65FR 42735
July 11th - 2000
(1)

UNIVERSITY OF ILLINOIS
AT CHICAGO

Biometric Laboratory (MC 912)
1601 West Taylor Street, Fourth Floor West
Chicago, Illinois 60612

Robert D. Gibbons, PhD
Professor of Biostatistics

September 2, 2000

U.S. Nuclear Regulatory Commission
Office of Nuclear material Safety and Safeguards Washington DC
20555-0001

Gentleman,

I have read with interest your draft report NRC NUREG 1724, entitled *Standard Review Plan for the Review of DOE Plans for Achieving Regulatory Compliance at Sites with Contaminated Ground Water Under Title I of the Uranium Mill Tailings Radiation Control Act.* Since one of the principal references upon which the document appears to be based is my book *Statistical Methods for Ground Water Monitoring*, (Gibbons, 1994, John Wiley and Sons), I felt obliged to provide you with a brief critique of your guidance. Upon reading the draft report, I was disappointed to find that it is riddled with both conceptual and statistical errors and omissions, some of which are quite serious. I strongly encourage you to seek help from a professional statistician working in this area. Furthermore, the report is largely based on old USEPA Guidance in this area which has been completely replaced both by ASTM Standard D6312 and the new USEPA Unified Statistical Guidance document, which is soon to be released. Neither document is even referenced in your draft report which represents a serious omission. In the following, I will highlight a series of issues raised in just a brief review of the draft report.

1. page 1 - The null hypothesis of "no difference between background and on-site water quality" should be changed to the null hypothesis of "on-site water quality is less than or equal to background" so that the alternative hypothesis of "above background" makes sense.
2. page 2 - The recommendation of 24 measurements based on limiting Type II errors is only one piece of the puzzle. Statistical power (*i.e.*, $1 - \beta$) is a crucial consideration, but the discussion ignores effect size. In the Unified Guidance, power of .5 at a 3 sd unit increase over background and power of .8 at a 4 sd unit increase over background are required, irrespective of the method or number of samples. This is a far more sensible criterion than simply specifying a fixed number of measurements.
3. page 2 - The use of the term sample to refer to a collection of $n = 24$ measurements for a given constituent is poor nomenclature. Further-

E-RIDS-ADM-03
Add: W. FORD (WHF)
Template = ADM-013

more, it is unclear if this pertains to characterizing background or a compliance well.

4. page 2 - The minimal sampling interval of two weeks is highly dependent on existing ground-water flow conditions. In general, taking ground-water samples less than quarterly is ill-advised.

5. page 3 - In almost all cases, prediction limits are preferable to tolerance limits because tolerance limits have a built in failure rate. For example, a 95% coverage 95% confidence tolerance limit has a 5% failure rate with 95% confidence. This is typically unacceptable in a regulatory environment. By contrast, a prediction limit has 95% confidence of covering 100% of the next k measurements and is therefore far better suited to ground-water monitoring problems.

6. page 3 - Use of tolerance limits for comparisons to standards should not be used. The authors are confused by USEPA's 1992 reference to the use of a lower confidence limit (LCL) for an upper percentile of a distribution, which is the same as a lower tolerance limit. This is not what is referred to in the draft report. This is now completely clarified in the Unified Guidance.

7. page 3 - The inherent assumption of parametric statistical methods is not normality. For example, Poisson prediction limits are parametric and have nothing to do with normality.

8. page 4 - You are always better using a statistical adjustment such as Aitchison's method or Cohen's method than imputing a value such as PQL or PQL/2. The point is that if the percentage of nondetects is less than 15%, it really doesn't matter what you do. Don't preclude use of more sophisticated methods when the detection frequency is 85% or more.

9. page 4 - The Poisson prediction limit is not a nonparametric method! Here you should introduce the idea of nonparametric prediction limits which is the best approach when the detection frequency is low.

10. page 4 - The discussion of the Kruskal-Wallis and Wilcoxon nonparametric tests is incorrect. They both assume homogeneity of variance. Furthermore, it is now very well known that ANOVA is a very poor method for analysis of ground-water monitoring data because it is quite sensitive to small consistent differences such as spatial variability and insensitive to highly variable data commonly observed in a contaminated well (see Gibbons, 1994 for a list of other reasons to avoid ANOVA). The Unified Guidance no longer recommends ANOVA.

11. page 4 - What is a "random interval"

12. page 5 - Tolerance limits as defined in the draft report should *never* be used for comparisons to standards.

13. page 5 - Equation 4.1 is completely incorrect. The tolerance limit is not based on Student's t-statistic. This a glaring error clearly highlighting the need for professional statistical consultation.

14. page 5 - Simple substitution of $s_{\bar{x}}$ into a tolerance limit has no statistical justification and is clearly incorrect. Here, use of a prediction limit for a future mean value should be used (see Gibbons, 1994 and the forthcoming book *Statistical Methods for Detection and Quantification of Environmental Contamination* (Gibbons and Coleman, 2000, Wiley).

15. page 5 - You cannot use the term prediction interval to refer to a one-sided upper prediction limit. They have different confidence levels.

16. page 6 - There are perhaps 30 different types of control charts. It is totally unclear what is being referred to here. The most useful one for intra-well ground-water monitoring applications is the combined Shewhart-CUSUM control chart (see Gibbons R.D. Use of combined Shewhart-CUSUM control charts for ground-water monitoring applications. *Ground Water*, **37**, 682-691, 1999).

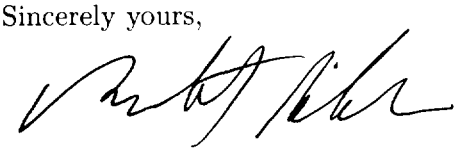
17. page 6 - The discussion of "Strategies for Multiple Comparisons" is extremely poor. First, the problem results from a comparison of multiple compliance wells and constituents to a *common* background. Second, the solution is to incorporate verification resampling as an integral part of the statistical test. There are numerous articles and books by myself and Charles Davis on this topic. Remarkably, verification resampling strategies are not even mentioned. None of this works without verification resampling.

18. There is nothing in the draft report that discusses experimental design issues of ground-water monitoring well networks. At the very least, there must be a minimum of two upgradient or background wells, otherwise potential contamination and spatial variability are completely confounded.

Unfortunately, I found the draft report to be of little value, statistically flawed and potentially misleading at best. I strongly encourage you to seek professional statistical help in establishing guidelines for this important problem of

protecting human health and the environment.

Sincerely yours,



Robert D. Gibbons
Professor of Biostatistics