Reviewer’s report

Title: Application of three statistical models for predicting the risk of diabetes

Version: 0 Date: 24 Apr 2019

Reviewer: Nicole Hobbs

Reviewer's report:

With high rates of undiagnosed diabetes worldwide and the associated complications of uncontrolled diabetes, an accurate method to at-risk individuals is valuable. This manuscript describes three models used to predict the occurrence of type 2 diabetes and risk factors identified to be significant are discussed. The developed neural network model outperformed the decision-tree and logistic regression model in the reported performance metrics.

Abstract

1. Please revise the background section to state the context or purpose of the study. Currently, this section provides the same information as the methods section. Please explain the rates of undiagnosed diabetes, the risks of untreated diabetes, or something along those ideas.

2. Please give the full words prior to abbreviations (ROC, BP, TG, TC).

Background

3. Please describe the rates of undiagnosed diabetes and the risks associated with leaving diabetes untreated in this section.

4. Page 1, Line 46: World is misspelled.

5. Page 1, Line 49 and entire manuscript: Please use the terminology "people with diabetes" rather than "patients with diabetes"

6. Page 1, Lines 56, 61: Please replace "it is easy to" with "it can."

7. Please change the phrasing of your discussion of diabetes related complications and death. These outcomes are not inevitable for all people with diabetes and occur due to poor glycemic control and genetics.

8. Page 2, Line 4: Use the term "Prevention-based" rather than "Precaution-based"
9. Page 2, Line 9: Give statistics on the number of people with pre-diabetes that respond to treatments and can reverse their diagnosis.

10. Lifestyle factors are only a factor in a subset of people diagnosed with type 2 diabetes, so please be careful to not place blame people with diabetes for their diagnosis. You may find the following paper to be an interesting view of this, "Ahlqvist E et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018 May 1;6(5):361-9."

11. Page 2, Line 10: Start a new paragraph at "In recent years, …" and add citations for the claims made in this new paragraph.

12. Page 2, Line 14: Remove the phrase "the main methods of data mining technology are"

13. Page 2, Line 33: The claim is made, "the established statistical prediction model is not suitable for China." Please explain this, and if applicable, give appropriate references.

14. Please explain and reference the existing literature on using data mining techniques to predict the occurrence or risk of type 2 diabetes. Please explain how your work surpasses their contributions to the area. A few examples of such to reference include:


Methods

15. Please change the term, "Objects" in your heading.

16. Page 2, Line 52: Please explain what it means for a case to be valid and why ~500 cases were deemed invalid.

17. Page 3, Lines 11-12: Please clarify the sentences on height and weight measurement. Were 2 m and 150 kg the maximum measurable values?

18. Page 3, Line 16: How was the BMI cut off of 28kg/m2 selected? The WHO guidelines for these values are &gt;25 is overweight and &gt;30 is obese.

19. Please define metric for abdominal obesity.
20. Please describe what portions of the data are used for training and testing your models as well as fitting any hyperparameters of the models. Was a technique such as leave one out cross-validation or bootstrapping used? If not, please reassess the performance.

Results

21. Please give the full word prior to the abbreviation (OR)

22. Please describe the meaning of the values given for the sensitivity and specificity of the ROC curves. Are those the average values? Sensitivity and specificity are the x and y-axes of the curve.

23. In the model comparison, please compare to existing models of predicting the risk of T2D.

24. Performance measures such as Bayesian information criterion or Akaike information criterion should be added to table 8 to assess the benefits of increased model complexity.

Discussion

25. Do the values reported based upon your study support other large questionnaires in China?

26. In figure 3, the x-axis should be "100-specificity".

27. Page 10, Line 50: Please state that there was a higher percentage of men that were classified as smokers and alcohol consumers rather than stating men prefer to drink and smoke more.

28. Page 10, Line 51: Was there any portion of your survey related to diet? Can you reasonably conclude that meat in their diets was the factor leading to higher TG?

29. Please remove the sentence, "For men, you can increase the control of diabetes diet control, reasonably consume meat, control tobacco and limit alcohol, and at the same time participate in exercise to control waist circumference." Please be careful to only make statements which are directly supported by your results.

30. Page 11: Please discuss limitations of neural network models such as overfitting.

31. Page 12, Line 30: Please rephrase to state "rapid and effective prediction of the risk of type 2 diabetes can allow for preventative actions to be taken by members of high-risk groups."
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 recommend additional statistical review

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 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.