Author's response to reviews

Title: Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study

Authors:

   David C Wheeler (wheelerdc@mail.nih.gov)
   Anneclaire J De Roos (aderoos@fhcrc.org)
   James R Cerhan (cerhan.james@mayo.edu)
   Lindsay M Morton (mortonli@mail.nih.gov)
   Richard Severson (severson@karmanos.org)
   Wendy Cozen (wcozen@usc.edu)
   Mary H Ward (wardm@mail.nih.gov)

Version: 2 Date: 31 May 2011

Author's response to reviews: see over
May 31, 2011

Professor David Ozonoff, MD
Editor-in-Chief
Environmental Health

Dear Professor Ozonoff:

We are pleased to resubmit our manuscript entitled “Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study” for consideration as a research article in Environmental Health.

We have addressed all the comments of the two reviewers in a point-by-point response in this letter and have made revisions to the manuscript.

We believe this manuscript makes a novel contribution to the NHL literature. Results of this study will lead to future investigations to evaluate possible reasons for the significant clusters in Detroit, Iowa, and Los Angeles that may lead to new hypotheses about the etiology of NHL.

We look forward to hearing from you regarding your decision on publication.

Sincerely,

David Wheeler, PhD MPH
Cancer Prevention Fellow
National Cancer Institute
6120 Executive Blvd
Rockville, MD 20892
wheelerdc@mail.nih.gov
phone: 301.435.4702
fax: 301.402.1919
Reviewer: Jared Aldstadt
Reviewer’s report:

Minor Essential Revisions
1. There is a clear discussion of the overall number of cases and controls included in the analysis and the exclusion criteria used to arrive at this number. I did not find the number of cases and controls included for analysis at each SEER center. I think this should be included in the manuscript. How does the sample size at each center relate to the ability to find clusters in each area?

We added a descriptive table for the study (Table 1) that includes the percent of total cases and controls per study area and a comment about the table (p. 9). The distribution of cases and controls across the study centers is fairly even.

Discretionary Revisions
2. The authors note that the time periods used in the analysis are, “...in a time frame of etiologic relevance.” More support for this statement would make for a stronger contribution, but I realize that it may not be possible. I know that latency and exposure for most cancers are complex and highly variable. Issues in data reliability and availability are a factor as well. However, seeing the result that residence at 20 years prior to diagnosis is the most significant spatial predictor, naturally makes one curious to see the analysis repeated for 25 or 30 years prior to diagnosis.

We agree that latency for NHL is complex and variable. There is a lack of published information on latency for NHL related to environmental exposures. Current evidence for NHL latency exists for organ transplant patients from start of associated immunosuppressive therapy to NHL diagnosis and for time from HIV infection to NHL diagnosis (as discussed in Engel et al., Cancer Epidemiol Biomarkers Prev, 2007). Based on our findings, we agree that it is worthwhile to explore longer lag times of environmentally related exposure for NHL in future analyses.
Reviewer: Veronica Vieira  
Reviewer’s report:

Minor Revisions  
1. I’d like to see data describing the number of cases and controls for each location, perhaps in a table. Over the 20-yr residential history, what was the average number of addresses per participant? Was there a lot of movement in the population?

We added Table 1 to describe the study characteristics and the distribution of cases and controls by study center.

The mean number of residences per subject was approximately 2 in each study center over the 20 year history. We added this information to the text (p. 9-10).

2. Did geocoding results differ by year of residency? For the different latency periods, it would be interesting to know the percentage of geocoded addresses. Also, among the addresses geocoded to the ZIP code level, were they evenly distributed across space? By disease status?

We looked at the percent of addresses that were excluded by residence end year (those matched to a populated place, county centroid, or county centroid). The percent of addresses excluded generally increased with years before study enrollment. The trend continues looking farther back than a 20-year time lag. We added a comment about this geocoding accuracy to the text (p. 9).

Most of the addresses geocoded to the ZIP Code level were located in Iowa, likely due to a larger prevalence of PO Box and rural route addresses in this study center. The spatial pattern of these addresses appears random. Assignment of case addresses to the same ZIP Code does not appear to explain the cluster of elevated risk identified in south-central Iowa, as there was no co-location of any of these cases at the ZIP Code level. We added a statement about the lack of pattern in the ZIP Code based matches to the text (p. 9).

3. Recent work has shown that significance tests for GAMs have an inflated type I error rate (Young et al., Comput Stat Data Anal, 2011). The authors should add a disclaimer to their discussion that the p-values they’re using to evaluate the significance of their latency periods are likely biased. As most of the results are not statistically significant, this isn’t really a concern here but it’s something to keep in mind.

We added this citation and a disclaimer statement about the p-values (p. 15). Thank you for the reference.

4. For your comparison of different latency periods (i.e, Figure 6), did you use the same span size or the optimal span for each latency period? Please add this to your figure title and text.

We used the optimal span size for each lag time. We have added this to the caption for Figure 6 and the main text (p. 19).

5. The sensitivity analysis for selection bias is a good idea. The authors might also want to consider using a GAM model with participation as the outcome to determine if participation varied by location.
Shen et al. (2008) analyzed nonresponse in the NCI-SEER NHL study spatially through logistic regression and a cluster analysis and did not find a significant cluster of nonresponse after adjusting for demographic factors. We have expanded the description of this study to make it clear that nonresponse status has been analyzed spatially in this study (p. 22). We believe our sensitivity analysis using subjects who did and did not respond is the best evaluation of the impact of nonparticipation on the spatial-temporal analysis.

6. The second paragraph of the discussion could use some clarifying. Specifically, I find the wording of lag times treated as a variable somewhat confusing and would suggest revising that. Assuming you have no time-varying covariates, only the smooth term will vary across latency models so the difference in p-values reflects differences in the address smoothing. Also, the text concerning reference 27 is somewhat misleading as that paper also used a single address per subject in their 20-yr latency in addition to analysis of several lag times that included calculating the global and local p-values.

We changed the wording when discussing the selection of the lag time to convey that the smoothed function of the residential locations may be considered a term to select in the model (p. 20). We believe this is less confusing now. Thank you for the suggestion.

We added a sentence to state that reference 27 included an analysis of one residence per subject using the residence of longest duration (p. 20).

7. The paper concludes with a sentence about generating new hypotheses but it doesn’t actually provide any follow-up thoughts about the areas of significantly elevated risk. In the results section, I found it interesting that the cluster in the LA analysis is in an area of high HIV+ prevalence but the analysis excluded HIV+ participants. Is there some community variable that would result in the clustering of these two outcomes?

In the Discussion section (p. 22), we mention datasets that we are going to explore to investigate the clusters found in this study. Evaluation of variables from these datasets could lead to new hypotheses about NHL etiology. This research is underway now.

For the NCI-SEER NHL study, individuals were instructed that they must reside in the study center at the time of diagnosis and not be HIV positive to be eligible to participate. They were then asked if they were eligible to participate. Therefore, we do not have an HIV variable to adjust for in the models. We are not aware of a possible area-level variable to include in the models for each study center to adjust for HIV prevalence, but we would consider an appropriate one in future work.

8. It would be helpful if the figure legends included the span, p-value, and confounders included in the model.

We added the span, p-value, and adjusted variables for each map to the figure captions.

9. As this journal is online, please consider using a color Odds Ratio scale.

We feel that grayscale is appropriate and effective for showing differences in quantities such as odds ratios and would like to keep this choice of shading.