

PRELIMINARY COMMENTS BY DR. RICHARD SMITH

First of all, my apologies to the Chair and committee members that I am late delivering these comments and that my review has not been as thorough as I would have hoped. At the time I originally agreed to participate in this panel, I assumed that the Christmas period and first half of January would be a relatively light time for me, but in fact things have turned out rather differently. I still hope I'm able to make an effective contribution to the committee.

A bit about my own background. I am a professor of statistics and biostatistics at UNC and have long featured air pollution as one of my research interests. I was appointed to EPA's Science Advisory Board (SAB) in 2017 by Administrator Scott Pruitt and reappointed by Administrator Andrew Mitchell in 2020. Prior to my appointment to the SAB, I had virtually no contact with EPA on any topic other than air pollution, and even while on that board, I have largely focused on topics related to air pollution, since on most other topics, there are other members of the SAB with more expertise than I have. So the question of how to control nuclear radiation is largely a new topic for me, though I have encountered it in a couple of consulting activities, which sparked my interest in finding out more of how EPA actually approaches these issues.

Regarding the document under review, I have already stated that I didn't spend as much time as I could have liked, but I did spend several hours trying to understand the sections that seemed most relevant to the charge questions, and I found them heavy going. Even if I had spent another 10 or 20 hours trying to read it, I am not sure I would have been better informed what the document is trying to do. There is barely a page of the document where you don't have to cross-reference something else to find out what they are talking about – couldn't it have been written in much more of a linear order? When reviewing papers for journals, I usually follow the principle that if I can't understand what the paper is about without referring to the appendices, the paper needs to be rewritten, and if this were a journal review, I am sure I would say something like that.

However beyond purely writing aspects, nowhere does the document specify exactly what it is trying to achieve. I would expect a discussion of these rules to be informed by (a) some broad agency-defined specification of objectives, e.g. the probability that someone gets a certain type of cancer as a result of nuclear contamination should be no more than one in a million or some similarly low number, (b) the model by which that criterion is turned into an action level (e.g. logistic regression would be a simple example of a statistical model relating the probability of a health outcome to a level of contaminant). (a) should logically not be part of this committee's remit to review. As for (b), I think it is appropriate for this committee to review the models that EPA is using, but we are not being asked to develop a rule from first principles, but rather, to review a set of rules that EPA has been developing for many years. So I think it is fair game to ask, what are the models that EPA is using in proposing these rules?

Charge Question 1.4. The question asks whether the proposed requirements for “areas of elevated activity” are accurate and appropriate, and specifically whether the “unity rule” is appropriate for areas of elevated activity. I don't find a clear definition of what is meant by

“areas of elevated activity” so it is difficult for me to answer this question in the abstract. Maybe the issues will become clearer to me after we have had an initial discussion in committee.

Regarding the “unity rule,” we are being asked whether this should apply to comparisons of elevated measurements (EMCs) but I see no discussion of the justification for the unity rule itself. To my eyes, the unity rule may seem like an intuitively reasonable thing to do, but I ask again, exactly what is the objective and how does the unity rule achieve that? Although I have not tried to derive this rule for myself, I can see how one might end up with such a rule based on, for example, an additive logistic regression rule for the probability of a particular adverse health outcome as a function of multiple pollutants. But if that was the justification, why would I do something different for EMCs? It seems to me that in asking whether the same rule should apply to areas of high contamination, that is exactly the situation where one would want to be rigorous in applying the rule!

However, I actually think this question is getting at something else. There is an obvious concern that if one is applying a sequence of tests to a geographically small area (e.g. a single house), there could be come unnecessary duplication on the testing. This seems to be more a question about the spatial distributions of the contaminants than the health effects per se. Possibly one could justify some weakening of the unity rule if it were supported by evidence that contamination levels were likely to be highly correlated at nearby locations, *even in the tails of the distributions* (an important caveat). There are some statistical tests that have been focused on that type of comparison, and I could give references if our discussion goes in that direction.

Charge Question 3.1. This question is impossible to answer without a clear definition of the rule we are being asked to evaluate. We are being asked to comment on “the revised description of how to set the Lower Bound of the Grey Region (LBGR)” but as a citation of where we can find this description, we are referred to “Chapter 4 and Section 5.3” which together occupy 105 pages of the document! If we cannot start with a clear, unambiguous definition of the rule, how on earth is any outside user supposed to make sense of it?

To my eyes as a statistician, the definition of the LBGR and corresponding UBGR all come down to the power and size of the hypothesis test, regardless of Scenario A or Scenario B, which are well-known concepts in statistics and obviously are familiar to the authors of this document. From that perspective, I entirely agree that a crude rule of thumb, like setting the LBGR to be half of the DGCLw, is unlikely to be adequate for the job.

As a general comment about the notion of power in statistics (this is not just confined to studies of the nature being considered here – I have often been asked to review biostatistics research proposals, which usually include some requirement to evaluate the power of the proposed test) I think trying to define a specific level of the alternative hypothesis is the hardest part of a power calculation. Once the null and alternative hypotheses are clearly defined, the rest is either a mathematical calculation or (more likely with complicated tests) a simulation. So, echoing my initial remarks in this review, this is another place where I would like to see a clearer statement of what the objectives are in formulating the rule. Under Scenario A, where the null hypothesis is that the level of contamination is at or above the DGCL, the obvious choice for an alternative

hypothesis is that the level of contamination is at or below a level that toxicologists and nuclear physicists agree to be safe. Deciding that level is not primarily about statistics, but deciding you whether you have achieved the target is a statistical question.

So I would say, the LBGR is something determined independently of the UBGR, and should be informed by what the relevant experts agree to be safe. Power calculations logically follow from this, on such questions as how large a sample is needed, but they shouldn't determine the LBGR itself.

Charge Question 2.2. This question is essentially asking me to review Appendix E, on ranked set sampling (RSS). This material is much better written than the other sections I have been asked to review – in general I find it a comprehensible and accurate summary of the methodology and will be useful to readers with an appropriate statistical background, which should include some concepts of different sampling designs and order statistics as well as basic probability and the Wilcoxon-Mann-Whitney (WMW) and sign tests (though this appendix focuses mostly on the sign test in conjunction with RSS).

It seems to me the major question about RSS is “When should you use it?” Perhaps we should have some side discussion about that. RSS designs are more efficient than simple random sampling (SRS) designs, but they are also more expensive to implement, so there is a trade-off.

My queries are not about the basic principles of the approach, but much more the sorts of things that I would routinely expect to bring up in a review for a journal or book chapter:

1. Page E-2: reference to software produced by the Pacific Northwest National Laboratory (PNNL). I think it would be helpful to clarify exactly what this software is and for what purpose it is being recommended (I don't see any later reference to PNNL – was this just “mentioned in passing” or does EPA intend that use of this software would be a major part of the recommendation?).
2. Page E-5 – I thought this was a nice example that illustrates the concept very clearly.
3. Page E-6 – reference to a book by Chen. I wasn't aware of this book before but I looked it up and it seems the authors are referring to *Ranked Set Sampling: Theory and Applications* by Zehua Chen, Zhidong Bai and Bimal Sinha, published by Springer in 2004. I recommend giving the full citation. Also I note that in the subsequent writing, the authors focus on the sign test as the main test they are recommending, while referring to the Chen book for the alternative WMW procedure. This is okay so long as they are not expecting the WMW method to be widely used – otherwise it might be a good idea to include an explicit section about that as well.
4. Section E.2.2 and Tables E1-E3: I would like to see a clear statement of assumptions and also an attribution for these tables – did you calculate them yourselves or were they taken from some other source? If the latter, please cite.

Regarding the assumptions behind Tables E1-E3, I think most likely two things are missing that would be good to spell out. Do the calculations rely implicitly on the assumption of a normal distribution? If so, it might be worth saying so. A practical point here is that if the fit to a normal distribution is improved by making some transformation (e.g. power law or logarithmic), it might be beneficial to assess Δ and σ on those scales as well – in other words, even though normal distributions are not required for the sign test, if the assessment of power does assume normal distributions, this might affect the way the test is applied in practice, so it would be worth including some discussion of that. My second point is: from the way this is described and the formulas on page E-10, I think the authors are assuming the ranking is perfect, in other words, that it does correctly identify which of the elements is smallest, middle and largest among the ranked samples. In practice the ranking won't be perfect and I think there should be some acknowledgement of that fact.

5. Section E.2.3: unless I am misunderstanding something, this seems to me pointing in a different direction from Section E.2.2. Section E.2.2 is all about deriving power from theoretical calculations, though as pointed out in the previous comment, I think this does rely on some assumptions that need to be made clear. However, Section E.2.3 seems to be recommending that the RSS sample be analyzed as if it were a simple random sample (SRS), which will, in particular, overestimate the population variance. It is argued that this is a good thing because the test will be more conservative, but is this not somewhat at variance with the concept of trying to do an exact power calculation? I would welcome some further discussion of the rationale behind this recommendation.
6. I think it might be helpful if during the committee session, we looked more closely at Tables E.4 and E.5 and assessed how much they are really different in practice. It seems to me that Table E.4 leads to smaller critical values but only just in many cases. This might be of practical importance in deciding when to use RSS, which I presume does involve some increased costs and therefore should not automatically be recommended.
7. Section E.3 and Example 5: this is a very clear worked example and nicely rounds off the section. One question I have which relates to the whole spatial dependence question: it seems to me that whether one employs SRS as on page E-20, or the more complicated RSS on page E-22, there is still an implicit assumption that the 15 or 45 observations being taken are independent. Is this correct and would that in any way affect the sampling recommendation?