From: bengt.glimelius@onkologi.uu.se

To: hill@umdnj.edu, drjhpitt@yahoo.com

CC:

Subject: Decision on manuscript ID SONC-2013-0125

Body: 06-Mar-2013

Dear Dr. Hill:

I write you in regards to manuscript # SONC-2013-0125 entitled "FORENSIC STATISTICS USED TO ANALYZE PRECLINICAL SURVIVAL DATA IN NUCLEAR MEDICINE" which you submitted to the *Acta Oncologica*.

Your manuscript did not obtained the necessary support for acceptance in competition with other good manuscripts.
It has been evaluated within the Editorial Board and sent it to two external reviewers who immediately responded that they did not consider *Acta Oncologica* as an appropriate journal. The evaluation is thus that it is not of sufficient priority for publication. We presently receive far many more manuscripts than can be accepted and published.

Thank you for considering the *Acta Oncologica* for the publication of your research. I hope the outcome of this specific submission will not discourage you from the submission of future manuscripts.

Sincerely,
Editor in Chief, *Acta Oncologica*

Date Sent: 06-Mar-2013

3/9/2013 9:49 AM

Dr Bengt Glimelius

Editor in Chief, *Acta Oncologica*

bengt.glimelius@onkologi.uu.se

Re: Decision on manuscript ID SONC-2013-0125

Dear Dr Glimelius,

We are very disappointed that you have declined to publish our paper # SONC-2013-0125 entitled "FORENSIC STATISTICS USED TO ANALYZE PRECLINICAL SURVIVAL DATA

IN NUCLEAR MEDICINE" which we submitted to the *Acta Oncologica*. In our paper, we analyze numbers that were used to provide data in 8 publications. We show that the numbers supplied to the article in your journal (Bishayee, A; Rao, DV; Howell, RW Radiation protection by cysteamine against the lethal effects of intra-cellularly localized Auger electron, alpha- and beta-particle emitting radionuclides. ACTA ONCOLOGICA 39: 713-720, 2000) are extremely likely to have been manipulated and thus, that publication along with the 7 others, is unreliable and should not be depended upon by other researchers in subsequent related studies. It is especially distressing that you feel that our paper is not appropriate for your journal since this apparently was not the case when your journal decided to publish the original paper by Bishayee, *et al.*

Lest the results recorded be relied upon by others in the field, we believe that it is imperative that the questioned paper be retracted. I believe that it is your responsibility to see that this occurs. I point out that results using the following radionuclides: ^{210}Polonium citrate, $^{3}H_2O$, ^{125}IdU and ^{131}IdU were documented in the paper in question. We noted that the average of triplicate colony counts appeared as one of the 3 counts very frequently in the results of those experiments with the 4 isotopes recorded in the paper. The probabilities for the average occurrences were all less than 10^{-15} (i.e. too small to measure) except for ^{210}Polonium for which the probability is 3.2×10^{-8}. We had many triplicate colony counts from other investigators in the laboratory to use as controls. For them, the averages were not frequent and the probability for the average to be present in the triple was greater than 0.9, i.e. those occurrences were very likely due to chance.

I implore you to reconsider. Scientific misconduct is being recognized as an ever increasing, wasteful, misleading and vexing problem. It is important to recognize it and repudiate it whenever we can. It is distressing that you and your colleagues at *Acta Oncologica* do not appear to agree.

I respectfully request that you supply us with the critiques of your Editorial Board members and the two referees that you mention.

Sincerely yours,

Helene Z Hill, PhD

11:18 AM

Dear Helen and Joel,
Of course I will reconsider the decision made and make a much more thorough evaluation. Sorry that I could not open your accompanying letter - the computers the county council in Uppsala have are not very updated. Likely I would then immediately have understood more than what I did from the title and the abstract. I asked two scientists in nuclear medicine who both mailed me telling that this was complicated statistics and not really within the scope of this journal. None of them said anything about your message of potential scientific fraud. I remember I checked the abstract again, reacted that it was "unusual" but full of statistics and very "polite". However, I made a too rapid decision, prompted by the many manuscripts we receive so the rejection rate is over 80%, which I of course regret.

Do you have plans to approach also the other journals in which these authors have published there results? The article published in 2000 was from a scientific meeting about Auger processes and there is always a tendency that symposia articles are less carefully evaluated than spontaneously submitted articles.
I will come back after a much more careful evaluation.
Kind regards
Bengt

3/9/2013 4:40 PM

My email reply:

Thank you so much for considering our work more carefully. There is no mention on my print copy from the journal of the 2000 *Acta Oncologica* paper that it was from a scientific meeting and, in fact, it is labeled as an *Original Article* in the heading. We are acutely aware that our paper is heavy with statistics and we tried to reduce that by putting the bulk of Dr Pitt's model in the appendix -- were you able to view that Appendix in the *Scholar One* submission? We have tried to be "polite". Statisticians (e.g. Dr Pitt) tend to be very precise and careful not to overstate their case.

I would like to call your attention to a series of papers and editorials that appeared in the journal *Anaesthesia* last year. Forgive me if you are already aware of them. The key paper with the heading of "Special Article" by J.B. Carlisle is also heavy with statistics - analysis of numerical data produced by Y. Fujii, an anaesthesia researcher in Japan. The journal's Editor-in-Chief, S.M. Yentis, formerly on the COPE Council, wrote an introductory editorial and J.J. Pandit followed with a rather detailed editorial explaining the analysis for readers who were not that familiar with statistics. Fujii was given the opportunity to rebut Carlisle's conclusions and other commentaries followed. To me, such an exchange of articles, editorials, critiques and rebuttals is science at its very best. Fujii, as of January 15, 2013, had retracted 183 articles according to the Retraction Watch and a follow up article by Carlisle showed that once the Fujii results were removed from the studies, the medication he was advocating had no effect.

I hope you will not only accept our paper, but will encourage similar exchange. The authors of the 2000 *Acta Oncologica* paper should have the right to respond. We did approach without success *Radiation Research* which published 4 of the questioned articles. That journal is not on COPE's list as adhering to its principles and so far has not taken any action.

We look forward to a favorable response this time.

Sincerely yours,

Helene Hill

4/17/13 4:36 PM

Dear Dr Glimelius,

It is now more than a month since the following exchange occurred. I hope it would not be too presumptuous to ask for an update. We are very anxious to move forward with our analysis. Is there any more information that you need from us? The raw data that we analyzed are all available on my website: helenezhill.com.

We are hopeful for a favorable decision to publish our article in *Acta Oncologica*.

Sincerely yours,

Helene Z Hill, PhD
Professor of Radiology
NJ Medical School

4/17/13 4:43 PM

Dear Helene,
I looked at your site yesterday and had only received 1 out of the 5 invited referees. They have got reminders every week and yesterday they got a more personal letter from me. So, sorry, no decision from us yet.
Kind regards
Bengt

From: bengt.glimelius@onkologi.uu.se

To: hill@umdnj.edu, drjhpitt@yahoo.com

CC: bengt.glimelius@onkologi.uu.se

Subject: *Acta Oncologica* - Decision on Manuscript ID SONC-2013-0125

Body: 02-May-2013

Dear Dr. Hill:

I write you in regards to manuscript # SONC-2013-0125 entitled "FORENSIC STATISTICS USED TO ANALYZE PRECLINICAL SURVIVAL DATA IN NUCLEAR MEDICINE" which you submitted to *Acta Oncologica*.

In view of the criticisms of the reviewers found at the bottom of this letter, your manuscript did not obtain the necessary support for acceptance in competition with other good manuscripts. We have tried to make a careful, complete, broad and open evaluation of your manuscript. I am sorry for our decision.

Thank you for considering *Acta Oncologica* for the publication of your research. I hope the outcome of this specific submission will not discourage you from the submission of future manuscripts.

Sincerely,
Prof. Bengt Glimelius
Editor in Chief, *Acta Oncologica*
bengt.glimelius@onkologi.uu.se

Reviewers' Comments to Author:

Referee: 1
Comments to the Author
This paper illustrates the power of statistics to assess the plausibility of data generated by one Research/training Specialist (RTS). The paper is well written and quite entertaining, but it uses language that may be disconcerting to readers of *Acta Oncologica*. It would not be very difficult to simplify the paper by eliminating the more elaborate statistical explanations, and focus on the simpler concepts. The Appendix seems out of place for a clinical journal – it would belong in a statistical journal. In addition, both the paper and the Appendix contain a large number of unnecessary details. Such details belong in a research report, not in a published paper (even in a journal devoted to statistical methodology). The paper could be drastically shortened and yet lose none of its impact.

Specific comments:
• I am unsure why the authors use the term "forensic statistics". Wouldn't "statistics" be sufficient? Baggerly has similarly used the term "forensic bioinformatics", but this precedent seems equally unfortunate to me. Forensic implies some legal context, but the detection of fraud (or simply errors) in data does not have to take place only in the case of litigation.
• Page 9-10: Section 8 (Summary) contains the overall message of the paper. Essentially, the authors have used four separate approaches to show that the data generated by the RTS were not compatible with either statistical expectation, or with the data generated by other workers in the same lab. I think this overall message could be delivered earlier in the paper, so that the reader knows what to expect rather than having to go through all details first. I also believe that a sharper distinction could be drawn between the two ways in which the data were shown to be (almost surely) falsified: it is elegant to demonstrate that the data differ highly significantly from statistical expectation, but it would have been sufficient, in fact, to show significant differences from data generated by other workers, under the

reasonable assumption that no factors other than data tampering by the RTS could be the cause for such differences. Finally, the authors could discuss whether other tests than the 4 they chose could have been (or were) considered to show the implausibility of the data generated by the RTS.
• Page 10: "one run rejected the hypothesis at the 0.000000001 level." Suggest to write this 10 to the (-9) power.
• Page 12: An Excel spreadsheet is available. To do what, specifically?
• Page 29: "…we find that the probability of 1 or more such triples is 0.7767973 not at all surprising given the probabilities concerned." This sentence is incomprehensible. What is the point of giving 7 decimals?

Referee: 2
Comments to the Author
Review of "Forensic Statistics used to analyze preclinical survival data.."

Summary: A worker in a research lab, identified as RTS throughout the paper, is suspected of having fabricated data in light of three main results.
First, looking at measures taken in triples, the median of the three measures was more often very similar to the average of the three than Poisson distributions would predict. This was not the case for the other workers.
Second, the last digit of the measures taken by ETS do not follow a uniform distribution, while those of other workers do.
Third, the last two digits are more often the same than they are with others.

Opinion.
Please note that I do not conduct research similar to that analyzed (i.e., that suspected of being fraudulent) and lack of contextual knowledge may impair my assessment.

I should say that am not as convinced by the authors analyses as they are. I am mostly persuaded RTS data is significantly different from that of others, i am not persuaded that necessarily means fraud. I could be convinced, i could be convinced by mere statistical analyses (do not hold the belief that only witnesses or confessions can demonstrate fraud), but I will need more than that currently reported.

Below I present a number list of concerns/opinions/recommendations.

1. Dichotomizing a continuous variable.
The variable the authors focus on, mid-ratio, is continuous.
Why is it dichotomized for the statistical analysis in sections 5.1 and 5.2?
It would seem fruitful to exploit its variation further.

2. The Poisson assumption.
Count data is very often modeled with Poisson distributions, but count data need

not be distributed Poisson. There is a large literature on count data, which I am not very familiar with, but which provide guidance on this issue distribution (see e.g., issues of overdispersion when Poisson is replaced with other distributinos).

I am particularly worried about this issue because the results may be quite different with another distribution. For example, in section 5.1 and 5.2 the authors assess how unlikely the pattern they depict in Figure 2A vs 2B is by relying on Poisson distributions.
In particular, in section 5.1 obtain the maximum expected proportion of mid-ratios in the .4-.5 range, which is 42% across all possible values of λ, and then do a binomial test to assess how unlikely getting 716 out of 1394 bernoulli trials with this most conservative of possible λs is, arriving at a truly unlikely probability (8.31x10-32). Seen this way the gap between observed and expected is astronomical. But we could ask, how large would that .42 need to be to make 716 out of 1394 be reasonable, say p-value=.05. It turns out that the p-value they are computing is quite sensitive to the .42. If it were .5 instead of .42, for example, then the observed 716/1394 leads to a p-value of p=.14. So we better be sure that .42 is the actual upper bound, but we are not, because that upper bound relies on the Poisson assumption. Section 5.2 is similarly affected by that assumption, but I did not calibrate it by how much as it is not as simple as the previous example.
If one does a difference of proportions test for RTS and the other workers, what kind of p-value does one get?

What to do?
1) Try other count models for robustness
2) Plot the data, and make the case for it being ~Poisson
3) Estimate λ from the data and show how far it is from λ=3, the value needed to get that .42
4) Try really hard to find a way to do a bootstrap test rather than one based on asymptotic distributions. I couldn't come up with one quickly, but I suspect it must exist. Are the data inconsistent with the data? That's the best possible test of wrongdoing.

3. May other paths may lead to mid-ratio~.5
The authors write: "The quickest and easiest way to construct such a triple [a fake one] is to record the desired mean (or a close approximation) as one of the three count values and then, using two roughly equal constants, record the desired mean less the first constant and the desired mean plus the second constant as the two other count values of the triple."

I don't' find that argument very persuasive. For instance, adding a constant to each observed reading seems just as easy a way to fabricate data, and one that seems more likely to tempt someone. Notably, adding a constant to observed readings does not greatly influence the mid-ratio. This does not mean it is the one people would use, but the previous argument is just too speculative, and undeservedly focuses the reader attention on fraud from the pattern that is reported.

I wonder which ways other than fabrication, if any, may get the median of three observations to match the average. Would clumsy rounding of values do it? (see point 4) Could clumsy notes transcribed later get you there? I would like to see a more thorough consideration of alternative explanations for the data patterns that are observed. Not being familiar with the studies or paradigms (I do not read nor publish in oncology journals), I cannot formulate many alternatives, but I suspect that conversations with workers like RTS would result in a few, reporting such interviews would anyway be useful. Whether they confirm no other sensible explanation exists, or whether they result in some candidates that are subjected to the data for testing.

4 Rounding

I would like the authors to discuss the possibility that mere rounding of numbers could explain their patterns. In Table III we see excessive frequencies for last digits 0, 5 and 9, could that be all that's going on? I am not proposing this is an un-addressable alternative explanation, the opposite, it is one the authors ought to discuss and which I suspect is easy to rule out.

From Figure 4 it seems that most studies by RTS were not that abnormal, why may that be? Lack of power for detecting it? Perhaps a similar graph with the relative prevalence of digits 0,5,9, rather than the p-value associated with the $\chi 2$ test, would be more informative.

5. Testing for fraud.

After considering possible alternative explanations for the mid-ratio and rounding of numbers, and attempting to falsify them with data, I would propose attempting to find confirmatory evidence of fraud. For example, Figures 4 and 3 show heterogeneity in the prevalence of suspicious results across studies by RTS, is that variation correlated with something we may expect to correlate with fraudulent behavior? For instance, are the results that are most suspicious, those that "worked" with RTS but that could not be replicated by others those with mid-ratio~.5 and those that look less suspicious those that either did not work at RTS's hands, or that others can replicate?

6. More analysis of results.

Are the results of the experiments themselves in some statistical way impossible? I would like to see more discussion of the experimental results, are the effects too big? Are the effects too regular across experiments? Are the time-sequence of effects not consistent with biological principles or other experiments? Etc.

7. Equal digits tests

I am quite worried that this test is post-hoc, conducted only because the pattern is observed, and perhaps more importantly, that it follows directly from the non-

uniform result provided earlier. Uniform distribution will have the least repeated digits, so if RTS numbers are not ~U, then they must repeat more than those of others. I would suggest considering dropping these analyses.

8. Other comments.
8.1 Why is the excel spreadsheet available upon request only? Why not post it? I myself would really prefer seeing it before recommending publication, if the editor agrees, i would like to request it in an eventual resubmission.

8.2 The uniform digits test derives directly from Benford's law, worth mentioning.

8.3 It would be worth telling the reader more about how these analyses came about, why RTS results were being investigated, why the mid-ratio was what the attention was fixated on, etc/

8.4 I would similarly suggest posting the R code the authors refer to.

Referee: 3
Comments to the Author
General comments

The manuscript does not follow the traditional style, with Introduction, Material and methods, Results and Discussion. Furthermore, the Introduction chapter includes a description of the results of the study.

Due to the untraditional outline of the paper it is difficult to read and evaluate. Parts that usually are considered as discussion are mixed among descriptions of methods. A stricter outline had enhanced reading.

It is unclear if the manuscript contains original data, not previously published. The major part of the data seems to be published before, and much of the raw data seems to be available on an open web page.

The aims of the study are not clearly expressed. Is it to develop new statistical methods and/or use forensic statistics on radiobiological data? Is it to test if certain results from a certain researcher or research group are fraudulent?

The absorbed dose levels (or activity concentrations and exposure time) are not presented. The radionuclides studied have very different properties; those emitting alpha particles and Auger electrons will give a very inhomogeneous exposure of the cells dependent on the exposure level, while those emitting electrons with higher energy and/or photons will expose the cells much more homogeneously. This fact may influence the statistical analysis of data.

The research data presented in the eight papers studied are all experimental results

that seem to be related to bystander effects or effects of low dose irradiation. The clinical procedures used today do not take into consideration non-targeted effects and this type of dose rate related effects. It is therefore most unlikely that any of the results in these papers have yet influenced the patient handling in any case. However, more recent research on non-targeted effects, and then especially some of the in vivo studies, has demonstrated that non-targeted effects and low-dose-rate related effects may play a role in clinical use of radiation, both for diagnostic and therapeutic use. Then data from the present researchers (in the 8 examined papers) may be of some interest, although in vivo studies will be of most importance, since the radiobiological response depends to a high degree on physiological factors from the entire organism, e.g. immune and inflammatory response. If non-targeted effects will be taken into consideration in the future, I would foresee modification of risk models for low doses and dose rates, and maybe modifications in risk estimations on normal tissue toxicity in radiation therapy, although in the latter case such effects might be already included in the empirically defined tolerance doses in external radiotherapy.

Detailed comments

Page 3 line 34: A “to” is missing in the text.

Page 4, line 29: Two words are written together.

Page 5, line 6-8: The denotations of the radionuclides are not correct and not consistent. Further, the chemical form of the radionuclide should not be given only for some of the radionuclides.

Editor
Comments to the Author:
This is a very special manuscript. I have considered the comments by the three referees who above have given their written opinions and in addition a few others. I consider that they all have high integrity and have tried to make a fair evaluation. In its present format, the manuscript can not be accepted. It is extremely difficult to read and does not give a good presentation of the issue. Even if the text may be entirely rewritten and in a format that is much more easy for readers to appreciate and understand, I am very hesitant to recommend publication.

Date Sent: 02-May-2013