Author’s response to reviews

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

Authors:
Liu Siyu (736200685@qq.com)
Yue Gao (yueg17@mails.jlu.edu.cn)
Yuhang Shen (shenyh17@mails.jlu.edu.cn)
Min Zhang (minzhang18@mails.jlu.edu.cn)
Jingjing Li (jjli18@mails.jlu.edu.cn)
pinghui sun (sph621102@163.com)

Version: 1 Date: 18 Jul 2019

Author’s response to reviews:

Dear Reviewers,

Thank you very much for your efforts to review our manuscript, and especially for your deep insight, detailed comments and explanations. We tried our best to answer your questions and made revisions clearly. Please let us know if you have any more comments or suggestions.

Response to Reviewer 1

Reviewer’s comment:
1. The rationale for this study was not exposed clearly enough. Many previous studies with large sample reported risk models for screening undiagnosed diabetes in China population, such as Gao, W.G. et. al. (2010), Li, W., et al. (2018), and Wang, A., et al. (2016) etc. The same population (China) was studied in this study, but what gap between previous studies and clinical practice would be bridged wasn't illustrated clearly. This point is crucial and should be documented while considering publication in a medical journal.

Our response:

Thank you for your careful reading of our manuscript and sorry for the confusion. We have added the advantages of this study and the differences from previous studies in the last paragraph of background (Shown in line 8-19, page 3).
Reviewer’s comment:

2. Only one dataset was used. The validation of modeling is absent.
Our response:
In this study, 70% of subjects randomly selected to provide a training data set for constructing models and 30% of subjects were selected to provide a validation data set for comparing the three models. We did not provide this part of the description at the time of writing and have added to ‘Statistical Analysis’ now (Shown in line 27-29, page 4).

Reviewer’s comment:

3. In Abstract section, please identify abbreviations or acronyms should be explained when first used. (such as ROC, OR, TG, TC, AUC, and BP). In results of abstract, the description of risk factors should be TG≥1.17mmol/L and age≥70 years, instead of TG and age.
Our response:
Thank you for this comment. We have explained all acronyms when first used in this essay, and modify the description of risk factors. (Shown in line 28-37, page 1).

Reviewer’s comment:

4. In Background section, please indicate what gap between previous studies and clinical practice would be bridged by this study.
Our response:
We have explained the modify results of the revision of this article in the innovation point.(Shown in line 8-19, page 3).

Reviewer's comment:

5. The outcome definition is not clear in terms of what kind of diabetes patients was predicted, because the exclusion criteria of study subjects did not mention. In general, the patients with known diabetes should be excluded in a prediction study for screening undiagnosed patients or for incident cases. Providing a flowchart of study subject's recruitment in Objects and Methods section is suggested.
Our response:
We have provided the exclusion criteria for the study in the text. (Shown in line 29-30, page 3).

Reviewer’s comment:

6. Please identify the year of study and the study sample drawn by using probability or non-probability sampling in Objects and Methods section. Furthermore, the definition of dyslipidemia by only using laboratory tests is not appropriate. The dyslipidemia history or dyslipidemia medication should be considered in this study.
Our response:
Thank you for this comment. We have identified the year of study. This study was conducted in 2018 (Shown in line 22, page 3). We may not have explained clearly in the text that the multi-stage hierarchical cluster sampling method is a type of probability sampling, and the article uses probability sampling. We have modified the screening indicators for people with dyslipidemia through reviewers' opinions (Shown in line 18-19, page 4).

Reviewer's comment:

7. In the analysis of logistic regression model, all lipid markers (TG, TC, LDL-C, HDL-C) were considered synchronously in one mode. Collinearity of these markers may happen because of high correlation among them.

Our response:
Thank you for this comment. We added the content of the collinearity test to the Method section and performed a collinearity test before the logistic regression. No collinearity was found between the included variables. (Shown in line 30-34, page 4)

Reviewer's comment:

8. The discussion of risk factors (the second to fifth paragraphs in Discussion section) is not appropriate. Authors used the results of bivariate analysis instead of multivariate analysis. Gender, level of LDL-C, smoking and stroke were not an independent risk factor according to the results of multivariate logistic regression model. It is not correct to report those factors were associated with diabetes.

Our response:
Thank you for this comment. We chose binary logistic regression as a multivariate analysis method because the dependent variable in the study is type 2 diabetes, meaning the dependent variable is a binary variable. A multivariate dependent variable, such as the severity of diabetes, was not used. Based on the reviewer's recommendation, in the discussion of influential factors, only the factors that were significant in the multi-factor analysis were included.

Reviewer’s comment:

9
1) Please identify the first letter of variable name in upper case in all tables.
2) It would be helpful to provide readers with the marginal distribution of all independent variables in Table 1, 2, and 3.
3) In table 3, the cut-off point of TC, TG, LDL-C, and LDL-C should be provided.
4) Figure 1 provided the same information which was shown in the last column of table 7. Please delete figure to avoid redundant expression.

Our response:
Thank you for this comment. We have identified the first letter of variable name in upper case in all tables. The reason we did not mention the marginal distribution of the independent variables in the table is that the marginal distribution of all independent variables was proposed in the “method”. We have delete figure 1 to avoid redundant expression.

Reviewer’s comment:
10. Please consult a Native American to let the readers to easy reading. Please do English revision.
Our response:
This manuscript was revised for priper English language, grammar, punctuation, spelling, and overall style by native English speaking editors.

Response to Reviewer 2

Reviewer’s comment:
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.
Our response:
Thank you for this comment. The background section has been revised to explain the incidence of undiagnosed diabetes, the risk of untreated diabetes, and other topics (Shown in line 21-25, page 1).

Reviewer’s comment:
2. Please give the full words prior to abbreviations (ROC, BP, TG, TC).
Our response:
According to the reviewer's recommendation, the complete word was added before the first capitalization, similar to below.

Reviewer’s comment:
3. Please describe the rates of undiagnosed diabetes and the risks associated with leaving diabetes untreated in this section.
Our response:
The incidence of undiagnosed diabetes in this study is discussed in 'background' and 'results'.

Reviewer’s comment:
4. Page 1, Line 46: World is misspelled.
Our response:
We have corrected the spelling mistakes.

Reviewer's comment:
5. Page 1, Line 49 and entire manuscript: Please use the terminology "people with diabetes" rather than "patients with diabetes".
Our response:
According to the reviewer's recommendation, we used the terminology "people with diabetes" rather than“patients with diabetes”, similar to below.
Reviewer’s comment:
6. Page 1, Lines 56, 61: Please replace "it is easy to" with "it can."
Our response:
Based on your comments, we have corrected the wording.

Reviewer’s comment:
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.
Our response:
According to your recommendation, the statement about the complications of diabetes has been changed from “inevitable” to “early detection can be avoided through treatment” (Shown in line 14-17, page 2).

Reviewer’s comment:
Our response:
Thank the reviewer for the kind suggestions. We have modified the default word.

Reviewer’s comment:
9. Page 2, Line 9: Give statistics on the number of people with pre-diabetes that respond to treatments and can reverse their diagnosis.
Our response:
Thank you for this comment. We have added this section to the essay. We summarized the number of people with pre-diabetes that respond to treatments and can reverse their diagnosis through consulting literature materials. The reason why we can't get results from our study is that our study is a cross-sectional study. It is not possible to follow up patients to see if they can improve their condition through treatment. Secondly, we have not found an indicator that can Identify whether a patient can respond to treatment or improve.

Reviewer’s comment:
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."
Our response:
Thank you for this comment. We have changed expression from "by intervening in behaviors and lifestyles of high-risk groups" to “by targeting intervention in high-risk groups” (Shown in line 23, page 2).

Reviewer’s comment:
11. Page 2, Line 10: Start a new paragraph at "In recent years, …" and add citations for the claims made in this new paragraph.
Our response:
Thank you for this comment. We started a new paragraph here and added supporting references. (Shown in line 31-40, page 2).

Reviewer’s comment:
12. Page 2, Line 14: Remove the phrase "the main methods of data mining technology are".
Our response:
We have removed the phrase “The main methods of data mining technology are Logistic Regression (LR), BP Neural Network (BPNN), and Decision Tree (DT)”.

Reviewer’s comment:
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.
Our response:
We have added the advantages of this study and modified statement in this section (Shown in line 8-19, page 3).

Reviewer’s comment:
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:
Our response:
Thank you for this comment. We have added the advantages in last paragraph of background (Shown in line 8-19, page 3)

Reviewer’s comment:
15. Please change the term, "Objects" in your heading.
Our response:
Thank you for this comment. We have changed “objects” to “participants” (Shown in line 20-21, page 3).

Reviewer’s comment:
16. Page 2, Line 52: Please explain what it means for a case to be valid and why ~500 cases were deemed invalid.
Our response:
We explain here why there are approximately 500 invalid cases. Because we excluded a number of participants based on the exclusion criteria. (Shown in line 30-31, page 3)
Reviewer’s comment:
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?
Our response:
Thank you for this comment. The height measurement tool used in this study has a range of 2 m, and the weight measurement tool has a range of 150 kg. The results show that there was no case outside the range.

Reviewer’s comment:
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.
Our response:
Thank you for this comment. The WHO defines BMI ≥ 25 kg/m2 as overweight and BMI ≥ 30 kg/m2 as obese. However, considering ethnicity, WHO defines BMI ≥ 24 kg/m2 as overweight in the Asia-Pacific population. BMI ≥ 28 kg/m2 is defined as obesity. Specific literature has been added to the manuscript.

Reviewer’s comment:
19. Please define metric for abdominal obesity.
Our response:
Thank you for your suggestion that a man's waist circumference &gt;94 cm and woman’s waist circumference &gt;80 cm are the standards for obesity, but these standards are more suitable for European populations. For Asia-Pacific populations, a man's waist circumference &gt;90 cm and woman’s waist circumference &gt;80 cm are recommended as the standards for obesity. The background literature has been added to the manuscript.

Reviewer's comment:
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.
Our response:
Thank you for this comment. In this study, 70% of subjects randomly selected to provide a training data set for constructing models and 30% of subjects were selected to provide a validation data set for comparing the three models (Shown in line 16-18, page 4). We used cross validation to verify the model (Shown in line 27-29, page 4).

Reviewer’s comment:
21. Please give the full word prior to the abbreviation (OR)
Our response:
Thank you for your suggestion. We have given the full word prior to OR (Shown in line 32, page 1).

Reviewer’s comment:
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.
Our response: Thank you for this comment. We explain the horizontal and vertical axis in the text (Shown in line 8-13, page 10).

Reviewer’s comment: 23. In the model comparison, please compare to existing models of predicting the risk of T2D.
Our response: After searching a large amount of literature, we found that the existing prediction models for type 2 diabetes were mostly logistic and BP neural networks, which are also useful for decision tree research. The purpose of our research is to compare the current models. When more than one model is applied to the same data, which one has higher prediction accuracy and is more suitable for predicting type 2 diabetes? We have considered comparisons with models in other literatures, but the results are meaningless because: 1 the variables selected are different. 2. We are different from the sample population in other studies. No other models have been introduced. However, our future research will explore other models.

Reviewer's comment: 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.
Our response: We have add Akaike information criterion (AIC) to table 9. The results show that BP neural network is the best model.

Reviewer’s comment: 25. Do the values reported based upon your study support other large questionnaires in China?
Our response: The questionnaire we used in the study is a unified questionnaire for China, which can support other large questionnaires in China.

Reviewer’s comment: 26. In figure 3, the x-axis should be "100-specificity".
Our response: Thank you for this comment. We have modified the wrong wording in Figure 2.

Reviewer’s comment: 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.
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.
Our response: Thank you for this common. The section related to gender has been deleted cause gender were not an independent risk factor according to the results of multivariate logistic regression model.
Reviewer’s comment:
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?
Our response:
Thank you for this comment. The section related to TG in the diet has been deleted.

Reviewer’s comment:
30. Page 11: Please discuss limitations of neural network models such as overfitting.
Our response:
Thank you for this comment. We have discussed limitation of neural network models (Shown in line 41-53, page 11).

Reviewer’s comment:
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."
Our response:
Thank you for this comment. We have rephased 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." (Shown in line 21-22, page 12).