Reviewer’s report

Title: Development and Validation of Algorithms to Classify Type 1 and 2 Diabetes According to Age at Diagnosis Using Electronic Health Records

Version: 0 Date: 02 Dec 2019

Reviewer: Kenney Ng

Reviewer's report:

The manuscript presents and evaluates a set of T1D vs T2D classification algorithms on a large registry data set of diabetic patients from Hong Kong to quantify performance as a function of age of diabetes diagnosis.

The development of automated methods to accurately classify diabetes diagnosis is an important problem with the increasing use of electronic health record data.

Overall, the manuscript is well written and easy to understand.

However, there are several issues that need to be addressed to improve the manuscript:

1) One major concern is that the authors do not seem to be properly using the separate derivation and validation cohort data sets to avoid over-fitting. All algorithm hyper-parameter tuning and model selection should be performed on the derivation data set (including the rule combinations and the age specific analysis). Once the final algorithm is selected (based on the derivation cohort), it should then be applied to the validation cohort. There shouldn't really be dozens of experiments performed on the validation cohort (Table 2) - there should just be one using the final model. Otherwise, you are using the validation data to select the methods that work best on the validation data - thereby defeating the original purpose of the validation data set. The Table 2 analyses should really be done on the derivation cohort, not the validation cohort.

2) A second major concern is related to comparisons of the approach presented in the manuscript to other published T1D vs T2D classification algorithms.

2a) The performance comparisons can be significantly improved by implementing and running the alternative algorithms on the author's derivation and validation data sets and then comparing the performance metrics on the same data.

2b) It is not obvious why "algorithms developed for European populations cannot be directly applied to Asians" - especially since the classification rules seem to be straightforward and based on similar information (diagnosis codes and drug prescriptions). Can the authors clarify in more detail (or show via comparative examples) how the algorithms developed for Asian children differ from the ones developed for European children and why?
Other comments/questions:

1) abstract, page 3 line 10: There are numerous diabetes phenotyping algorithms for non-European subjects (TEDDY: https://teddy.epi.usf.edu/, PheKB: https://phekb.org/).

2) page 5, line 15: "diabetes subtype" may not be correct terminology: T1D and T2D are not really considered the same disease. Most phenotyping algorithms explicitly differentiate between the two.

3) page 5, line 40: The study [11] included more patients (43K) than 210, but only 210+100 (randomly selected) patients were chart reviewed to measure performance. This is from a large healthcare provider in the Boston area (Atrius Health), which is not European although the race is mostly white.

4) page 7, line 37: What is the justification for the 31 December 2015 end date criteria? Why not include additional patients with a more recent date?

5) page 7, line 37: How well does the HKDSD criteria work? There may be false positives, but there will be no false negatives included in the HKDSD data set. Any comments?

6) page 7, line 40: How does this exclusion criteria (missing subtype data in the HKDR) impact the selected study population? In reference 11, they included patients that were flagged based on the diabetes criteria but were not classified as either T1D or T2D.

7) page 9, line 36: The characteristics of the validation cohort should be listed somewhere (either in Table 1 along with the derivation cohort or in a supplementary table)

8) page 9, line 41: The average age of T1D dx (22.7 years) seems very high. Please comment. Were a large number of subjects undiagnosed for a long time? Was there sufficient observation history for the subjects before their t1d dx?

9) page 10, line 58: I assume you me "incidence of T1D" not "prevalence of T1D" - since T1D is still not curable and is a chronic disease.

10) page 12, line 36: "highly accurate in classifying T2D" - this can be due in part to the significant class imbalance since it is very easy to do well with the majority class in terms of sensitivity and ppv.

11) page 12, line 54-57: Can the authors discuss more about the age specific performance difference. Is it surprising? Is it expected? Why? T1D is mainly a childhood disease - so it does not seem surprising that you need a younger age group for it to even work. The number of t1d cases for the older
population is very small (and I wonder if they are really just undiagnosed cases - versus real incident cases that occur at the older age).

Minor suggestions:

- In abstract, page 3, line 36: include count (n) for T1D and T2D.

- In abbreviations, page 4, line 24: typo, "positive predictive value" should go with PPV.

- In "study population" (page 7), mention appendix figure 1 which nicely illustrates the impact of the inclusion/exclusion criteria

- On page 8, line 20, should the end date year be 2015 (since that is the end date of the study population mentioned earlier)?

- In Appendix Figure 1: include count (n) of T1D and T2D cases.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

Yes

**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.

No

**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:

Acceptable
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.