Reviewer’s report

Title: Probabilistic Walking Models Using Built Environment and Sociodemographic Predictors

Version: 0 Date: 09 Jan 2019

Reviewer: Roy Burstein

Reviewer's report:

This was an interesting paper, looking at associations between individual and neighborhood characteristics and two measures of physical activity. The authors did a tremendous job compiling data from various sources to create built environment (BE) indicators. Using survey data from individuals in the greater Seattle area, the authors fit several statistical models and tested associations between these potential predictors and self-report walking or moderate-to-vigorous physical activity (MVPA) times. BE factors are seen as mutable by policymakers, unlike socio-demographics, and thus they were interested in determining which BE factors could explain differences in activity. They developed a four-stage approach: first they kept only certain covariates based on univariate correlation coefficients; second they tested for non-linear effects; third they used a backward elimination method to keep variables that remained correlated in a multivariate analysis; fourth they validated their predictive models. Ultimately, they found that adding BE variables to a model based on individual sociodemographic factors did not improve the model - in other words their BE variables did not explain variation in the probability of MVPA or walking.

Minor revisions:

* Background, line 6: I was confused by what 'objective data' means here. After some context I guess it means actually measured versus self-report. Is this correct? You use this term a lot in the paper and it confused me so it may be worth explicit defining.

* Background, line 32: 'Additionally, the increasing availability of detailed objective data on BE can help avert statistical errors due to the "modifiable areal unit problem"'. I find this sentence a bit confusing, though I think you mean that you have precise GIS data on things like market locations, and this is better than, say, aggregated data on total number of markets in a neighborhood? If so, could be made more clear.

* A map of the study area would be a good addition to this paper.

* Line 149: I think a useful figure to add here would be a table or heatmap of univariate correlation coefficients across all covariates and models.
Table 3: The effect sizes on the transportation count variables are super small, maybe consider log-transforming?

Broader questions and suggestions for the authors:

* Line 57, when describing the self-report measures for physical activity, you should also mention the validation of such measures. You bring this up later in the discussion (line 286). It would be good to move this up, and possibly in a supplemental document give more information on this side-analysis for validation, since it sounds interesting and could serve to instill more confidence in the self-report data.

* It seems you have good justification, based on recommendations, for dichotomizing MVPA at 150, but I see no similar justification for walking to be dichotomized at this level. Have you done any sensitivity analyses to adjust where this threshold is? Or maybe using a continuous outcome. It does seem like you needlessly lose information by dichotomizing here. It would be interesting to see a sensitivity analysis as a part of a supplement, if possible.

* You use household income, and not income per capita. Why is this? I would assume a HH of 7 making $50k is very different than a HH of 1 making $50k.

* Can you explain how novel it actually is to develop a multivariate model that includes both individual and BE covariates? Perhaps devoting a short paragraph to this in the intro section would be useful. I do not work in this field so it would be good for an outside reader like myself.

* The title of the paper and the language throughout suggest you are interested in coming up with something like a score for walkability, based on probability of walking greater than a certain threshold, though most of the paper reads like you are more interested in the effects of BE characteristics. It seems like you are trying to toe the line between making a predictive model, and doing inference on potential determinants. For policy relevance, I would think the latter is more useful. And if you wanted to predict as good as possible, you would likely choose other methods. I think this paper would be stronger if there wasn't this ambiguity in the goal of this analysis. Ultimately a diagnostic like a ROC plot is about prediction: can your model discriminate walkers and non-walkers, maybe this isn't actually what you want in this case? Besides, you say your intent is on making an index, but in the end there was no index. To me an index sounds like it would be some reduction of a set of BE attributes that you can use in various situations to determine something like walkability or MVPA-ability. Would you ultimately like a policymaker to determine walkability of a place based on the attributes chosen by your analysis? Or would it be more useful to understand the magnitude of the effect of each characteristic?

* Line 27: "Only a consistent treatment of environmental predictors and activity outcomes will help untangle their relationship for different populations in different settings". In this part of
the background you are setting up the problem that you may overcome in this paper. Are you suggesting that you have overcome this in this paper? Is there a reason the questions used for the outcome here should be used more widely? Should others mimic the bespoke GIS analysis you did to create your BE covariates? If so, you should come back to this in the discussion, if not maybe consider removing it from the Background.

* Is the use of the twins data out of convenience because it already existed or is there some other justification for it? There seems to be no benefit here to use data from twins so this caused some confusion for me as twins keep getting mentioned. If there is a specific reason to use twin data please be more clear about it, because I am not seeing it.

* Univariate selection: had you considered the possibility that a covariate is not univariately correlated, but would be correlated in a multivariate model?

* Line 151, Can you explain how you chose which variable to keep if two were highly correlated? If ultimately the purpose of this analysis is to let policymakers know what BE attributes lead to physical activity, it seems strange that you would just choose which attributes to drop at this point. If you are concerned with multicollinearity affecting your effect sizes, you could also look at variance inflation factors (VIF).

* Line 160: not enough to say 'did not improve predictive ability', can you give more detail on precisely what this means? You can always include ancillary findings like this as part of a supplement.

* Line 165: I am confused what you did here. Did you categorize your BE variables into quantiles and then use those quantiles as continuous variables in your models? If so, I really do not understand why you did that, please explain.

* Some of the choices in the backward elimination seem somewhat ad hoc, for example the p<=0.20 may need some more justification. It would be interesting to also see the full model fits before backward elimination to see how much some of these effects change - perhaps in a supplement.

* Given the potentially sensitive nature of these data, which include some health information as well as exact household locations, did you receive IRB clearance/exemption for this study? Assuming you did, please explicitly state this in the paper.

* To the extent possible, you should share their code and data. Especially since you have developed a new method which you ostensibly would like others to use in other scenarios, it would be good to provide other researchers with the means to do so.

* I am skeptical of the policy relevance of this approach. You will always be in danger of finding spurious correlations here. For example, you kept in youth sport facility, but we see it has
a negative effect on MVPA. What level of evidence would you need, for example to suggest to a policymaker that a youth center should be removed in order to increase MVPA?

- Have you considered doing some sort of random permutation test? Consider, for example randomly matching the outcomes of individuals with the BE of others. We would expect everything null in this case, but I am sure some covariates would still come up as significant. This would perhaps give us a better understanding of how spurious some of these results are.

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

An article of limited interest

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

Needs some language corrections before being published

**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