Reviewer's report

Title: Highway proximity associated with cardiovascular disease risk: the influence of individual-level confounders and exposure misclassification.

Version: 2 Date: 14 May 2013

Reviewer: Sandrah Eckel

Reviewer's report:

Thank you for your responses to my initial comments, your revisions, and the opportunity to re-read this manuscript. Note that the updated version of the paper I received did not have changes highlighted – which limited my ability to detect changes from the previous version. I still have several major issues which remain from my initial review.

Major comments:

1. Distance variable:
The authors have rich continuous distance data, but coarsen to a categorical variable. Due to the lack of data between 450-1000m, I agree that it is sensible to perform primary analyses with their proposed categorical grouping. However, when trying to identify the “exposure-response” relationship between distance and their CVD outcomes, it would be much more interesting to subset the data to those participants with distance <450m (near highway group, approximately N=207) and fit a generalized additive model (GAM) which allows for a smooth effect of the continuous distance variable (GAM does not require the effect to be linear). GAM will avoid the “problems with maintaining sufficient sample size” in distance categories. Given the “steep pollution gradients within 200m of a highway”, one would expect to see an interesting shape of the exposure-response relationship in the 0 to 450m range of exposures. The authors have gone through a lot of trouble to reduce exposure misclassification by using refined geocoding methods and they are throwing away information by coarsening the continuous distance variable. The authors might also see the most interesting impacts of the “rarely included” confounders on the distance effect within the steep exposure gradient very close to the freeway.

2. Confounding:
I still have two major issues with the author’s emphasis on and treatment of confounding, though I agree with the authors that it is nice to have information on “rarely included” confounders (e.g., time activity and time spent on highways, residential window opening, recent combustion exposures, and air conditioning). Unless the authors can show very compelling evidence of novel/dramatic impacts of adjusting for these “rarely included” confounders, I think the paper would be much stronger if the authors included these potential confounders along with the usual set of confounders (and listed this as a strength of the study in the discussion), but made the improved geocoding methods the main emphasis of
the paper. In this case, the more interesting geocoding results should be moved from the supplement (e.g., misclassification table) and, for example, the figure of unadjusted results (Figure 2) could be cut.

Issue 1: The authors’ model building procedure is ill-suited to their purpose. I agree with the initial screening of variables to identify potential confounders from a long list of covariates, and forcing inclusion of age, sex and smoking status. I do not agree with using the forward stepwise selection procedure (p of 0.15 as entry/exit criteria) to decide whether to include/exclude the potential confounders. Potential confounders should be included if they change the effect estimate of interest (not if the potential confounder variable has a significant slope). It is reasonable to implement a quasi-“step-wise” procedure to investigate a long list of potential confounders where one defines sets of potential confounders and includes/excludes each of these sets based on whether the effect estimate of interest changes. For easy interpretation, I prefer a parsimonious model that includes only the essential confounders (usually choose same set for all related outcomes being considered), but I will sometimes include other covariates that are strongly significant if they explain a large proportion of variation in Y and hence reduce the standard error of the estimate of the effect of interest.

Issue 2. The authors’ have not addressed how the inclusion of the “rarely included” confounders (in addition to the “usual” confounders) changed the effect estimates of interest when compared to a model that controlled only for usual confounders. For example, Table 2 compares an unadjusted model, a model adjusted for “rarely included”, and a model adjusted for “rarely included” and “usual”. It would be more interesting to see a model with “usual” only than a model with “rarely included” only. Which of the rarely included confounders were most important? Which were least important? Currently, the authors have not convincingly shown “the importance of controlling for confounders rarely included in roadway proximity analyses” and haven’t provided information on whether these variables should be collected in future studies and which should have high priority. Supplemental Table 7 only addresses the association between Y and the “rarely included” covariates, not the confounding effect of the “rarely included” covariates.

Minor issues:
- Manuscript still says: “All statistical tests, with the exception of Wilcoxon, were two-sided” even though that authors had said in their first response to review that “We conducted a two-sided test and have changed one-sided to two-sided in the text.” The text should be changed to: “All hypothesis tests were two-sided.”
- Table S1 column heading should be “>1000m” not “<1000m”?
- In Supplemental Table 4, which set of adjustments are used? (“exposure adjusted” or “fully” adjusted)? Clarify in table legend.

Level of interest: An article of importance in its field

Quality of written English: Acceptable
Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

No competing interests.