Reviewer’s report

Title: Patterns and correlates of physical activity in adult Norwegians: a forecasted evolution up to 2025 based on Machine Learning approach

Version: 0 Date: 23 Jan 2018

Reviewer: Charles (Chad) Heilig

Reviewer's report:

The authors have conducted an ambitious analysis of 8-15 waves of serial cross-sectional data on Norwegians' physical activity habits. The sequence of steps in the analysis is complex and occasionally difficult to follow, and in some places omits details that an astute reader would need in order to make full sense of the analysis. This reviewer uses "page;line" notation to refer to the line and page numbers in the authors' PDF.

General comments:

This manuscript poses 3 broad challenges for assessing its methods: First, although the title and much of the text emphasizes machine learning, the approach includes enough classical methods that the machine learning emphasis seems a bit off. Second, it is often difficult for this reviewer to tell whether or how the authors attended to the assumptions in their models or motivated some of their steps. As 2 examples, at 6;98, we learn that ANOVA was used with a 6-level ordinal response variable, and at 6;117, we learn that features were standardized as z-scores with no discussion about whether this is appropriate for all nominal and ordinal features. Third, where the authors use an autocorrelation model to assess secular trends across serial cross-sections, this reviewer wondered why the authors didn't choose to consider a model regressing the response against year. These are discussed further in the specific comments below.

Specific comments:

5:82-89 and Appendix A: When the 3 physical activity components are introduced, it would be good to indicate that they are analyzed over different time periods, as indicated in later exposition and in figure 2. Please consider adding information, perhaps through a supplement, on the number of surveys each year by gender; if the number contributing to each component differs, then consider including that further breakdown. The exposition could be trimmed if the information about response values were removed from this paragraph in favor of appendix A.
5;91-92: Please clarify what is meant by, "Answer options … in the independent features were deleted from the dataset." Does this indicate case-wise deletion, or does it indicate some other method for dealing with missing values? It would be useful in the results to show the extent to which missing values affected the analysis.

5;97-107: This paragraph seems to describe an analysis in which the response variables are the original 6- or 8-point ordinal scales. What did the Shapiro-Wilk and Mauchly tests indicate for the ordinal-scale inputs to these ANOVA models? Where the authors assess interaction between gender and year, is year modeled as a linear score? Since frequency was assessed over a longer period than duration and intensity, do any analyses assess the PA frequency component over the same duration (1999-2013) as the other components for comparability? Where the authors model autocorrelation, to what extent does revealed autocorrelation undermine the previous ANOVA models (which depend on conditionally independent residuals)? This reviewer would have found a regression model for secular trend across serial cross-sections to be more compelling than an autocorrelation model (assuming that the interpretation of the 3-level linear score is valid), especially where the authors seeks to project to future years. Please justify the use of an autocorrelation model in this instance.

6;114-123 and table 1: This reviewer found the notation to be confusing. In particular, both the number of respondents (variable m) and the number of features (variable h) are indexed through n, the number of surveys. If the number of features is k = |S|, then the feature vector m_i should be h_1 .. h_k rather than h_n; then matrix F_S is an n-by-k matrix. Please review the subscript ranges for h at line 114, c at line 120, and Item n on the first row of table 1. In addition, vector s_i is a concatenation of m_i and c_i. None of this notation appears later in the manuscript. This reviewer would encourage the authors to consider whether this level of detail is needed. If it is, then they should remove ambiguity in the multiple uses of n.

6;116-117: Please say why features were standardized as z-scores and how this applies to categorical variables (e.g., Q46 pension vs holidays at 22;452-456). Furthermore, how does standardizing within year affect analyses that extend over years, such as the autocorrelation analysis?

6;119: The way that responses are analyzed could be made clearer and more unified if the details were moved completely to appendix A rather than in footnote 2. Besides more clearly connecting the 6- or 8-point scales with the 0/1/2 scales, this could help readers more readily understand which analyses use which scales.

7;132: Please state how CMIM and RFECV were used to select and eliminate features. For example, was CMIM implemented with an initial threshold for the number of features to retain? Based on a statement at 6;116-117, it appears that CMIM was calculated as though all features were on a continuous scale. Where RFECV is applied using ordinal regression, does this use the
same module as the ordinal classifier (mord.LogisticAT)? Consider addressing in the discussion section: Did the authors consider using random forests to assess variable importance?

8;143: How were these 2 classifiers (ordinal regression and random forests) chosen?

8;148: Were precision and recall calculated based on exact predictive matches? For example, if the target class is 2, is a predicted class of 0 treated the same as a predicted class of 1? For a public health audience not familiar with precision and recall, consider relating these to sensitivity and predictive value positive.

9;156-157: Please clarify whether (1) these correlations are restricted to the time periods for which all 3 PA components were available and (2) they were calculated based on the original 6- or 8-point component scales. Where the authors say the results show "no relationships", it would be more correct to say something like "no monotonic relationships", or to remove the interpretation and just say that the components are not correlated.

9;158: This is more accurately stated as an interaction between gender and age group, rather than "between men and women among different age groups".

9;167: It is difficult to tell which interactions are being considered here. Figure 2 suggests either 2-way gender by survey year, stratified by age group or the 3-way interaction of gender, age group, and survey year.

11;206: The authors correctly claim that the ordinal regression model has superior performance. But they have not shown that they can claim "high validity" for the ordinal model. Please clarify.

11;210+: Where the authors use the word "significance", it is not clear if that word is warranted as compared to "importance". Based on the usual statistical interpretation of significance and on the methods section, it seems that importance is more appropriate. Furthermore, much of this text repeats information that appears in figure 3; much the repetitive information could be removed from the text. (See also the comment on figure 3 below.)

13;257-264: Following on the previous comment at 5;106, the autocorrelation model doesn't seem completely appropriate for the authors' purposes. Assuming that it is appropriate to model the 3-level ordinal response in this way, the slope (or similar parameter) of a regression model would allow one to test for a nonzero trend and to project into the future.

15;282: Where the authors claim that their model is more accurate than another published model, in what sense do they mean "accurate"? What is the basis for this claim?

15;287: The Spearman rank correlation result shows that the 3 physical activity components are uncorrelated; to conclude that they are independent requires additional assumptions.
17;342: Note that not all machine learning algorithms are "black box" algorithms that obscure the relationships among features. Is this a comment on all of CMIM, RFECV, random forests, and ordinal regression?

Table 1: See comments at 6;114-123. In the table caption, what is meant by "player's items"?

Table 2: Please clarify whether (1) these correlations are restricted to the time periods for which all 3 PA components were available and (2) they were calculated based on the original 6- or 8-point component scales.

Table 3: Consider indicating superior performance in boldface, as is conventional in machine learning literature.

Table 4: Please state in the methods and the table how the 2025 values were projected.

Figure 1: Please clarify whether these are based on the original 6- or 8-point component scales. It would also be useful to restate in the figure caption for which tests significance is indicated (e.g., Tukey's HSD).

Figure 2:

Figure 3: This reviewer was unable to read figure 3. It seems that the same message could be conveyed more effectively if these results were presented as a table, since the signed bars require a great deal of vertical space and don't show much contrast (presumably because the selected features are the top performers, hence somewhat similar to each other in performance).

Minor comments:

5;96 (and elsewhere): "Spearman-Rho's" should be "Spearman's rank correlation coefficient" or "Spearman's rho".

6;108: Please consider making Python analysis code publicly available, whether the data are made public or not.

11;225: "trials" should be "trails".

15;276: The word "wave" seems more appropriate here than the word "year".

15;297-298: The phrase "on the other hand" appears in 2 consecutive lines.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

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