Comments on the paper by Pitt and Hill.

I do not regard this paper as a good discussion of the issues associated with the use of statistics in the detection of potentially fabricated data. It reads more like an attempt to publish their analysis of the data of RTS, thereby getting TAS support for their case of misconduct against this person. A broad-ranging, critical discussion would read very differently. For example, the details of the specific tests the authors devised are not really of interest.

Why don't I think it is a good discussion? For three principal reasons: it is one-sided, it is uncritical, and the authors seem to believe that rejection of a null hypothesis implies the acceptance of the alternative they offer. More fully, the authors seem either unaware or uninterested in the possibility that statistics might be misused in this context. Further, they exhibit a faith in the applicability of statistical models to real world data generated by humans that in my view is quite unjustified. And finally, as I saw it, they made no effort to seek explanations other than fabrication for the phenomena they identified and studied.

Lest they think I am speaking hypothetically, I invite them to read either the full transcript leading to the decision 1996.06.21 DAB1582 Thereza Imanishi-Kari, Ph.D., the decision itself (available on the web), or the statistical parts of the book *The Baltimore Case* , by D J Kevles, Norton 2000. I am not sure where the entire transcript can be found, but Kevles obviously found it. Most of the statistical analysis put forward by the ORI (contrary to their statement on line 8 of page 3) was carried by JR Mosimann, whose work they cite. I critiqued it. Mosimann seemed to proceed exactly as described in the authors' paper, and exactly as they themselves proceed in their analysis.

The first step in using statistical techniques to identify fabricated data is to look for anomalous patterns of data values in a given data set (or among statistical summaries presented for separate data sets), patterns that are inconsistent with those that might ordinarily appear in genuine empirical data. That such patterns are, indeed, anomalous may potentially be confirmed by using genuine data sets as controls, and by using simulations or probabilistic calculations based on appropriate models for the data to show that they would not ordinarily occur in genuine data. (ms, p.3)

They go on:

The existence of these anomalous patterns in given suspect data sets may be indicative of serious issues of data integrity including data fabrication (Al-Marzouki, Evans et al. 2005), but they may also arise as a result of chance. (ms p.3)

In other words, it is either chance, as assessed the way the investigative statistician chooses to assess chance, or "serious issues of data integrity including data fabrication." What can go wrong here? Well, in my view a lot went wrong in the case

I mentioned, and I invite the authors of this paper to reflect on this matter. They may conclude that my critique of Mosimann's analysis led to a miscarriage of justice, that he was right and I was wrong. In that case, they need another reviewer's opinion, as my comments will then be disregarded. On the other hand, they might agree with me that Mosimann went too far, that he misused statistics in his efforts to demonstrate misconduct, in which case they will want to revise their paper.

I will now present a few quotes from the paper that highlight my assertion that the authors (like Mosimann in the above case), put too much faith in statistical models.

"Ideally one would like to have a probability model for the underlying randomness in the experimental data and use it to show that the distribution of terminal digits of counts values in data sets consistent with that model will be uniform."

"Finally, one could try to validate the assumption that terminal digits of counts in legitimate data sets are uniform, empirically, by testing the uniformity of terminal digits in indisputably legitimate experimental data sets of exactly the same type, constructed using the same protocols, as that of the suspect data."

"None of these patterns were evident in any of the data sets reported by the nine other investigators in the same laboratory, or in data sets that we obtained from three other independent, outside researchers."

"Random variation in these triplicate data that are common components of pharmacological, cell biological and radiobiological experimentation, can be analyzed by modeling the triples as sets of three independent, identical Poisson random variables."

"Having observed what appeared to us to be an unusual frequency of triples in RTS's data containing a value close to their mean, we used R to calculate the ratios for all of the colony data triples that were available to us. We then constructed histograms of the resulting data sets. "

Let me ask two questions. Did the authors consider taking large bodies of data from each of the other investigators, conscientiously searching for anomalous patterns therein, finding something in each case, devising custom test statistics that highlighted the patterns found, and finally calculating p-values for these statistics for all investigators? Did they take any steps to see whether the empirical p-values they calculated from their control data using their test statistics actually followed a uniform distribution (as it should if all these hypotheses were null). In my view, this is the least they should have done.

I close with the remark that RTS may well have fabricated his or her data. At issue here is the way in which statistical methods are used to demonstrate, support or suggest that fact. In this respect, the present paper is not fundamentally different in outlook and approach from that of JE Mosimann in the case above. If they or the editors concede the possibility that my critique of Mosimann had some validity, then more caution is required. I should not have to repeat those arguments here, as it is all on the record.

Terence P. Speed, March 17, 2014