tests to detect differences given the limited quantity of data with high proportion of nondetects has not been established.** The size of difference that can be detected reliably by the statistical tests was not examined. This deficiency should be corrected before RPRT-0053 is adopted as an appropriate procedure for evaluating coworker models.

> *NIOSH response (p 10): In this comment SC&A may be referring to a post-hoc power analysis, which is the determination of power after the data are collected and the test performed. A post-hoc power analysis is an attempt to extract something useful from a null hypothesis test where we fail to reject the null hypothesis. Unfortunately, this analysis provides no additional information beyond that given in the confidence intervals of the estimated parameters and its use is generally discouraged (see [Ellis 2010, pg 58] and [Hoenig 2001]).* [emphasis added].

SCA Notes:

(1) A confidence interval for the difference between the two distributions would provide as much information as a power analysis (see Figure 1). **But confidence intervals are not mentioned in RPRT-0053.**

(2) Not all statisticians agree on this topic. Bayesians tend to have more heretical views toward classical hypothesis testing, and favor estimation over hypothesis testing.

FINDING 6: SMALL SAMPLE SIZES (continued)

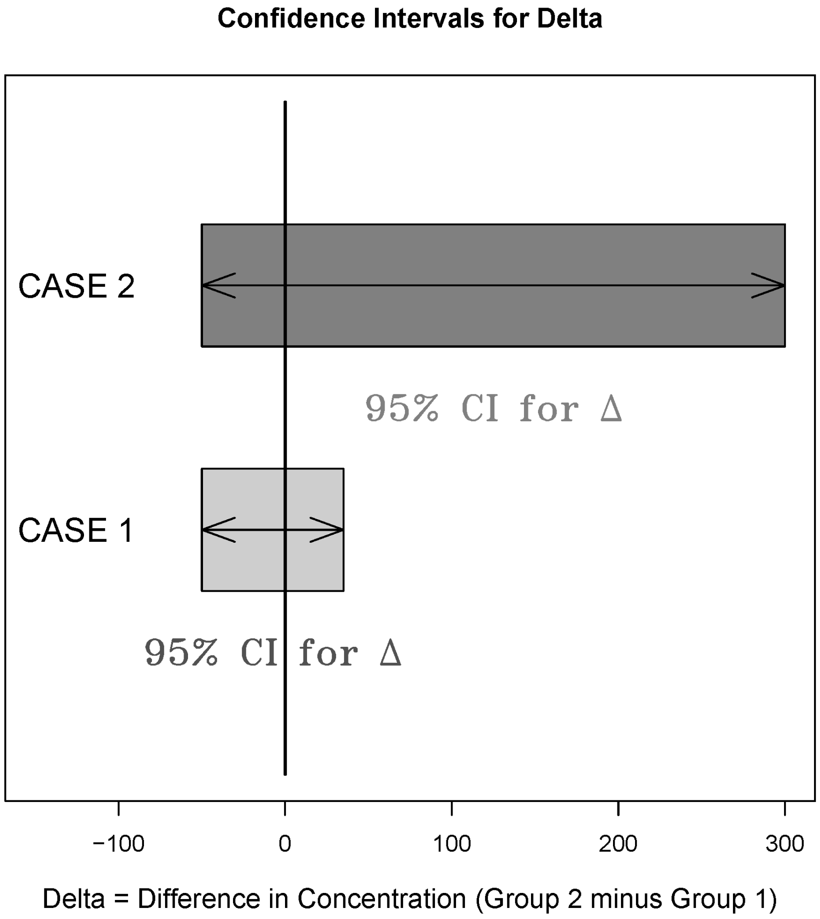

Figure 1. Confidence Intervals for Delta (Δ) with Large (Case 1) and Small (Case 2) Sample Sizes

FINDING 6: SMALL SAMPLE SIZES
(continued)

Gelman and Weakliem (2009): "There are cases in which it is difficult or impossible to obtain more data, and researchers must make do with what is available. We offer two practical recommendations for such situations:

First, whenever possible, ***researchers should determine plausible effect sizes based on previous research or theory, and carry out power calculations based on the observed test statistics.*** *…*

*Second****, researchers should exercise more ingenuity in efforts to obtain additional data from outside of their primary sample.*** *…"* [Emphasis added.]

FINDING 6: SMALL SAMPLE SIZES (continued)

- Another approach to confirm the test has adequate power is to determine if the sample size was adequate given the observed degree of variability. This approach was adopted in EPA 2006 as guidance for data quality assessment:

Box 3-32: Directions for the Wilcoxon Rank Sum Test

COMPUTATIONS: Rank the pooled data from smallest to largest assigning average rank to ties. Sum the ranks of the first population and denote this by R_1. Then compute

$$W_0 = R_1 - \frac{m(m+1)}{2}$$

STEP 1. Null Hypothesis: $H_0 : \mu_X - \mu_Y = 0$ (no difference between population means)

…

STEP 5. b) Conclusion: If p-value < significance level, then reject the null hypothesis

STEP 6. If the null hypothesis was not rejected, then the sample sizes necessary to achieve the DQOs should be computed. If the sample sizes are large, and only one false acceptance error rate (β at δ_1) has been specified, then the false acceptance error rate has probably been satisfied if both *m* and *n* are at least

$$1.16 \cdot \left[\frac{2 \cdot \mathrm{var}(W_0) \cdot \left(z_{1-\alpha'} + z_{1-\beta}\right)^2}{\delta_1^2} + \frac{z_{1-\alpha'}^2}{4} \right]$$

NOTE: The value of α′ is α for a one-sided test and α/2 for a two-sided test. The large sample normal approximation is adequate as long as *min(m, n)* > 10.

FINDING 6: SMALL SAMPLE SIZES (continued)

NIOSH response (p 31 & 33): Every comment in this section that refers to a priori power in the context of MARISSIM, ProUCL, and DQO implicitly assumes that one can ask questions and then design a sampling program that is capable of answering these questions. This is not possible for coworker studies so we feel that the information in these documents is not relevant. …

… The a priori power of a statistical test is considered during the design phase of the data collection procedure (e.g., the experiment or survey) so that the information collected is adequate to answer the questions being asked. The worker monitoring data used for developing coworker models were collected in the past to demonstrate compliance with the applicable occupational dose limits that were in place at the time. We are provided with these retrospective data and are asked to perform statistical analyses on the data to answer questions being asked in the EEOICPA program today. We did not have the opportunity to select the workers and monitoring programs needed to ensure that we could develop definitive answers to these contemporary questions. In summary, in coworker modeling we are presented with a predetermined dataset and cannot collect more data, so it is not useful to perform an a priori power calculation.

FINDING 6: SMALL SAMPLE SIZES (continued)

- In our review, we examined SRS Logbook Np OPOS data for 1961–1989 (Table 2). The WRS test shows slightly less power than the t-test for these datasets. We found there are sufficient data in 1961–1963 to detect differences as small as a factor of 2. The year 1985 produced anomalous results in this analysis. **In many years, the WRS test cannot reliably detect differences smaller than a factor of 4 to 10 in the CTW/NCW ratio of GMs.**

- Larger differences have a 95% or better chance of detection. Smaller differences cannot be detected reliably with the available data.

NIOSH response (p 8): The retrospective data used to develop coworker models "are what they are" and we have no opportunity to change them. Failure to reject the null with retrospective datasets is inherently neither "claimant favorable" nor "claimant favorable" [sic] and is not indicative of "bad" data or inappropriate statistical methods. ***The small CTW datasets mentioned in this comment argue for the use of an unstratified coworker model, perhaps used in conjunction with the 95th percentile intake rates if there is evidence that a particular construction trade worker had potential for exposure on a par with the higher exposed workgroups***. [Emphasis added.]

FINDING 7: WORKERS CHANGING JOBS

- 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 period, the OPOS values are correlated for this worker. (Note that OPOS aggregation periods can be as long as 3 years.)
- 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. If CTWs and NCWs are being compared, it is essential that the job designation has not changed during the period of OPOS aggregation.
- NIOSH has not investigated whether changes of workers from one stratum to another occurred during the period of OPOS aggregation and, if so, how such data are to be handled.

NIOSH response (p 9): First and foremost, we consider the technical benefits realized by using the OPOS statistic to far outweigh relatively rare problems like the one mentioned in this comment. Second, to stratify coworker models one has to be able to assign individuals to specific and meaningful job titles (i.e., develop a job exposure matrix). The difficulty in determining an individual's job title, as postulated by SC&A in this comment, is a general problem associated with assembling a job exposure matrix and really has little to do with the use of the OPOS statistic.

FINDING 8: POWER CONCERNS

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

<u>NIOSH response (p 10)</u>: The a priori power of a statistical test is usually considered during the design phase of the data collection procedure (e.g., the experiment or survey) so that the information collected is adequate to answer the questions being asked. In coworker modeling, we are presented with a predetermined dataset and cannot collect more, so it is not possible to perform an a priori power calculation…

To perform an a priori power analysis, an acceptable level of power 1 - β has to be defined. ***To define β we must first define the size of the effect[4] that we want to detect, i.e., the size of the effect that is of practical significance.*** *If we could define practical significance* ***(we tried and were unsuccessful)****, we would perform an equivalence test [Streiner 2003], which tells us if the difference in the two groups is of <u>practical</u> significance, rather than a null-hypothesis test, which tells us if the difference in the two groups is of <u>statistical</u> significance.* [Emphasis added.]

[4] The magnitude of the difference between the two groups.

FINDING 8: POWER CONCERNS (continued)

- Accepting the null hypothesis could be claimant unfavorable to the more highly exposed group if the available data do not provide adequate power for the test.

NIOSH response (p 10): Tthe Peto-Prentice is the most powerful test available that can be used for comparing two groups with left-censored lognormal data, and while the power of this test will vary depending on the actual data used, there is no better statistical test that can be used for this purpose.

- NIOSH has stated that 30 samples in each strata is sufficient for a valid comparison.
- Figure 2 shows results of simulations performed to show the power of the WRS test when using 30 samples to compare lognormal distributions which differ by a factor of 2.73. The samples include 24% nondetects. The GSDs are assumed the same for the two distributions.
- Results: **Type 2 error rates can be very high (15%–35%) when using the 95% confidence level (α=0.05) if the GSDs exceed 4.** If the confidence level is reduced to 90% (α=0.10), the Type 2 error rate is maintained below 20% up to a GSD of 6. If the confidence level is 80% (α=0.20), the Type 2 error rate maintained is below 10%.

FINDING 8: POWER CONCERNS (continued)

- **Overall, SC&A concludes that the NIOSH method of determining that there are no significant differences based on the available data would often lead to very claimant-unfavorable results to the most highly exposed strata**.

NIOSH response (Appendix A): Simulation results reported in Figures 1 to 12 of Appendix A.

SC&A Note: All examples in Appendix A have GSD no higher than 3. Our concerns were for GSDs of 4 and above. (See Table 1 and Figure 2)

Table 1. Type 2 Error Rate of WRS Test using 30 Samples from Two Lognormal Distributions: LN(0,1) and LN(1,1)

	GSD				
α	6	5	4	3	2
0.05	0.35	0.27	0.16	0.04	<0.001
0.10	0.22	0.16	0.09	0.02	<0.001
0.20	0.11	0.07	0.04	0.006	<0.001
0.25	0.09	0.05	0.02	0.003	<0.001

n1=n2=30, 24% nondetects, and GM2/GM1 = 2.73

Shaded region of table has Type 2 Error rate ≤10%

FINDING 8: POWER CONCERNS (continued)

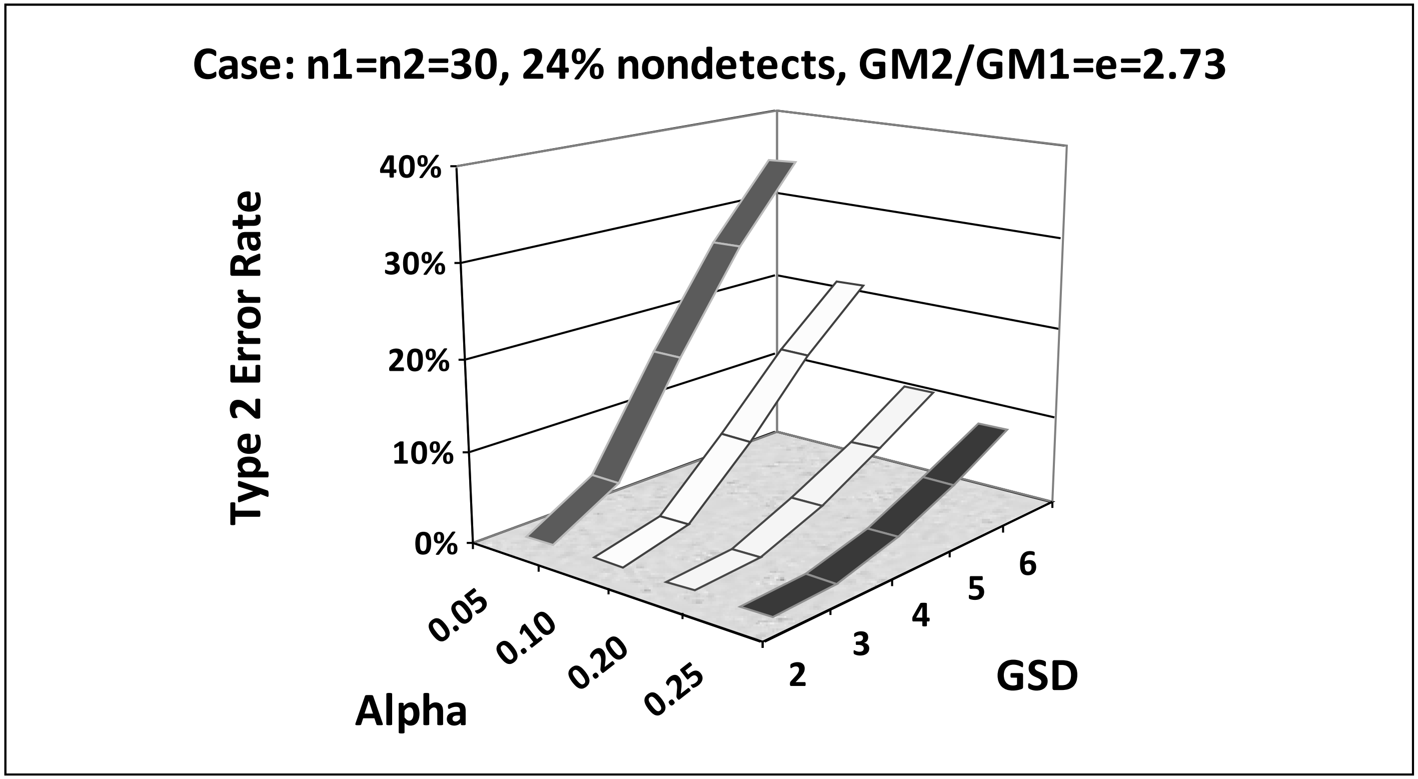

Figure 2. Type 2 Error Rate of WRS Test using 30 Samples from Two Lognormal Distributions: LN(0,1) and LN(1,1)

RECOMMENDATION 1: 1-SIDED VERSUS 2-SIDED TESTS

- RPRT-0053 recommends using 2-sided tests to determine if there is a significant difference between groups of workers. The null hypothesis for these tests states there is “No Difference” between the two groups. **This form of test is not claimant favorable to highly exposed workers at SRS, as it places the burden of proof on the CTW claimants to prove that a significant difference exists.**

<u>*NIOSH response (p 11-12)*</u>*: RPRT-0053 was designed to test for non-directional differences. So, for example, a stratified coworker model would be considered when the dose to Group A is significantly different than the dose to Group B, regardless of which is larger. This led us to use "two-sided" hypotheses tests in RPRT-0053…*

We feel that the decision to stratify a coworker model should be based on significant differences, not just the specific differences that are of interest to a given concern. This difference in philosophy is noted here because we feel that it is the basis for a number of the comments offered by SC&A. The ultimate resolution of this issue may fall under the category of a policy decision on when, why, and how coworker models should be stratified.

- In the specific case of the SRS SEC for CTWs, a 1-sided test is more appropriate, as it addresses directly the question at hand: **Are the CTW samples higher than the NCWs, or not?**

NIOSH response (p 46): In theory, a better approach to achieving what SC&A appears to be after here would be to define a difference d between the CTW and non-CTW empirical cumulative distributions that is considered to be of practical significance and rearrange the hypotheses to form an equivalence test [Streiner 2003, Wellek 2010]:

H_0: { $F_{ctw}(x) - F_{nctw}(x)$ } $\geq d$ *for all x*
H_A: { $F_{ctw}(x) - F_{nctw}(x)$ } $< d$ *for at least one x*

Here, H_0 is that the distributions are not equivalent and H_A is that they are. Thus, the data can prove that the CTW and non-CTW are equivalent, i.e., that there is no practically significant difference between the two, by rejecting the null hypothesis. If the null is retained we can proceed under the assumption that the CTW distribution is significantly different than the non-CTW distribution.

The difficulty in implementing an equivalence test is that we have to define doses or bioassay results that are of practical significance to the compensation decision before looking at the data. During the development of RPRT-0053 we tried to define d and were unsuccessful, which is why we used the null hypothesis test of statistical significance rather than the equivalence test of practical significance.

RECOMMENDATION 1: 1-SIDED VERSUS 2-SIDED TESTS (continued)

- NIOSH might consider using a 1-sided hypothesis test instead of the 2-sided test now used. The non-parametric Peto-Prentice test is more generally applicable than the parametric MCPT, and may be applied using a more claimant-favorable 1-sided null hypothesis stating that groups with high exposure potential, such as CTW at SRS, are more exposed. This is more likely to result in a claimant-favorable coworker model for highly exposed groups of workers.

NIOSH response (p 46): Adopting this approach means that we could go looking for a difference in strata, fail to detect that difference, and then develop a model that incorporates the difference anyway. We see fundamental difficulties associated with this approach. If the data are not adequate to demonstrate a significant difference between the strata then it is not clear to us how incorporating this difference into the coworker model will improve the estimates of the intake rates. This is critically important when there are in fact no significant differences between the strata. In this case stratification (the default action taken when we fail to reject the null) will always result in poorer estimates of the intake rates because it unnecessarily reduces the size of the sample used to estimate the intake rate.

RECOMMENDATIONS 2 & 3: NUMBER OF STRATA AND NUMBER OF SAMPLES FOR SUBGROUP COMPARISONS

- **NIOSH has not made comparisons of CTW subgroups.**

- Analysis of SRS CTWs by job type and by area of work (SC&A 2010a, 2010b) indicates that subgroups of CTWs have unique distributions of exposure and are not from the same distribution as NCWs or other CTW subgroups.

- Multiple pair-wise comparisons would be required for the CTW analysis. Sufficient data (**at least 30 samples for each category**) would be required in each job/area category for which a coworker model is to be constructed. **The hurdle is sufficient data for such comparisons.**

NIOSH response (p 47): Small strata that represent small samples of some larger group can have large uncertainties in the estimated parameters. On the other hand, if a small stratum is basically a census of all workers who should have been monitored and all of the data are uncensored, the situation may not be as bad. In the end, we recommended a minimum of 30 OPOS statistics, i.e. data from 30 individuals, in each stratum. However, because the procedures in RPRT-0053 were going to be used only by degreed statisticians, we gave them the latitude to exercise professional judgment on this subject.

REFERENCES

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

Gelman, A. and Weakliem, D. 2009. “Of Beauty, Sex and Power,” *American Scientist*, July–August 2009.

ORAUT 2013. *Response to SC&A Comments on ORAUT-RPRT-0053*, August 2013.

ORAUT-RPRT-0055. *A Comparison of Exotic Trivalent Radionuclide Coworker Models at the Savannah River Site*, July 2012.

ORAUT-RPRT-0056. *A Comparison of Neptunium Coworker Models at the Savannah River Site*, August 2012.

ORAUT-RPRT-0058. *A Comparison of Mixed Fission and Activation Product Coworker Models at the Savannah River Site*, September 2012.

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

Streiner 2003. ‘Unicorns Do Exist: A Tutorial on “Proving” the Null Hypothesis’ *Canadian Journal of Psychiatry*, (48) No 11.

Wellek 2010. *Testing Statistical Hypotheses of Equivalence and Noninferiorty*, Boca Raton: CRC Press.