Reviewer’s report

Title: Probabilistic Walking Models Using Built Environment and Sociodemographic Predictors

Version: 0 Date: 07 Jan 2019

Reviewer: Abraham Flaxman

Reviewer's report:

This paper investigates the ability of logistic regression to predict physical activity (PA) levels of individuals, based on demographic variables alone or in combination with built-environment (BE) variables.

Despite the optimistic claim of the abstract, the findings are primarily negative: the authors were able to produce only a highly imprecise predictor whether they included BE variables or not. The ROC AUC values for all their best models were less than 60% (50% is equivalent to random guessing).

In my opinion, a negative finding of this nature is still worthy of publishing, but must be presented in a cautious manner. It may be tempting to claim that this work shows that BE factors do not predict PA, but what the authors have shown is that the methods they explored were unable to predict PA. Perhaps different methods or more/better data would have succeeded.

With sufficient revision and clarification, I think that this will be an acceptable paper for POHM.

Detailed Comments:

Intro says effects are nonlinear. Sounds like methods will only capture linear effects, though, since it uses logistic regression with continuous variables.

Methods say outcome var was dichotomized, to 150 MET, but survey questions make it hard to report 150 minutes per week! Would alternative thresholds for PA be more predictable?

The walking outcome is similarly subject to noise and bias; is the potential for misreporting systematic?

I would like to know more about the sample frame: what is the response rate for twin survey? Which twins were selected as individuals for your study?

More detail on all of the BE variables would be valuable in the methods section: how was slope quantified? What boundary was used for Seattle city limits? What precisely constitutes a neighborhood commercial center? All of the other variables could be made more precise as well,
and examples would help the reader understand the potential for predicting PA from these factors individually or in combination.

Many of these factors change over time, as well. How did you deal with temporal misalignment?

P 6, line 133: what is a twin's InfoUSA destination?

P 7, line 147: Pearson correlation will miss effects that are too nonlinear, e.g. a U-shaped relationship

P 7, line 157: Step two considers nonlinearity after all the non-monotonic relationships have been removed, so it is not surprising that it doesn't find anything.

P 8, line 161: Distance from home to single nearest feature seems highly sensitive; I recommend some more robust version of distance.

P 8, line 170: How did you used the sibship of twins? Was it part of the sampling design?

P 8, line 172: backwards selection with an ad hoc process to force inclusion of certain predictors seems non-standard. Would a more common approach, such as gradient boosting machines, do better?

P 10, line 213: perhaps including interactions between age group and BE factors would improve predictions; BE features seem likely to be of different value to older and younger twins.

P 15, line 326-328: Related to my comment on the careful framing of this null result, I think it is important to make replication data available, so that other researchers can attempt alternative prediction schemes.

Line 250: if you found something, this would be good. but since you didn't find anything, I'm not convinced that you reduced the predictor space effectively.

Line 257: what is walkability, if not the factor that leads to more walking?

Line 288: r=0.46 is not that correlated. Maybe the mis-reporting noise drowns out any signal here.

Line 298: this method failed to find any BE factors that increase predictions of PA. But that does not imply that they do not exist.

Could the twins help resolve this in some way? It seems like you have not actually used this highly unique aspect of the data design.

Could the change over time help? Either the change of address for a twin, or the change in the built environment around the twin could be used to develop a quasi-experimental design.
It seems like you have many sources of noise: self-reporting with social desirability; the changing neighborhood; the 4 stage estimation procedure you have developed. Concluding that none of these BE factors affect PA seems risky.

If the point of this work is to identify modifiable factors that could increase physical activity, there must be some attempt to identify the causal pathway---does choosing Seattle over the suburbs cause an active life, or does an active lifestyle cause choosing Seattle?

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