Reviewer’s report

Title: Importance of Details in Food Descriptions in Estimating Population Nutrient Intake Distributions

Version: 0 Date: 15 Nov 2018

Reviewer: Geraldine Lo Siou

Reviewer’s report:

This is an interesting paper exploring the use of random forest algorithm to identify the less important characteristics of food descriptions (facets) for the purpose of improving the efficiency of data collection and handling in 24-h dietary recalls. The analyses were carefully done but for clarity (and to avoid confusion), the reviewer has the following comments:

Lines 122-123: There were two 24 HRs, however, it is not clear whether analyses were based on both 24 HRs or only one of them? What was the rationale (using both vs. choosing one 24 HRs over the other)? If both 24HRs were used, how was the two 24 HRs analyzed (i.e., average, or other methods)?

Lines 123-125: How many interviews were telephone vs. face-to-face? Were results/estimates obtained from telephone vs. face-to-face interviews comparable (i.e., both formats of interview could be used interchangeably?)? The paper seems to suggest that both 24 HRs (from a same participant) were always conducted in the same format depending on the age of the participant (i.e., adults vs. children), however please confirm this.

Lines 128-130: Was the average time (42 minutes) needed to complete one 24HR for telephone or face-to-face interviews? Was there a difference in time regarding these two formats of interview?

Lines 133-134: What was the rationale behind deciding on the 16 facets? Was it a data-driven or expert knowledge decision?

Lines 137-139: Please could you explain how you came up with 26,679 unique combinations of foods with descriptors? Was that number obtained from the 2,389 food items and the total number of descriptors from Table 1 (total number of descriptors=244)?

Lines 139-140: If 1,599 codes of 2,389 food items were linked to the consumption data, what happened to the remaining 790 (=2389-1599) food items?

Lines 143-144: The reviewer suggests adding an appendix listing the food items included in each food group.
Lines 150-151: The reviewer is confused about the meaning of "when data for that facet are permuted in the dataset (implying no predictive power for that facet in predicting the nutrient content)". Please add details or reformulate the sentence.

Lines 182-189: The reviewer is confused regarding the explanation about the NEVO code reassignment protocol. Since those detailed explanations are not included in Figure 1, the reviewer suggests adding them into Figure 1 for clarity.

Lines 191-195: It seems both 24 HRs were used at the second step of the simulation study (see "averaged over two days"). Does this mean: only one of the two 24 HRs was used in random forest algorithm, and both 24HRs were used at the second step of the simulation study? Please add details to avoid confusion. Lastly, did the authors consider using the NCI method (https://epi.grants.cancer.gov/diet/usualintakes/method.html)? If yes, please add details. If not, please discuss in the discussion section.

Lines 195-197: Were sampling weights used in the analyses in order to get population nutrient intake distributions?

Lines 201-202: The correct values should be "50 out of 111 facets".

Lines 209-210: 'Packing medium' was unimportant for 6 food groups (because of the value "0.02" (column 1 - Potatoes; Row 5 - Packing medium)).

Lines 217-218: The reviewer is not sure how the value "350,369 facet descriptors" was obtained, knowing there were 16 facets, a total of 244 descriptors and 26,679 unique combinations of foods with descriptors. Please explain.

Table 2: There should be an asterisk for value "0.02" (column 1 - Potatoes; Row 5 - Packing medium). Consequently, the "# of omitted/# of original" for column 1 (Potatoes) should be 3/8, and the "# of omitted/# of original" for row 5 (Packing medium) should be 6/9. Lastly, the "# of omitted/# of original" for column 2 (Vegetables) should be 0/5.

Discussion section: Although information collected from 24 HRs in this study was not self-reported but by telephone or face-to-face interviews, there still exists challenges related to mis-reporting or other types of measurement error. Please add this topic into the discussion section.

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.