Reviewer's report

Title: Diabetes, Obesity, and Recommended Fruit and Vegetable Consumption in Relation to Food Environment Sub-Types: A Cross-Sectional Analysis of Behavioral Risk Factor Surveillance System, United States Census, and Food Establishment Data

Version: 2 Date: 24 January 2015

Reviewer: Colin Rehm

Reviewer's report:

In general the authors have improved their manuscript. Specifically, the methods are much more clearly described and I (and likely others) can now follow what was done. I have a number of comments that would improve the manuscript further that follow. While improvements have been made to the reporting of methods, the data are still not presented in a manner that is as clear as it could be.

Major comments requiring attention of the authors

Was k-means implemented on the number of stores or the relative percentage of stores? This is an important point to clarify, as the proportion of stores of a certain type within a neighborhood will be largely influenced by the total number of stores. This could introduce problems of interpretation similar to interpreting proportionate mortality. This issue also rears its head in Table 1-2.

Results of abstract: Should first describe the clusters identified before jumping into results related to health/behavioral outcomes. This was the first objective, but receives little attention in the abstract. P-values of difference should be presented in the abstract to support the authors use of the term “significantly”. Also, the authors should avoid using a simple dichotomous indicator of significance (e.g., * p < 0.05) and should instead present either exact p-values or a more complex set of indicators (e.g., *** < 0.001, ** 0.001 < p-value < 0.01, * 0.01 < p-value < 0.05).

Line 57-63, this is a complete non-sequitor. Please create new paragraph if you think this information is useful. I am not sure it adds much given the stated study objectives.

Line 22, “a food environment classified as “healthier options” or “good”. I would avoid “good”/”bad” terminology. Please change to “a food environment classified as having a greater proportion of “healthier options”.

Line 173-75ish, please specify what the input to the k-means procedure was: counts of food establishments vs. % of food establishments. As noted above, using proportions is dependent on the overall denominator. I would think that using the counts would be preferred given the purpose of this paper. Also, earlier
in the paper the authors note that their RFE measure is based on counts rather than proportions (as compared to the CDC RFEI measure), so I am somewhat confused.

Line 185, from Table 1 it seems that percentages are presented yet the authors describe them as means. Please be as explicit as possible in correctly labeling all quantities. Also, a t-test would be a short-hand, but sub-optimum approach at testing for differences in prevalence/percentages.

Table 1, if the values for the different food sources are percentages as indicated by footnote 1, one would expect these percentages to add up to 100, but for RFE < 1 they add up to 74.2 and for RFE >1 they add up to 92.5, respectively. Please check these numbers and note why they do not add up to 100. Also, density may not be the best term to describe this quantity. It is more accurately the proportion of food stores of that type among all food stores; density would generally refer to number of stores per area or population.

Table 1, presentation of dichotomous indicator of statistical significance is not ideal. If possible could you present the p-value with greater precision (e.g., use different symbols to indicate p<0.001 and so on).

Table 2, title. Why are percentages described as means? Please clarify this title.

Table 2, either in the text describing this table or in a footnote (or both) please indicate that the label of the cluster refers to the food establishment that is the mode. For example, the unhealthier restaurants cluster has a large proportion of fast food stores, so to only interpret based on full-service restaurants would be unwise.

Table 2, if I understand the table correctly all values in a row should sum to 100, but they do not. Please check the calculation or indicate in a footnote why they do not sum to 100.

Table 3, similar to Table 1 it would be helpful to present the p-values with greater precision. A footnote to this table should also be added that describes how this analysis accounts for difference in socio-demographics. My understanding is that these variables were accounted for in the imputation model, but please add a footnote to reduce concerns that these differences may be due to confounding.

Figure 1-2 present very interesting information but do not fit in the current paper. I do not see characterizing the spatial distribution of obesity/diabetes as one of the primary or secondary aims of this study. If the authors would like to present a map, a map of the food environment clusters would be much more informative. Similarly, the discussion of the distribution of obesity/diabetes in lines 209-219 distract from the actual study purpose. This is an area of personal interest, so I find this data interesting and compelling, but not in the context of the current paper.

Line 271, here it would helpful to refer to maps of the RFE or food environment clusters.
Line 310, “If these observations lead…”, observations from numerous studies and from different environments would be needed to develop potential policy mechanisms. The point is well taken, but the impact of this work alone is over-stated.

Line 314, pretty sure the authors mean billion. $90 million would be about $0.25 per person, which seems pretty manageable. A quick search revealed that ADA has updated these figures to include $245 billion (indirect + direct costs). Would focus on societal costs, the fact that diabetics have higher health care costs, while interesting is not nearly as compelling as the societal cost.

Line 346-347; If I recall the authors do not directly compare the prevalence of obesity/diabetes/FV5 between the low/higher RFE categories. A quick back of the envelope calculations suggests that these quantities are not different (6.3% diabetes in RFE >1 vs 6.4% diabetes in REF<=1, assuming equal population size in each strata). The authors should focus on the data presented.

Minor comments

Line 88 and throughout, please drop “this analysis” here and other places. It is clear that you are describing what you did.

Line 92-93, was the metropolitan area defined based on a priori definition, such as certain counties that are of important to governance, or were they defined by the authors. This should be included.

Line 122, confirm that obesity was defined as greater than or equal to 30 kg / m2 by using the proper symbol.

Line 130, it’s important to note that the BRFSS questions only assess frequency, not servings. Please add this clarification in this section.

Line 148-50, specify the scale of this analysis, assume block-group level, but unclear.


Discussion introduction; it would be helpful for the authors to summarize the three most interesting findings of their study at this point. From the results section many different pieces of information are presented that relate to two distinct goals, and the reader is left sorting through all of this data. Distilling the results into the most compelling findings would provide some focus to the Discussion.

Line 274, the authors use “metropolitan” and “urban” interchangeably, which is not appropriate. If the data are not stratified by population density, the findings of this study do not suggest what interventions would be mostly likely to succeed in “urban” areas. For example, if the analysis was limited to areas with a certain population density this conclusion would be justified.
Line 299-301; this alternative explanation can likely be explored with available data, specifically by incorporating a population density covariate.

Line 270, delete “aggregate”, the DC metro area is implicitly an aggregated area.

Line 305, unclear what the “limitations” refer to.

Highly Recommended Text edits

There were a number of sentences that read awkwardly or where word choice was not ideal. Please re-read carefully before re-submitting. Please note that this is a “convenience” sample of text edits and does not represent a complete effort at copy-editing.

Line 78, change “imputation” to “imputed”.

Line 79, change “data was geocoded to specific” to “data was geocoded to the specific”

Line 84, data is plural, change “study is” to “study are”

Line 106, same clause repeated twice.

Line 141, period typo following “race”.

Line 143, “include” is somewhat better than “contain” and specify “residents” over “people”.

Table 1, please add % to either the title or after each variable; e.g., Male (%)

Line 236, change “in prevalence” to “in the prevalence”.

Line 236-242, it might be more clear to call the categories “broad” rather than “large”.

Line 362, change “about” to “at the”.

Line 365, change “informative towards policies” to “informative to policy makers”.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

I declare that I have no competing interests.