Reviewer's report

Title: Pancreatic Cancer Clusters and Arsenic-Contaminated Drinking Water Wells in Florida

Version: 4 Date: 9 December 2012

Reviewer: How-Ran Guo

Reviewer's report:

From the cover letter, I believe the authors have tried to address my major concerns that they failed to cover. My comments are as following:

Major Compulsory Revisions

1. The figure provided in the response is helpful and thus should be included if the Editor decides to publish the manuscript. The formula I was looking for is for the logistic regression model, not for the likelihood ratio test, nor for the SaTScan program. There is no harm to include those descriptions, but they can be omitted if the Editor feels that the size of the manuscript should be cut down. While the authors still failed to give the formula of the logistic regression, the description in the revision is sufficient for the readers to understand the models. Nonetheless, they stated in their response that “The dependent variable was the binary outcome that whether the block groups categorized as having either an excess or expected pancreatic cancer incidence by the cluster detection software described above.”; but in the text they stated “the dependent variable was a patient with pancreatic cancer living in a neighborhood with a higher than expected pancreatic cancer incidence (a “cluster”) versus being diagnosed not in a cluster.” These two are different. To begin with, one used a “block” as the unit of analysis, but the other used a patient as the unit. Furthermore, even using a patient as the unit, as I pointed out in the first review, the authors should discuss briefly the issues involved in extrapolating the model used in this study to the model corresponding to the research question directly (Why the association between arsenic exposure and being diagnosed in a cluster of pancreatic cancer can be explained as an association between arsenic exposure and the occurrence of pancreatic cancer itself? Were there no alternative explanations?).

2. My previous comment was that since the model has included smoking as an independent variable, the confounding effect of smoking cannot be used to explain the association, unless there were residual confounding effects. In their second reply, the authors explained that it might due to an interaction between arsenic and smoking. If so, they should demonstrate it by including an interaction term in the model. As an alternative explanation, the authors admitted that their “data on smoking was [were] both incomplete and subject to potential misclassification.” If so, they should not include the variable in the model.

3. While the authors tried to explain why there seemed to be a protecting effect of smoking in the cover letter (again, they offer the fact that data on smoking were both incomplete and subject to potential misclassification as an explanation) but
did not discuss this issue in the revised manuscript.

4. The revised Figure 1 still shows a discrepancy between the distribution of arsenic contaminated wells and distribution of higher than expected pancreatic cancer clusters: many green spots were not in a red area, and many red areas had no green spots. Can the authors calculate the proportion of green spots in red areas (blocks) and the proportion of red areas with green spots?

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I declare that I have no competing interests.