Reviewer’s report

Title: A multivariate analysis of CalEnviroScreen: comparing environmental and socioeconomic stressors versus chronic disease

Version: 0 Date: 28 Aug 2017

Reviewer: Lynne Messer

Reviewer's report:

Environmental Health manuscript number: ENHE-D-17-00227

Title: A multivariate analysis of CalEnviroScreen: comparing environmental and socioeconomic stressors versus chronic disease

The authors sought to assess the relationships among the variables that contribute to CalEnviroScreen. They assessed the measure overall (20 variables), the 12 environmental variables, and the 5 sociodemographic variables, all in association with their constructed chronic disease outcome measure. They found the use of CalEnviroScreen supported by their analysis.

Thank you for the opportunity to review this paper. It promises to make an interesting contribution to the literature, and especially to folks using CalEnviroScreen (CES). I have some suggestions I think the authors should consider in a resubmission.

In general, I am unclear exactly what the authors are trying to accomplish. If they are just trying to describe the associations among the variables that go into CES, I am surprised that work wasn't undertaken as part of its development and refinement. If they are trying to justify its use with health data, they could use a lot fewer words to accomplish this task. The goals, approach and utility of their work needs better clarification.

The justification for analysis approach seems inadequate. I am not clear if authors are using PCA because that is how CES constructed? If so, why is it meaningful to say the PCA of all 20 variables is associated with the CES? If the CES was not constructed with PCA then why did the authors use PCA to empirically summarize the variables? If they wanted to understand the respective amounts contributed by various components, why did the authors not construct those components in a manner consistent with CES and then do a decomposition analysis to see how much of the overall score each component part was responsible for?

Regarding the outcome the authors constructed (DB), they used LBW in the outcome construction, which is also included in the exposure (CES) construction, which is problematic.

Also, it seems the authors had to undertake rather a lot of manipulation to make their outcome variable (DB) usable in their statistical models. Why not focus on a subset of the outcomes? Or
each outcome individually? The outcomes they included are not the only categories plausibly associated with environmental factors. Further, since this is hospitalization data, at the zipcode level, this is not a terribly specific measure of some health outcome. The authors could have made it easier for both themselves and the reader. Since the outcome categories were somewhat arbitrarily chosen, why not make choices that make it easier to describe/model exposure-outcome relationships? And yes, I saw the justification for combining the outcomes (to avoid type I error), but given the multiple testing undertaken, having a few outcomes seems like the least of the statistical problems the authors should be concerned with.

If there are no meaningful differences in the CES across versions, then it would be better for the authors to pick one and describe that. If there are meaningful differences, then the authors should consider describing those in more detail. As it is, it is messy to try to follow.

Why are there equations in the results? Shouldn't the description of the approach be presented in the methods?

Given the lack of focus as to what the authors are trying to accomplish, I am unable to derive meaning out of the results. The authors have a lot of text, describing somewhat simple results (correlations). I think the paper could be pruned to the specifics of what the authors want to contribute and it would be more useful and easier to follow.

**Level of interest**
Please indicate how interesting you found the manuscript:

An article whose findings are important to those with closely related research interests

**Quality of written English**
Please indicate the quality of language in the manuscript:

Acceptable

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal