Author’s response to reviews

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

Authors:

Alessio Rossi (alessio.rossi2@gmail.com)
Giovanna Calogiuri (giovanna.calogiuri@inn.no)

Version: 1 Date: 15 Apr 2018

Author’s response to reviews:

Dear Editor Natalie Pafitis,

Encouraged by your letter and the favourable tone of the Reviewer, we are submitting a revised version of the manuscript entitled “Patterns and correlates of physical activity in adult Norwegians: a forecasted evolution up to 2025 based on Machine Learning approach”.

We thank all the referees for the appreciation of our work and their insightful comments, which we feel have contributed improving the quality of our paper.

The entire text has undergone a revision of the English language and style. Moreover, prompted by a Referee’s insight, we have made some changes in the Machine Learning approach predicting future physical activity patterns in the Norwegians’ population. This led to a slight change in the results for this part of the analyses, and consequent changes in their discussion. The changes linked to specific comments of the reviewers’ are highlighted in red-coloured text, but please notice that we have not highlighted the general changes due to language proofreading.

We feel that revised version of this manuscript has improved and we are pleased about the quality that the manuscript has now achieved. We hope that you will agree. Please find attached a point-by-point response to reviewers’ concerns. We hope that you find our responses satisfactory.
TECHNICAL COMMENTS

- Please remove the title page as a supplementary file. It must be included in the main manuscript file.

Response. I included the title page into the main text

- Please amend the sub header 'Founding'

Response. Section Declaration, line 478, page 26: We have corrected the word ‘Founding’ with ‘Funding’

REVIEWERS' COMMENTS

Charles (Chad) M Heilig (Reviewer 1): 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.

Response. Thanks for your observations. In relation to your first point, the use of ‘classical’ statistics in this study served mainly to the purpose of investigating patterns and changes in the different PA components, which have then helped us making appropriated chooses on how to structure the Machine Learning approach, as well as how to best interpret and communicate the findings. We believe that the inclusion of this information is also useful to the readers, especially considering that the purpose of this study is to help researchers and public health actors to design
effective interventions to change people’s PA. Furthermore, as we explained in the manuscript, one of the purposes of this study was to propose Machine Learning techniques as a valuable statistical approach in the study of population PA (as well as other health outcomes). We understand however your concern, and revised the manuscript in a way that this concept is clearer. Specifically, while we kept the wording ‘machine Learning’ in the title, in the methods section we have explained why the classical statistics were performed. Moreover, in the discussions, we have emphasized the outcomes of the Machine Learning. We hope you can consider this as a satisfactory solution.

With regard to the other two points, please find below detailed responses.

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.

Response. Please, notice that we indicated the different time periods used in this study in the three PA components and the number of surveys filled in each year by gender in Section Methods, lines 118-120, page 6 and as: “To be noted also that PA Frequency was recorded since 1985, while PA Duration and Intensity were added starting from 1999. The response Frequency for the three PA components throughout all survey waves available are showed in Figure 1”. Moreover, Figure 1 shows a breakdown of male and female respondents (n) for each survey year – this figures show for instance that females provided higher number of survey answers compared to males for all the PA components, but the number of answers is consistent as the years go by.

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.

Response. We use a case-wise deletion approach to remove answer options such as ‘I don’t know’, ‘I’m not sure’, ‘I cannot answer’. We added this information on the text as “…from the dataset by a case-wise deletion approach” (Section Methods, line 131, page 7). Moreover, in order to explain how much this approach affected the analysis we added the following
statement: “Only 0.65% and 0.73% of the answers was deleted for females and males, respectively. Due to of the low rate of missing data answers, the analyses were not relevantly affected.” (Section Methods, lines 131-133, page 7).

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.

Response. We thank the Referee for the accurate reading of the paper and suggestion. The Shapiro-Wilks indicate that the PA components data are normally distributed, while the Mauchly test shows that the data meet the assumption of sphericity. The results of these tests were not showed in the text. To remedy this lack of information, we add the following sentence: “the assumptions of normal data distribution and Sphericity were met” (Section Method, lines 149-150, page 8).

To address your second comment, in our analysis, we considered year as a linear score.

We detected similar results for ANOVA e Autocorrelation when we performed analyses on PA frequency between 1999-2013 compared to the ones obtained from 1985 and 2013. Hence, we decided to shows the results on larger dataset for this PA component.

In order to avoid misleading results of the ANOVA due to the significant autocorrelation found in PA frequency (i.e., violation of the independence of observations assumption), we decided to perform a two-way ANOVA for repeated measures on the year factor. Based on this analysis we found that there is statistical interaction for all age groups (p<0.001). We added a description of the two-ways ANOVA for repeated measure in Section Methods, lines 155-156, page 8, as: “In order to avoid autocorrelation effects in the following ANOVA analyses, a correction on year factor was applied”. Moreover, we change the ANOVA results in Section Results. Some other minor changes is highlighted in red into the Results section.

Thank you for the suggestion about forecasting model in time series. After a careful evaluation, we decided to forecast the PA component upon 2025 by using an autoregressive model that
reflect the secular trend on PA components instead of autocorrelation model. In particular, we decided to use the autoregressive model because this process forecast future values based on a weighted sum of past values. An autoregressive process operates under the premise that past values have an effect on current values. We added the description of this approach in Section Methods, lines 205-214, pages 10-11, as: “In order to forecast the PA components upon 2025, an autoregressive model [23] was created, which predicted the evolution of the three different PA components based on a weighted sum of past values reflecting the secular trend. More specifically, for both genders, a time series of the means of the different PA components in each survey wave was created. For each time series, the model was trained on n-3 elements of the time series (i.e., data recorded until 2007) and tested on the remaining 3 survey waves (i.e., 2009, 2011 and 2013 surveys). The accuracy of our model was assessed by computing the mean squared error (MSE) between the observed and predicted values. To test the stationarity null hypothesis of autoregressive model, an F-test was performed. Finally the model created was used to forecast the future changes for each of the three PA components separately for both genders and the different age groups”. By using this approach, the forecasting results change. We provided the new results as: “Table 5 shows the outcomes of the autoregressive model predicting future changes of the three PA components up to 2025 based on the secular trend detected upon 2013. According with this model, men and women in all the age groups are expected to increase their PA Frequency of about 14.65±6.11% (MSE=0.004±0.001; F(1,6)≥3.6e+6, p<0.001) and 41.10±23.12% (MSE: 0.007±0.002; F(1,6)≥3.6e+9, p<0.001) from 2013 to 2025, respectively. All the age groups in both males and females increased their PA Frequency from at least one time every two weeks (<1.02 arbitrary unit) in 1985 to one time a week in 2013 (1.10-1.27 arbitrary unit), and will reach in mean more than one time a week in 2025 (1.22-2.07 arbitrary unit). The model predicts also an increment in PA Intensity of about 7.98±30.10% (MSE=0.004±0.002; F(1,6)≥3.6e+12, p<0.001) and 6.19±29.46% (MSE=0.003±0.003; F(1,6)≥3.6e+5, p<0.001) from 2013 to 2025, in men and women respectively. To the contrary, the model predicts a reduction of PA Duration of about 14.78±14.16% (MSE=0.005±0.002; F(1,6)≥3.6e+8, p<0.001) and 1.89±5.08% (MSE=0.007±0.003; F(1,6)≥3.6e+7, p<0.001), in men and women respectively.” (Section Results, lines 301-317, pages 15-16). Some other minor changes is highlighted in red into the Results and Discussion sections.

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.
Response. Thank you for the accurate revision. We have modified this part in order to avoid 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?

Response. This is a misprint from previous version of the paper. We did not standardized the categorical data in z-score. We deleted this sentence from the text.

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.

Response. In accordance with your suggestion, we moved the information of footnote in Appendix A to be clearer about the survey answer subdivision.

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?

Response. We state that we did not set an initial threshold for the number of features to retain when we perform a CMIM, but we select the subset of features that minimizing the entropy function $\widehat{H}(y|h_1,\ldots,h_k)$. Moreover, we state that RFECV was performed using the same Ordinal classifier module used in the prediction step. This information are added in the Method Section, line 182-184, page 9.

We decided to compute the importance of the features by using Ordinal regression due to the higher performance recorded by this classifier compared to Random Forest (Section Results, lines 254-255, page 13).
How were these 2 classifiers (ordinal regression and random forests) chosen?

Response. Several other classifiers could be used for this multiclass classification problem. In analysis process, we tried some other algorithms, but we decided to show the Ordinal and Random Forest because this provided the best classification results. Moreover, the choice of these algorithms is also due to their characteristics: Ordinal is a type of regression analysis used for predicting an ordinal variable, while Random Forest is an ensemble learning method for classification and regression that operate by constructing a multitude of decision trees outputting the class that is the mode of the classes of the individual trees.

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.

Response. We defined precision and recall in accordance with exact predictive match. We decided not to insert sensitivity an predictive values positive because we think that could be difficult concepts to explain to public health audience that are usually unfamiliar with predictive performance evaluation. We hence add additionally explanation of precision and recall in Section Methods, line 195-197, page 10.

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.

Response. Thank you for the suggestion. We modify this sentence as: “Table 2 shows the findings of the correlation analysis among the different PA components from 1999 to 2013, showing that weak positive linear relationships among the independent features. Even though this relationship was statistically significant, the small correlation coefficient suggests that these components are largely independent from each-other” (Section Results, lines 217-220, page 12).

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

Response. Thank you for noticing this. We modified the sentence at Section Results, line 230, page 12.
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.

Response. To better clarify what this sentence was referred to, we modify the results in Results section, lines 241-248, page 13.

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.

Response. This is a typo. We mean ‘Accuracy’. However, we deleted this sentence during the revision process. To a clearly explanation about our assertion, we added more details about our results in the Results section.

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.)

Response. Thanks for your suggestion. You are right. We mean importance. We changed all the word ‘significance’.

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.

Response. As showed in the answer at the question 5;106 we use a autoregressive function to predict future evolution on PA components.

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?

Response. Thank you for the comment. We decided to delete this phrase because could be misleading. However, we added some details about the advantage of using ML approach
compared to ‘traditional’ statistics in ‘Strengths and limitations of the study’ (Discussion section, lines 416-438, pages 23-24).

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.

Response. Thank you for this comment. To better explain our correlation results we modified this sentence as: “Table 2 shows the findings of the correlation analysis among the different PA components from 1999 to 2013, showing that weak positive linear relationships among the independent features. Even though this relationship was statistically significant, the small correlation coefficient suggests that these components are largely independent from each-other.” (Section Results, lines 217-220, page 12).

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?

Response. Yes, it is referred to our specific approach.

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

Response. This was a typo. We have now removed the wording ‘player’s item’ from Table 1 caption.

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.

Response. We modified the Table 2 caption in accordance with your comment.

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

Response. We modify the Table 3 in accordance with your comment.
Table 4: Please state in the methods and the table how the 2025 values were projected.
Response. Please, refer to our response to comment 5:97-107.

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).
Response. We stated this information in the Figure caption.

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).
Response. Thank you for the suggestion. We deleted that figure and added a new table (Table 4) providing this information.

Minor comments:
5:96 (and elsewhere): "Spearman-Rho's" should be "Spearman's rank correlation coefficient" or "Spearman's rho".
Response. Thanks for this suggestion. We change ‘Spearman-Rho’s’ to ‘Spearman's rank correlation coefficient’.

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

11:225: "trials" should be "trails".
Response. Thanks for noticing this. We corrected this typo.
15;276: The word "wave" seems more appropriate here than the word "year".
Response. Thanks for suggestion. We changed “year” into “wave” throughout the paper.

15;297-298: The phrase "on the other hand" appears in 2 consecutive lines.
Response. Thanks for noticing this. We have rephrased these sentences.

Steven Cohen (Reviewer 2): This is an interesting and well-presented manuscript using innovative statistical methods to address a major public health issue facing many countries today. That said, there are a number of elements that could be fixed or improved upon in a revision. I also recommend that the manuscript be proofread by a native English speaker to improve the numerous, but extremely minor grammatical and syntactic issues throughout the manuscript. Specific elements that could be fixed are listed by section here:

Abstract

1. In the introduction of the abstract, it would be helpful to have a sentence motivating/justifying the sex-stratified analysis.
Response. Thanks for your suggestion. It was difficult however to include such a sentence, while staying within the word-count limit of 350 words, especially considering that if we include a sentence to justify stratifying the analyses by gender, we should then also have provided a similar sentence regarding the stratification by age group. Thus, in order to address this comment, we have made the following changes:

1. In the introduction of the abstract, we added the following sentence: “In order to effectively increase PA levels in the population, in depth understanding of PA habits in different subgroups is vital” (Abstract section, lines 5-6, page 1). This should provide some explanation of why we chose to stratify the analyses by gender and age group.

2. In the Results section, we added the wording “In line with previous literature, …” (Abstract section, line 17, page 1). This should hint to the fact that different PA patterns between men/women and among different age-groups are known in the literature.

Notice also that in the Methods section of the main text, the following sentence has been added to explain why the analyses were stratified by gender and age group: “Because of known
differences of PA patterns between genders and among age groups [6], all analyses were
stratified by these variables. Age, which was available in the Norsk Monitor survey as a
continuous variable, was categorized in four groups: <25 years old (yo), 25-44 yo, 45-64 yo, and
≥65 yo” (Methods section, lines 125-128, page 7).

2. The authors could probably delete "using a machine learning approach" in the conclusion
subsection.

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we
had the entire text undergo a revision of the English language and style.

Background

3. It would be helpful to have a sense of what percentage of Norwegians (or other relevant
gographic area) meet the current WHO requirements.

Response. This information is now more explicitly reported

4. (minor) Paragraph 2: Should be "work" not "works" in the second line.

Response. Thanks for suggestion. We modify this typo.

5. I would suggest rewriting some of the second paragraph of the background section, in
general. For example, the message of the two parts of the second sentence (beginning with
"National surveys…" are not linked easily. I would recommend splitting this sentence into
two smaller sentences. Also, the last sentence is difficult to interpret. Some overall
numbers would be helpful here (see earlier comment).

Response. This paragraph has been rephrased and re-structured.

6. Third paragraph: Rewrite first sentence. It is not grammatically correct.

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we
had the entire text undergo a revision of the English language and style
7. This sentence should be edited: "As a matter of fact, thanks to the huge quantity of data that are now recordable by survey and technologies, Machine Learning approach could help to discover disease risk factors related to lifestyle or the environment helping people to understand and then modify their wrong behaviours." In addition to some minor grammatical issues, the sentence seems to be a bit biased, colloquial, and subjectively written. Consider substituting for words such as "huge" and "wrong".

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.

8. Next sentence: Delete "actually".

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.

9. Machine Learning should be defined clearly somewhere in the beginning of this paragraph.

Response. Thank you for the suggestion. We add a Machine Learning definition in Section Background, lines 74-77, page 4.

10. Relatedly, please provide some additional references for the use of Machine Learning in previous studies. Currently, it appears as if one is referenced (ref 8).

Response. We have now added seven more references for previous Machine Learning studies (see ref n. 11-12-13-20-21-22-23)

11. The approach of using those specific measures (frequency, duration, and intensity) should be better justified previously in the Background section.

Response. We have added a (thorough) background for these measures trying to emphasize how these reflected the definitions and recommendations provided by the WHO and the Norwegian Directorate of Health.

12. Last paragraph: Delete or reword "thanks to this approach".

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.
Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.

13. Last paragraph: Consider specifying what "deeply investigated" means.

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.

14. (minor) "Physical activities" should be abbreviated as PA (see last sentence).

Response. Thanks for suggestion. We made appropriated corrections throughout the paper.

Methods

15. For the statement "Answer options such as 'I don't know', 'I'm not sure', 'I cannot answer', in the independent features were deleted from the dataset." What percent were deleted? Were they consistent across cases? In other words, were respondents who answered in this way to one question more, less, or similarly likely to respond this way to other questions? This information would be extremely important to have to assess bias.

Response. “Only 0.65% and 0.73% of the answers was deleted for females and males, respectively. Due to of the low rate of missing data answers, the analyses were not relevantly affected” (Section Methods, lines 131-133, page 7).

16. (minor) Consider rewording "First of all" in the Statistical Analysis paragraph.

Response. Thanks for suggestion. We addressed this comment in the text. Please note that we had the entire text undergo a revision of the English language and style.

17. The methods described under the subheading "Experiments" should have appropriate references.

Response. Thanks for the suggestion. We have provided the specific reference for both Preliminary Analyses and Machine Learning approach.
18. What software was used to conduct this extensive, computationally-intensive analysis?

Response. We have performed all the analysis by using Python 2.6. The library used was provided in the footnotes. Please, note that this information is provided in the Methods section.

19. The explanation of the "feature" and "label" values found in the last paragraph of the methods section is difficult to follow. Some additional explanation/more basic terminology would be helpful for interpretability.

Response. Thank you for the suggestion. We modified this section in accordance with your comment.

20. More information on the predictive models would be helpful in the last paragraph.

Response. We provided more information about our predictive model: “In order to forecast the PA components upon 2025, an autoregressive model [23] was created, which predicted the evolution of the three different PA components based on a weighted sum of past values reflecting the secular trend. More specifically, for both genders, a time series of the means of the different PA components in each survey wave was created. For each time series, the model was trained on n-3 elements of the time series (i.e., data recorded until 2007) and tested on the remaining 3 survey waves (i.e., 2009, 2011 and 2013 surveys). The accuracy of our model was assessed by computing the mean squared error (MSE) between the observed and predicted values. To test the stationarity null hypothesis of autoregressive model, an F-test was performed. Finally the model created was used to forecast the future changes for each of the three PA components separately for both genders and the different age groups.” (Section Methods, lines 205-214, pages 10-11).

Results

21. In general, there is so much information contained in this section that it is hard to follow how the results flow logically from the study aims. Consider using subheadings to better bridge the aims with the results.

Response. In accordance with your suggestion, we added three subheading: ‘Physical activity patterns from 1985 to 2013’, ‘Correlates of the physical activity changes in 1985-2013’ (this in turn contain three sub-heading: ‘Correlates of PA frequency’, ‘… duration’, ‘… intensity’), and ‘Future physical activity patterns’.
22. (minor) The first paragraph is one standalone sentence.

Response. We change these sentence as: “Table 2 shows the findings of the correlation analysis among the different PA components from 1999 to 2013, showing that weak positive linear relationships among the independent features. Even though this relationship was statistically significant, the small correlation coefficient suggests that these components are largely independent from each-other” (Section Results, lines 217-220, page 12).

23. I recommend condensing some of the text in the descriptive results (first two paragraphs).

Response. We have now condensed the text in the results section, emphasizing the most important

24. (minor) Add the phrase "statistically significant" in the first sentence of the results section.

Response. Thanks for suggestion. We added this wording in accordance with your comment.

25. Second paragraph: Consider rewording "males reduce". This sounds causal and should not be.

Response. Thanks for suggestion. We changed ‘males reduce’ with ‘males have reduced’.

26. The results in the third paragraph should be stated more clearly and succinctly. I recommend listing the significant interactions and providing a brief interpretation of one of them.

Response. Please, refer to our response to comment #23

27. For the PA Frequency, Duration, and Intensity paragraphs that follow and reference Table 3, the presentation of the specific results can be confusing and hard to see where the findings are listed in the table. Numbers representing the associations observed should be cited in the text to support the explanation.

Response. In accordance with the reviewer 1, we indicated superior performance in boldface in order to highlight the classifiers with best performance. We also decided not to provide numbers
in the text avoid redundancies. However, please note that we make explicit reference to this table in the text to address the reader to where they can find this information.

28. Although I recommend condensing and clarifying much of the results section to better focus on the key results (and relevant interpretations of those results), I recommend discussing the predictive models presented at the very end of the results section in more detail.

Response. Thanks for suggestion. We have now condensed the text in the results section, emphasizing the most important. We have also provided a specific subsection at the end of the Results section to show detailed results of the autoregressive model.

29. Relatedly, are the confidence bounds or any other measure of uncertainty on these estimates for Table 4 (predictive model)?

Response. We assessed the accuracy (i.e., measure of uncertainty) of our autoregressive model to predict PA components in the following years by Mean Squared Error (MSE). In particular, this value reflects the quality of an estimator. We described in the text the accuracy of our models as: “The model predicts also an increment in PA Intensity of about 7.98±30.10% (MSE=0.004±0.002; F(1,6)≥3.6e+12, p<0.001) and 6.19±29.46% (MSE=0.003±0.003; F(1,6)≥3.6e+5, p<0.001) from 2013 to 2025, in men and women respectively. To the contrary, the model predicts a reduction of PA Duration of about 14.78±14.16% (MSE=0.005±0.002; F(1,6)≥3.6e+8, p<0.001) and 1.89±5.08% (MSE=0.007±0.003; F(1,6)≥3.6e+7, p<0.001), in men and women respectively.

The low MSA values (i.e., close to zero) and the statistically significant results found in F-test regression analysis indicate that the forecasted evolution of the three PA components predicted by our autoregressive models is highly accurate for both gender in all the age groups” (Section Result, lines 309-317, page 16).

30. (minor) Delete the use of "like" in the presentation of the results.

Response. We have deleted the word “like”, as you recommended.
Discussion

31. Third paragraph: What are potential explanations for the increase in PA levels (slightly contradicts some of the arguments in the Background section). Were there events/historical changes that could account for these trends?

Response. As this study is based on an observational design, it is difficult to fully understand what are the causes of these predicted changes. The Machine Learning analysis has identified a series of factors (environmental as well as in beliefs and values) associated with the evolution of the PA patterns throughout the past 30 years. These factors have been changing alongside (and maybe explain) the Norwegians’ PA habits, and are likely to keep changing with them in the future.

32. The fourth paragraph that describes the machine learning approach and the findings contained therein would benefit from a discussion of some public health approaches and theories that might explain or corroborate the specific findings of social determinants and PA.

Response. It was our intention, in this section, to make such discussion (notice for example that we mentioned the socio-ecological model of PA advocated by Sallis et al, 2006, and Nelson et al., 2008). We realized however that this intent was not explicit enough, so worked to improve it. We have revised this part, making more explicit reference to the socio-ecological model as well as emphasizing the impact of environmental and social factors on the population’s PA.

33. I would also recommend providing a bit more detail about Machine Learning and its contributions to other public health issues in Norway or elsewhere. Some limitations are listed in the limitations paragraph, but since this is a fairly new approach, it would be helpful to have more detail.

Response. We have now extended the discussion of the Machine learning model both in terms of its accuracy and in terms of its potential in Public health research. This information has been provided in different sections of the discussion, in particular under the section ‘Strengths and limitations’ (line 416, page 23) and in a new section entitled ‘Machine learning approaches and its contribution to Public Health’ (line 390, page 22) added at the end of the discussions.

Kind Regards, on behalf of all authors