

Reviewer 1

Thank you for the opportunity to review this manuscript. Despite the technical nature of the study, the text is highly readable, aided by the quality and clarity of the figures and tables. Although I cannot comment on the appropriateness of the modifications made to the linear regression and the choice of learning algorithms, the statistical analyses, despite their complexity, are particularly clearly reported. I have read this paper carefully, and in my opinion it is mostly very well presented, however I have a small number of suggestions that may improve the overall clarity.

Introduction

Revise the use of the term nutrient when referring to dietary/food components more generally.

Thank you for bringing this to our attention. We have revised lines 47-50 in the introduction to describe the outputs of the ASA24 and NCC database as both nutrients and food components.

“studying the link between diet and LP genetics is limited...” expand a little on why this area of research is important as context to your choice of lactose for your case study.

Thank you for this comment. We have added lines 63-67 to clarify why lactose is of interest in the context of lactase persistence genetics and provided an example for why lactose consumption by LNP individuals would be biologically interesting to study.

‘American cheese food’ – not clear - please amend or explain.

The full name for this food in NCC is “American cheese, processed cheese food” (NCC FoodID 3204), but we have shortened it to “processed American cheese” for clarity and added the NCC FoodID to line 74 for clarification. American cheese is a mixture of conventional cheese (such as cheddar or Colby) with other ingredients such as cream, milk, and emulsifiers.

Lines 78-86. This paragraph could be edited down to the most relevant 1-2 sentences.

Thank you for this suggestion. This paragraph has been edited to remove detailed descriptions of the types of record linking for conciseness (Now lines 86-88).

Lines 111-114. The novelty of the current study is explicitly stated here. However, the length of the sentence obscures the meaning – re-phrase into two sentences for clarity. The purpose of the study could more clearly be stated as an overall aim and a set of objectives.

This sentence has been rephrased as suggested by the reviewer (now lines 114-116).

Materials and methods

‘at least 0.1 cup equivalent..’ please translate to g/mls for non-US readers.

Thank you for this comment. We have added lines 140-148 and 150-152 to describe how D_TOTAL is calculated and the purpose of using the cut off of $D_TOTAL \geq 0.1$

Unfortunately, there is no way to convert the D_TOTAL value for all foods to any unit other than ‘cup equivalent’ since D_TOTAL is the sum of the total milk (D_MILK), total cheese

(D_CHEESE), total yogurt (D_YOGURT), and whey reported by the subject. D_TOTAL is sometimes greater than the sum of D_MILK, D_CHEESE, and D_YOGURT because of miscellaneous dairy components composed of whey¹. The weight of one cup equivalent for each group is set by the Food Patterns Equivalent Database (FPED) and is as follows²:

Milk: 245 g

Yogurt: 245 g

Cheese: 1-4.5 oz (28.3-127.6 g), depending on the specific type of cheese

Whey is not included in the FPED as a separate variable and does not have a weight associated with one cup equivalent.

As an example, D_TOTAL can be easily estimated from D_MILK for ‘simple’ dairy foods such as plain 2% Milk and could therefore be converted to g/mL. However, this is not possible for mixed foods (e.g. Macaroni or noodles with cheese), where D_TOTAL is comprised of multiple sources of dairy and whey. We have added the appropriate details to the manuscript (now lines 140-148 and 150-152).

The comparisons by which the machine learning and database matching are judged are stated but it would be useful if these ‘gold standard(s)’ were more clearly highlighted in one place.

We have removed the term “gold standard” from the text and simply state that the manual lookup lactose values were used to compare against values from our machine learning or database matching methods. We thank you for this comment; we think this is clearer.

Results

1 and 3.1.1: there is information here that reads as repetition of the methods.

Given the technical nature of the study, we wanted to include a short overview of how the data was generated for each results section to help guide readers in their understanding. However, we appreciate the reviewer’s comment and have shortened or eliminated the overview sentences at lines 472-473, lines 569-572, and 605-607.

2: as for item 9, but also appear to have discussion points here.

It appears as though the first part of this comment was cut off. We have edited the following sections to address the issue of having discussion points in the results section:

- Line 664-666: removed “better” and clarified the matching outputs. The sentence now reads: “For many ASA24 input foods, the matches were more similar for the Nutrient + Text algorithms than those of the Nutrient-Only algorithms, corresponding with the improved R² and MAE values.”
- Lines 679-680: removed “which is time consuming” from the end of the following sentence: “Therefore, the brand name is not apparent from the food description but can be found by inspecting the *MS* or *Responses* file.”
- Lines 681-684: Revised description of difference in lactose between New England and Manhattan style clam chowders. The section now reads: “For example, for Manhattan or New England style clam chowder, the Nutrient + Text algorithms indeed selected one type of clam chowder (Manhattan style), but it was not the type that was actually reported

(New England style). On the other hand, the Nutrient-Only matches were not related to clam chowder.”

In addition to the accuracy of the machine learning and matching methods, is it possible to also report the extent of the time-saving achieved in using these methods compared to manual look-up.

Thank you for this suggestion. We have added section 4.3 (Lines 861-871) to compare the time required for manual lookup of ASA24 foods into NDSR versus our computer-assisted methods. We estimate that on average, it took 28 minutes to look up one ASA24 food into NDSR. The actual time required for look up depended on the specific food and how many details needed to be entered into the system, or if user recipes needed to be made. In contrast, the computer-assisted methods took less than 30 minutes to execute and results for all foods are returned together, so the output can easily be reviewed. This represents a 99.7% reduction in the time needed to estimate lactose.

Discussion

A summary of the findings is useful here but repetition of the study context, aims, and methods is not necessary.

We appreciate the reviewer’s comment about repetitiveness but think that providing short summaries of the study aims and methods are useful to give context to the discussion. We have therefore shortened the study aims and methods (Lines 720-724, 764-765, 796-797).

Where integrating findings into previous research there is no need to repeat the descriptive detail of the studies reviewed in the introduction – the discussion points could therefore be much more concise.

We have removed some of the descriptions about FNDDS and LanguaL (now lines 757-760) since these details are initially described in the introduction.

Figure 4 is misplaced here and should be in the results.

Thank you for this suggestion. We have moved Figure 4 to Results where it is now Figure 3.

Review the discussion overall and remove points that refer to the study limitations to avoid repetition in the designated Limitations section.

Thank you for bringing this to our attention. We have consolidated the discussion about pre-processing the NCC files to lines 907-914. However, we feel that it is best to leave the suggestion (“future work”) to use tree or feature dropout immediately following the discussion of overfitting since these suggestions specifically address this discussion point (compared to the other suggestions in section 4.4, which are more general).

Avoid negative contractions i.e. use ‘do not’ rather than ‘don’t’.

We changed ‘don’t to ‘do not’ at line 890. We did not find any additional contractions.

Amend the conclusion so that it is a concise summary of the key points of the study, in particular removing reference to figures (and much of the second paragraph which is not needed).

We removed the first paragraph of the conclusion, which was a re-statement of the study findings, which was already summarized at the beginning of the discussion. We kept most of paragraph 2 as-is because we wanted to emphasize the utility of these methods in the two given scenarios but removed the sentence regarding the need to validate our methods for other nutrients or for full database mapping since this is already stated in the limitations.

References

1. National Cancer Institute. ASA24® Sample Analysis Files & Data Dictionaries. Available at: <https://epi.grants.cancer.gov/asa24/researcher/sample.html>. (Accessed: 22nd March 2019)
2. Bowman, S. A. *et al.* Food Patterns Equivalents Database 2007-08: Methodology and User Guide. (2013). Available at: https://www.ars.usda.gov/ARSPUserFiles/80400530/pdf/fped/FPED_1516.pdf. (Accessed: 22nd March 2019)