Article details: 2018-0142	
Title	**Developing a case definition for type I diabetes mellitus in a primary care electronic medical record database: an exploratory study**
Authors	Brendan Cord Lethebe, Tyler Williamson, Stephanie Garies, Kerry McBrien, Charles Leduc, Sonia Butalia, Boglarka Soos, Marta Shaw, Neil Drummond
Reviewer 1	Holger Kunz
Institution	University College London, London, UK
General comments (author response in bold)	1. Adapt title to include ‘supervised machine learning’. **Thank you. The new title includes ‘supervised machine learning’.** 2. Abstract could be me more specific. Which ML-algorithms have been applied? **Great point. We have modified the abstract to include the statement: “Three decision tree classification algorithms and LASSO logistic regression were used.” Given the word limits we felt this was appropriate.** 3. Explain the feature engineering and data cleansing in more detail. The feature ‘occurrence of “type 1” in any text field’ might be considered as data leakage. It might be useful to remove those features which directly express the class label. **The purpose of this study is to find the set of features that best distinguishes the type of diabetes. The reason that no case definition for type 1 diabetes have been developed in EMR settings is that the ICD9 codes used in ambulatory care by physicians do not distinguish between type 1 and type 2 diabetes. The main advantage we have in using a machine learning approach is that we can use the free text, since the codes may not be sufficiently specific. Occasionally, the only differentiating feature in a individual patient’s EMR that identifies that person as having type 1 diabetes rather than any other type is the free-text “type 1”.** **Perhaps there is some leakage opportunity with the fact that “type 1” doesn’t necessarily have to mean type 1 diabetes. It could, for example be indicative of herpes type 1. There are no patients in our sample that had the text “type 1” for any purpose other than diabetes, so our method would not be able to catch this important exclusion.** 4. Define the hyperparameter tuning of the classifier in more detail. Which hyperparameters were optimised? **It’s different for each algorithm. The LASSO only has one parameter (the penalty for the sum of the absolute value of betas). The C5.0 has several parameters (winnow, complexity parameter, max depth, minCases, loss matrix), as does the rpart (complexity, max depth, minCases, loss matrix). The CHAID algorithm has a few parameters (max depth, p_val threshold) which can be tuned. To make this clearer, we added the following text on page 4:** **“These complexity parameters include maximum depth of the tree, a confidence factor or complexity parameter, a minimum number of cases required to make a split, and a loss matrix.”**

5. Add a method for feature ranking (for example RandomForest or any other method)
We have added table 4 on page 14. This shows variable importance as per the Random Forest method with 500 trees. It shows the top 20 variables using the Mean Decrease in Gini Index.

We agree that it is useful to see which features rank as the most important, but we are not using the Random Forest algorithm for anything else in this study, since it is not interpretable. As this also is over the limit for the tables, we have indicated to the editors that this is acceptable material for the supplement.

6. Provide a definition for area under the curve (AUC). This is an imbalanced dataset and AUC might be a reasonable performance metric.

We agree that AUC is a useful metric in some applications. However, since we have carefully done our sampling of people with diabetes to reflect the true prevalence of type 1 among people with diabetes, we feel that the measures we have selected are appropriate. Furthermore, while it is definitely possible to get ROC curves for decision tree algorithms, they are designed to output a class label, not a predicted probability. For example, when using a loss matrix other than a 1:1 ratio for False positives vs False negatives, R's implementation of the C5.0 algorithm will only output class labels, and not offer probability predictions. This is because the loss matrix is overriding the predicted probability.

However, in response to this request, we have updated the manuscript to use Youden's J rather than the Naïve Mean. Technically, Youden's J is defined as sensitivity+specificity-1, which is mathematically different, but the maximum "naïve Mean" is equivalent to the maximum "Youden's J" in this setting. Youden's J is often used to find the optimal cut-point in the ROC curve. For our purposes, we believe that Youden's J is more informative than the AUC value.

As discussed in our results and interpretation sections, there are two types of case definitions that we find acceptable: 1) A case definition with high PPV (despite low sensitivity), this ensures that the cases we identify are real cases. This is obtained by maximizing the PPV directly, which we are comfortable doing because the prevalence of our sample is representative of the real world; and 2) A case definition that has good balance of Sensitivity and Specificity, which we get by maximizing Youden's J, which is often used in conjunction with the ROC curve.

7. Include a dummy classifier as a baseline algorithm

Thank you, we have added the following to the caption for table 2: "Note that a dummy classifier that assumes all cases are Type 2 diabetes would achieve an accuracy of 91.6%."

8. Add a confusion matrix

Thank you for this comment. Respectfully, we felt that a confusion matrix would be misleading, since we used 10-fold cross validation. A confusion matrix is used in the machine learning settings where we train on a percentage of the data and test on the remaining, without the repeat steps. Essentially, we have 10 confusion matrices for each condition, and for each metric that we maximized. We thought it better to aggregate these using estimates of the sens, spec, ppv, npv.

9. Add area under the curve + AUC-plot in a figure

As explained in point 6, we have reported Youden's J instead of the AUC.

10. Avoid the presentation of 'Accuracy' as this performance metric might be useless in the context of a highly imbalanced dataset unless a sampling method has been used. Focus on F1-score and AUC.

We identified a large cohort of patients that met the previously validated diabetes case definition and randomly sampled from that cohort. This means that our sample is meant to reflect the true prevalence of type 1 diabetes cases among all people with diabetes. Since we have carefully preserved this estimate, we feel that reporting accuracy (1-misclassifiation rate) and reporting PPV and NPV is appropriate. We also used the F1 score and Youden's J, which is often derived from the ROC, much like the AUC.

11. Present the generatxed decision tree (max depth of 3-5 is sufficient) in a separate figure

We have reported two final case definitions, one from the CHAID algorithm, which is a decision tree algorithm, and another from the LASSO, which is not a decision tree algorithm. We have decomposed both into simple rule sets in table 3. We felt that including the plotted decision tree would be redundant, as all the same information is reported in table 3.

12. Provide the optimized hyperparameters of the tuned classifiers

This is another great suggestion. We have created a new table 5. However, as we have exceeded the maximum number of tables, we will recommend to the editors that this be included as supplemental information.

13. Discuss whether the imbalanced dataset had an impact on the performance metrics. This is a highly imbalanced dataset with type 2 (n=1,110) and type 1 (n=110). It is advisable to focus on F1 and AUROC as these performance metrics are more robust against highly imbalanced datasets. Avoid accuracy for highly imbalanced datasets. Discuss whether over-sampling or under-sampling would have been useful.

We identified a large cohort of patients that met the previously validated diabetes case definition and randomly sampled from that cohort. This means that our sample is meant to reflect the true prevalence of type 1 diabetes cases among all people with diabetes. Since we have carefully preserved this estimate, we feel that reporting accuracy (1-misclassifiation rate) along with PPV and NPV is appropriate. We also used the F1 score and Youden's J, which is often derived from the ROC, much like the AUC.

Oversampling type 1 cases at the chart review stage would have been useful in the sense that we would get more cases to work with, but we would be falsely inflating the prevalence of the sample which would hinder our ability to use PPV and NPV, which are important measures for those who use these case definitions.

We have added the following to page 3 to help clarify this point: "By using this method of sampling patients, we attempted to preserve the true distribution of type 1

	and type 2 cases among the undifferentiated diabetes population. This allows us to comfortably report PPV and NPV metrics." 14. Discuss the newly generated feature ranking. Provide an interpretation why some features (variables) have a highly predictive power to distinguish between type 1 and type 2 diabetes as this might be of interest to many readers. **Thank you for the comment. We have added table 4 and the top 20 features as selected by the Random forest model are all clinically relevant at first glance. Obviously, the text for "type 1" is highly predictive. The differences in drug prescriptions (Insulin vs Non-insulin drugs) and age are also well established relationships that distinguish between type 1 and type 2 diabetes. The specialty clinic referrals are also indicative of a type 1 diagnosis.**
Reviewer 2	Kirstin Clemens
Institution	Western University, Medicine, London, Ont.
General comments (author response in bold)	1. I'd appreciate a more thorough description of the CPCSSN database. How often is this database used by researchers? Is it available to anyone? This may help with the generalizability of the study. What do you mean by the database being able to clean, process and code? **Thank you for the pointing out that we had not fully described CPCSSN in the methods section. We have added additional text to try and make this clearer. Given the limited space to describe this study we have directed interested readers to the CPCSSN website and provide references to relevant CPCSSN papers.** 2. Were the family physicians who determined whether patients truly had type 1 vs. 2 diabetes blinded to the machine learning results? You've checked this as yes on your checklist, but I think that this should be explicitly stated in the manuscript. **Thank you, we have added this explicitly in the manuscript on page 3: "These physicians performed this task prior to any machine learning classification"** 3. I wonder about the validity of your gold standard (i.e. having family doctors clarify whether a patient had type 1 vs 2 disease). Should a specialist have done this instead (i.e. an endocrinologist)? There are some characteristics in Table 1 that make me wonder if there was some misclassification of diabetes type (e.g. 12.7% of type 1's had a script for a blood glucose lowering drug excluding insulin in the previous year/26.4% at any time?). **Thank you for this comment. This comment touches on a point that we gave considerable thought to during the design stage. The objective of this research was to develop a case definition for type 1 diabetes using primary care EMR data. We feel that the appropriate reference standard is family doctor diagnosis particularly as type 1 diabetes has little diagnostic uncertainty (extreme thirst, urination, and weight loss), requires immediate medical attention, immediate insulin use and often presents with diabetic ketoacidosis (and a hospitalization). These characteristics make it less susceptible to misclassification errors. Ultimately, we felt that it was important to have the family physician provide the reference information because they had long-term relationships with these individuals and therefore new them best. Specialists would be challenged to review these records (which may be incomplete) and make an outside diagnosis. Further, a lack of insulin prescription in the primary care EMR is expected in some cases given that specialist prescribed medications may not be**

explicitly recorded in the prescribing section of the chart.

4. You validated the algorithm in Alberta, but the database is used across Canada. This is a limitation to mention.

Thank you, we have added the following on page 6: "Also, this study only used CPCSSN data from Alberta, and therefore needs to be validated in other provinces to ensure validity for all of CPCSSN."

5. In your discussion of previous studies, you highlight validation studies of undifferentiated diabetes. It would be more relevant to comment upon any other studies that have validated type 1 vs type 2 diabetes using EMR/administrative data.

Thank you, agreed. Added a few studies where this was the goal. (Citation 9,10)

6. I might reword the first sentence of the introduction to make clearer/more readable.

Thank you, we have reworded the first few sentences of the introduction to make it clearer and more readable as suggested.

7. Apart from Chronic Care models/funding, I think there are other important research gains to be made from validating type 1 vs type 2 diabetes (e.g. outcomes research, drug studies, quality of care etc). I think these should also be highlighted as reasons why this type of work is important in your introduction.

Great point, we have added a sentence at the end of the first paragraph putting a higher emphasis on these points.