

Author's response to reviews

Title: Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: An application of generalized additive models to case-control data

Authors:

Veronica M Vieira (vmv@bu.edu)
Thomas F Webster (twebster@bu.edu)
Janice Weinberg (janicew@bu.edu)
Ann Aschengrau (aaschen@bu.edu)

Version: 2 Date: 16 January 2009

Author's response to reviews: see over



**Boston University
School of
Public Health**

**Department of
Environmental Health**
Talbot 2 East
715 Albany Street
Boston, MA 02118-2526
TEL: (617) 638-4620
FAX: (617) 638-4857

Dear Editors,

We are submitting a revised manuscript “Spatial analysis of bladder, kidney, and pancreatic cancer on upper Cape Cod: An application of generalized additive models to case-control data” that addresses the comments of the reviewers and Environmental Health editors. We are including a detailed response to the reviewer’s concerns. Per the suggestion of the editors, we have rewritten the Conclusion and replaced Figure 1 with two figures (Figure 1 and 2) to allow for more detail. Lastly, we made the requested formatting corrections.

Best regards,

Veronica M. Vieira, DSc
Assistant Professor of Environmental Health
Boston University School of Public Health

REVIEWER 1

MAJOR COMPULSORY REVISIONS:

- 1. The analyses use multiple locations for individuals who move (p. 5 talks about 'case locations' and p. 9 about multiple residences for the same individual but this is not fully clarified until p. 15 in the discussion), thereby having more 'observations' than individuals. This does not seem well justified and the authors make brief reference to the possibility of bias at the bottom of p. 9, but I felt that the statistical issues were not sufficiently addressed. For one thing, the outcomes are no longer independent. For another, individuals who move a lot are weighted more heavily than those who do not move, with weight equal to the number of residences. Accounting for residential history is a difficult problem, but creating replicates of an outcome to assign to each address does not seem well-founded and needs further justification. What is the statistical underpinning of this? At the very least, I would like to see it stated clearly in the methods that the approach replicates each outcome with the same covariates but different residence for each residence.*

The authors note that for bladder cancer, in using longest residence, the optimal span is much larger than when using multiple residences, raising concern that the multiple residence results (which have a smaller span) may be biased away from the null. For the other analyses, it appears the sample size is sufficient only when using multiple residences; if the maximum residence approach does not have sufficient sample size, I am wary about analyses of the same data that rely on the multiple residence approach, which seems to artificially inflate the sample size. Does going from 37 cases to 49 case locations really make the analysis feasible with the larger size (49) but not the smaller (37)?

Response: Accounting for residential history is a difficult but important problem and we appreciate the comments from the reviewer regarding this issue. We acknowledge that keeping all of the addresses in the analysis may result in bias, e.g., creation of a spurious cluster by the same case moving within a small area. We have added text in the methods (pg 9-10) to emphasize that our multiple residence analysis replicates each outcome with the same covariates but different residence for each residence. We check for the possibility of bias by restricting to the residence of longest duration: if the restricted map appears similar to the map including all eligible residences, we judge that the use of multiple residences did not cause bias. On this basis, and as discussed in the original manuscript, we therefore believe that our bladder cancer analysis (with 15 years of latency) is not biased (pg. 10). To follow-up on the reviewer's comment, we conducted additional analyses of pancreatic and kidney cancers that restricted addresses to one per individual, based on longest residency. The resulting map for kidney cancer was quite similar to the original, suggesting little bias (text added on page 11). The resulting map for pancreatic cancer differed from the original however, suggesting that the pancreatic cancer analyses may be biased (text added on page 12).

Use of all residences and residence of longest duration are both imperfect solutions. Because the outcomes for an individual with multiple addresses are always the same, a correlated data approach is impossible. By including all residences, we examine spatial variation alone with no temporal consideration. When we consider space and time together, the odds of disease likely varies depending on which address is used because each address serves as a proxy for the location and timing of exposure. Multiple residences are ultimately a space-time issue, an important problem that we are currently working on. One possible solution is to smooth on both space and time (which can be done in many ways); we used on such approach in a prior, much larger analysis of breast cancer. However, this requires a larger sample size than we have available for the current analysis. Additional text has been added on this point on page 15.

- p. 8: It seems to me that since the span selection is part of the model fitting, in doing the permutation test, one should estimate the span from the data for each of the 999 permutations. Otherwise one conditions on the amount of smoothing selected based on the data, treating that as the correct value, when under the null, it surely is not. Could the authors either further explain their logic or consider estimating the span for each permutation? Is this a computational limitation?*

Response: We appreciate the reviewer's comment. This is an interesting statistical point that requires further work. We agree that our permutation tests are conditional on span size; we do this to maintain the same spatial resolution. We have added text to clarify this point (page 8). As discussed on page 13 of the original manuscript, choice of span size is one of the most important issues in smoothing; choice of span size to optimize the bias-variance trade-off is not necessarily the same as understanding the importance of various features of the map. Furthermore, while the use of different span sizes for each permutation is an interesting idea, the process of selecting an optimal span size cannot be automated using our current computational methods because `step.gam` finds local rather than global minima, and manual selection of 999 spans is impractical.

MINOR ESSENTIAL REVISIONS:

- Figures: The first plot (1a) shows the MMR as a separate 'town', but later plots do not indicate the MMR. I would like to see the MMR indicated on all the plots, since proximity to the MMR is a key exposure hypothesis.*

Response: We show the MMR in Figure 1a primarily to explain the sparse population distribution in that area shown in Figure 1b (Note: we have modified Figure 1 so that Figure 1a is now Figure 1 and Figure 1b is now Figure 2). We feel it is inappropriate to single out the MMR in every map when there are other exposure hypotheses including ground water plumes from landfills and waste water treatment facilities.

- p. 3: Wood's book (Generalized Additive Models; 2006) would be a very appropriate citation along with [13-17]; same comment for page 6, [13] citation.*

Response: We have added the Wood reference

- p. 4: Why were only cases from 1983-1986 used?*

Response: The current study is a secondary analysis of an existing case-control study that used cases from 1983-1986. We have clarified this point in the Introduction.

- p. 6: I wouldn't call $x_{\{1\}}$ and $x_{\{2\}}$ 'longitude' and 'latitude' since they are projected coordinates.*

Response: We acknowledge that the use of longitude and latitude is technically incorrect because we use a projected coordinate system. We feel the terms longitude and latitude are easier to understand than x- and y-coordinates. We have added a sentence (page 6) informing the readers that the measurements were technically not longitude and latitude.

- p. 7: How was the number of parameters in the AIC calculation determined - is this based on an estimated degrees of freedom for the smooth term?*

Response: Yes, the AIC is based on estimated degrees of freedom for the smooth term.

- p. 8: Strictly speaking, the permutation test is not 'with and without the smooth term' since the smooth term is in the permutation-based model fits, no? I would instead say something like 'with and without spatial structure'.*

Response: On p. 8, the text 'with and without the smooth term' refers to the two different models being fit for each permutation of the data. A deviance statistic is calculated comparing the model with and without the smooth term.

- p. 11: I would point out that the hotspot for kidney cancer with span 0.15 in the MMR is surely spurious as there are no addresses within the hotspot.*

Response: We have added a sentence (page 11) stating that the hotspot is likely spurious due to the sparse population in the area.

DISCRETIONARY REVISIONS

- Figures: To my mind, given that there is a null value, I would choose a color scheme with white as an odds ratio of 1 and blue and red for values less than and greater than one. As it is, the values near zero are at an intermediate color and one has to go back and forth from the legend to the map to determine which areas are near the null value. Plotting as log odds may also be an improvement to achieve symmetry of odds ratios larger than and less than one.*

Response: There are many ways to display maps, a point we discussed in earlier papers (references 16, 17). We believe that the use of odds ratios rather than log odds is more intuitive for the general reader to understand. Our current color scale uses light blue for odds ratio of one. We have considered many scales, including the use of white for OR=1, but we believe that the current scale is effective.

- p. 8-9: The local permutation test does not account for multiple testing, so I'm concerned that too many areas may be detected. On the other hand, by virtue of doing smoothing as the initial modeling step, there should be some protection against false positives (see refs below in the Bayesian context, as this is somewhat analogous to what is being done with the smoothing in the manuscript, thereby borrowing strength spatially). Is the false discovery rate approach a possibility here to be more sure about protecting against false positives? There is now a literature on using FDR techniques in the face of spatial correlation (including a paper I was involved in: Ventura et al. 2004, J of Climate 17:4343-4356). Another possibility is to use a cluster detection type algorithm such as SatScan to see if the hotspots seem robust (though I haven't looked to see if SatScan is appropriate for case-control data). At the least I'd like to see some mention of the possibility of a multiple testing problem or why the authors think this is not an issue (perhaps based on the effects of smoothing I refer to above).*

Response: We appreciate the reviewer comments on the multiple testing problem and have added text on page 15 to address the multiple comparison problem. We have not yet explored a FDR technique but we will examine the references provided by the reviewer for our future work. In a prior analysis (Ozonoff et al. Environmental Health 2005; 4:19), we compared our crude results to SatScan results. We saw good agreement when there is one circular cluster and

different results in other circumstance. SatScan does not easily allow for multiple covariates when using binomial outcome data.

3. *p. 14: Are there any mechanisms for the ground water plumes to be related to bladder cancer, such as known carcinogens in the Cape Cod plumes?*

Response: While common contaminants in landfill plumes include solvents, we do not have data on the exact composition of the plumes within the study area.

REVIEWER 2

MINOR ESSENTIAL REVISIONS:

1. *Case-control recruitment: When the study population is first introduced, on page 4, I would like more details about how the controls compare with the cases in terms of spatial recruitment. Am I to understand that each case (e.g., bladder cancer cases) has ~14 controls living in the same town around time of diagnosis? How does the original matching on town influence the spatial distribution of the case-control populations in this study? I worry that if the town-based matching was well executed, then these data may be over-matched on spatial factors, making it difficult to detect any underlying spatial factors that may be present. The latency analysis is one way to get around this problem, and may help explain why clustering was detected using that approach. I don't think this recruitment strategy would be a large problem using historical residences, but by including residences at time of recruitment in the analysis, this recruitment strategy might be biasing the results. The authors should consider and discuss how recruitment might influence their results.*

Response: The current study is a secondary analysis of an existing study population. Controls were chosen from the same study area which consists of 5 towns in Massachusetts. However, controls were not matched on town. Therefore, there is no matching by spatial factors and we believe recruitment does not influence our results. We have added a sentence on page 4 that clearly states controls were not matched on town.

2. *I wish there were greater numbers of cases available to enable something more than spatial-only analyses of temporally aggregated data. One of the great strengths of the GAM approach is the ability to identify spatio-temporal clustering using time-resolved mobility histories—it is too bad this dataset did not permit this type of temporally detailed analysis. Understandably, the authors chose not to consider the full range of temporal variability in their data because of small sample size. However, in the discussion, I would like to see mention that residential histories in case-control studies provide a valuable resource for generating hypotheses about location and timing of exposure. More temporally resolved analyses might provide greater insights into importing time periods of susceptibility. Similarly, mapping where people live at different ages, also might allow for observing different patterns that could lead to new hypotheses.*

We appreciate the reviewer's comment on the matter of space-time analyses. Our prior work includes such analyses with breast cancer, which had larger numbers of cases. We have added text in the discussion addressing the usefulness of residential histories to generate additional hypotheses for timing of exposure.

3. *At the end of the first paragraph on page 12, the authors state that there was a lack of clustering in the kidney and pancreatic cancer analyses. But I understood the results to indicate that*

clustering was present in these analyses. A hot-spot along the southern shore for kidney cancer, and a couple reddish areas for pancreatic cancer that were significant after accounting for alcohol-related behavior. Am I misinterpreting the results? This requires clarification.

Kidney cancer did result in a significant elevated risk area in the southern study region that we described as a sloped surface. Although there was little variation in the x-direction, the variation in the y-direction was significant. The sentence should have read that the kidney cancer cluster was less pronounced. We see how this can be confusing, so we have chosen to remove the statement.