Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-23-35

Reviewer A

Comment 1: 1453 patients were consecutively prospectively recruited – what is the recruitment strategy? The data runs from 2012 to 2016, so I don't see how this is prospective. Also, need to describe time frame and how this particular database was constructed.

Reply 1: We thank the reviewer for highlighting the need for more information about the study's methodology. This research was a retrospective cohort study but at the time of cohort formation the patients were all recruited prospectively. Line 122 and 355 have been updated to reflect this update. In relation to the time frame the data was collected from project inception to pathway commissioning in 2016 as added to the text in line 134. The database was constructed using excel.

Comment 2: How were the patients labelled as having ALD or NAFLD? Through a clinically documented process? Also, how well is the data standardized? If I wanted to validate this work in different parts of UK, would it be possible? Were any standardized data like READ codes used?

Reply 2: We thank the reviewer for highlighting this omission. The patients were recruited based on pre-specified READ codes and the relevant papers outlining this process have been referenced in line 117. Patients were labelled as ALD or NAFLD based on the READ codes by which they were identified. However, the sentence on line 139 has been deleted as the disease label was not used as a variable or a categorisation method in the sub-analysis and therefore not relevant. The spectrum of lifestyle related liver disease is being increasingly recognised in current literature and this hopefully allows for this.

Comment 3: Among the 81 variables that were used; how did you handle the correlated data?

Reply 3: We thank the reviewer for highlighting this aspect of analysis. A further supplementary table has been added and further description of the process has been added from line 161- 165.

Comment 4: If you divided 1453 patients down to 2/3, training, 1/3 hold out, and then further braking 2/3 down to testing and validation – it will likely have a small sample of cases in the cohort. Provide a precision recall curve for assessment; as model may not be stable. Also, provide PPV as well as F score.

Reply 4: Many thanks for opening discussion regarding the topic of cohort size. There are advantages and disadvantages of precision recall curves but for our initial approach in which we wish to evaluate the performance across various thresholds, considering both sensitivity and specificity, our view was ROC analysis was better. We do however take on your comment for further work. As a

result of reviewing our work with your comments we have re-labelled the 'Validation data' the 'Testing dataset' to hopefully allow further clarity throughout the paper as to dataset being used.

Comment 5: What was the percentage of the missing data? And what was the rationale for choosing KNN? What was the number of records that had fully filled out details without missing data? If you have missing data across the board and if there is a higher percentage of it, and then apply KNN, imputed results will be biased based on what is available. KNN also assumes that patients which lie in the circumferential region as similar; which may result in the entire dataset to conform to become similar – and may explain why similar results are obtained despite different geography.

Reply 5: We were aware of the challenges of missingness in this real-world data-set and hopefully have explained our methological rationale below and in the added text in the manuscript from line 197 -199. - .
There were varying percentages of missing data for different variables as shown in supplementary table S3. KNN imputation was chosen due to its ability to handle missing values by identifying similar instances and utilizing their values to impute the missing ones. In this case, a value of K = 3 was selected using the elbow method.

Comment 6: Were there differences in clinical practice among different regions? Did you try training a separate model for respective regions?

Reply 6: We felt the individual cohort sizes were not large enough to carry out individual analysis based on geographical location and have updated the wording to convey this in lines 291-294. We are hoping to explore this in further work.

Comment 7: What NLP techniques were used? Keyword search? Existing NLP pipeline? Manual abstraction? Need to describe this process.

Reply 7: We thank the reviewer for highlighting the need for more clarification of the NLP process used in this work. The below text has been added at line 201-209 to expand our explanation of the NLP techniques used.
Pre-processing of the medication and comorbidities included elimination of noise and irrelevant information by removing specific recurrent or obsolete characters and then using. WordNetLemmatizer to reduce words to base form. For comorbidities comma separation and generating n-grams of length 2 was carried out prior to running data through an ICD 10 application programming interface (API) to retrieve the parent code. (20). Patients were then assigned a positive classification if the comorbidity was present. Cosine similarity was used to compare medications with a pre-formulated list based on the British National Formulary (BNF) and assign a label for the parent class of medication. This was then reviewed by the clinical team to ensure correct categorisation.

Comment 8: ML: did you compare 1453 against the remainder of the cohort? How did you handle the imbalance?

Reply 8: 1453 is the entire cohort recruited between 2012-2016.

Comment 9: Table 3: some of the numbers are too small, numbers are jumping around – I wonder if the model is not stable. Did the model reach convergence? Overall: need to describe methodology in greater detail. For example, what was the k value used in nearest neighbour? Was it the standard 3? What were the classifiers that were trained and what were the hyperparameters? Were there any crossfolds used? Without these details, it is hard to tell whether the model was configured properly.

Reply 9: We have discussed and made changes relating to KNN in reply
The following text has been added from line 170 to clarify the classifier training:
When building the ensemble stacker a grid search was used to fine tune the hyperparameter's performance. GradientBoostingClassifier, the meta learner, was set up using several hyperparameters. Estimators were set to n= 1000 and due to this being a classification task the loss function was configured as exponential. The maximum number of variables used was 6 to ensure individual models were built using a diverse set of variables. To overcome overfitting the maximum depth of each tree was 3. To introduce randomness and reduce the correlation between models a subsample ratio of 0.5 was used with a learning rate of 0.001. A random state was applied in order to ensure result reproducibility.

To train the model, the StratifiedKFold technique was employed. StratifiedKFold is a variant of k-fold cross-validation that ensures the preservation of the class distribution in each fold and was used due to good performance in imbalanced datasets. By using StratifiedKFold, the dataset was divided into k equal-sized folds while maintaining the same class distribution as the original data. During training, the model was trained and evaluated k times, with each fold serving as the validation set once while the remaining folds were used for training. This approach helps to mitigate the risk of overfitting and provides a more reliable estimate of the model's performance on unseen data.

Reviewer B

Comment 10: The Ensemble Stacker model showed the best performance at classifying a patient as at high risk or low risk of disease, with a performance shown by an AUC of 0.72 in the validation set - what do the authors think of this as a reliable method of identifying significant fibrosis?
Reply 10: Many thanks for your comments on our paper and raising the below queries. In relation to this point we noted the Ensemble stacker's performance in this study at determining clinically significant liver fibrosis is better than FIB4's performance. Further analysis of any clinical benefits of the ML model will be key for indicating clinical relevance as a screening tool. This has been added to line 343.

Comment 11: The mean BMI was 29.7 kg/m2 - what was the median and the IQR? Were most of the patients overweight? Was there a good range of BMI's or was it mainly overweight/obese?

How does the algorithm perform in patients who have a BMI of between 19.0-24.9kg/m2? Does the algorithm identify lean NAFLD?

Reply 11: We thank the reviewer for highlighting the importance of BMI as a risk factor and the influence on this algorithm development. The median BMI was 27.7 with and IQR of 6.7. Unfortunately within our dataset we were unable to carry out validation of the model in a subgroup with a BMI range of 19-24.9 kg/m2 and there were no patients for whom the ground truth showed clinically significant liver fibrosis i.e. TE > 8 kPa therefore we felt it was not reliable.

Comment 12: I think the authors could have calculated the FIB-4 score for these patients? It would be interesting to compare the FIB-4 value with the algorithm.

Reply 12: We agree with the reviewer that even at this early stage the comparison to routinely used clinical tests is interesting to investigate. We have calculated the FIB4 and have added to the results section and discussion in lines 265 and 322 respectively. The Ensemble stacker does out perform FIB4 and we hope to further investigate this in future work.