Author’s response to reviews

Title: Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records

Authors:

Maria Pikoula (m.pikoula@ucl.ac.uk)
Jennifer Quint (j.quint@imperial.ac.uk)
Francis Nissen (francis.nissen@lshtm.ac.uk)
Harry Hemingway (h.hemingway@ucl.ac.uk)
Liam Smeeth (liam.smeeth@lshtm.ac.uk)
Spiros Denaxas (s.denaxas@ucl.ac.uk)

Version: 1 Date: 26 Feb 2019

Author’s response to reviews:

Rebuttal and Revisions

“Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records”

M Pikoula, J K Quint; F Nissen, H Hemingway, L Smeeth, S Denaxas

Editor Comments:

Thank you for your submission. I have read the manuscript and received one complete review and one cursory review. I would like to have received a third review but given the detailed first review and based on my own assessment of the work, I believe the work is clearly relevant to the scope of the journal, interesting, and needs some minor revisions to be presented in a manner that is informative and replicable. In your revision, please address each comment individually and describe exactly where you have made changes in a tracked changes version of the manuscript in relation to each comment. If you choose not to modify the manuscript in relation to a specific comment, please explain and justify that choice in your response.
Thank you for reviewing our manuscript and for recognizing the importance and relevance of our research.

Finally, please justify in more detail (a) the resampling of the training dataset; and (b) the approach for validating the clusters - perhaps referring to prior equivalent examples in the literature where appropriate.

Please find below our responses to your specific comments:

(All page, equation and figure numbers refer to the revised manuscript)

We performed three complementary analyses (described below) in order to evaluate the results:

The first approach (a) used the standard methodology of repeating the experiment on the test dataset (25% of the overall dataset) and comparing the characteristics of the resulting clusters with the characteristics of the clusters obtained from the training dataset.

The second (b) approach involved analyzing the stability of the results by repeating the experiment on random samples of the training dataset and comparing patient overlap using the Jaccard similarity score.

The third (c) approach addressed the challenge of assigning patients which have not been previously analysed into clusters (essentially the process which clinicians would need to perform if a tool was deployed in clinic). We trained a decision tree classifier using cluster labels as target outcome classes and compared the results obtained using the Jaccard similarity score.

All three approaches have been used in literature before and we provided references here and in the main manuscript. We have revised the manuscript to provide further details on these experiments (page 6, under header “Evaluation”) and included a new Figure (Figure 1) which illustrates the experiment in detail.

a) In order to validate the resulting clustering results, we left out 25% of the data as a test set (Step 1, Figure 1), and independently applied the MCA and k-means clustering process on this set (Steps 7 and 8, Figure 1). The characteristics of the 5 clusters obtained by clustering the test
set were then compared to the characteristics of the 5 clusters resulting from the analysis on the training set.

b) Similar to previous literature, we assessed the stability of the results obtained in the main analysis by repeating the experiment on 100 random samples of the training dataset. We compared the solutions obtained from each sample with the clusters identified in the main analysis by using the Jaccard similarity score which quantifies the proportion of patients that will be grouped in the same corresponding clusters. All combinations of cluster pairs were assessed against one another, and the solution with the best Jaccard similarity score was retained as the score for each sample. The number we report (89%) is the average of all 100 such scores.

c) We followed the process of Burgel et al., and trained a decision tree algorithm on the same features as the ones used as input to the MCA, using a subset of the training set and the cluster labels as target outcome classes (Steps 4 and 5, Figure 1). We used the trained decision tree algorithm to classify patients of the test set into clusters (Step 6, Figure 1) and compared the overall cluster characteristics to the distributions of the original features. As an additional step and in order to verify the decision tree algorithm produces a comparable output to the main analysis, we quantified the overlap of the resulting clusters with the ones obtained by applying the experiment on the test set using the Jaccard similarity score (Step 9, Figure 1).

Reviewer 1: Jenna Wong

Questions and comments raised by Reviewer 1 are answered as follows:

(All page, equation and figure numbers refer to the revised manuscript)

In this paper, Pikoula and colleagues use unsupervised learning - specifically k-means and hierarchical clustering algorithms - to discover, describe, and validate different subtypes of COPD based on electronic health record data. The objective of this paper is important, as the identification of clinically distinct subtypes of COPD could enable the development of more
tailored therapies for this heterogeneous disease. This paper is also noteworthy because it demonstrates a useful application of unsupervised learning in clinical research. To date, unsupervised learning has not been commonly used in clinical research and I suspect that many investigators are not familiar with this type of statistical learning. Below, I have included several comments and suggestions to help improve the current manuscript.

Many thanks for this positive outlook to our paper and for appreciating the essence and motivation of the work.

Major comments:

1. The authors have included sections in the paper titled "Exposure definition" and "Outcome". These headings would be expected for a study employing supervised learning, but these concepts do not exist for an unsupervised learning problem where there is no concept of an outcome nor an exposure. Including such terms in the paper is incorrect/misleading. For example, under the "Exposure definition" section, the authors describe how they identified patients with COPD. But having a diagnosis of COPD in this study is not really an "exposure" in the usual sense of the term, where exposed individuals are compared with unexposed ones. Instead, having a diagnosis of COPD was part of the study inclusion criteria, meaning that there were essentially no unexposed individuals in this study. I suggest the authors avoid using these terms completely and rename these sections to something like "COPD definition" (instead of "Exposure definition"), "Cluster-generating features" (instead of "Covariates"), and "Clinically relevant events associated with cluster assignment" (instead of "Outcome"), which will more clearly communicate to readers how this analysis differs from a supervised learning problem.

We agree that terms such as “exposure” and “outcome” can be potentially misinterpreted, and it was for the benefit of the clinical audience that we used them as they are standard terms used in the literature. We have now changed the terms according to your suggestion in order to better reflect this type of analysis. We have changed the manuscript accordingly:

Pages 4 and 5 (paragraph headers)
2. The authors excluded never-smokers from the analysis, rationalizing this exclusion because the prevalence of COPD in never-smokers is less than five percent in the UK (which I presume is much lower than the prevalence of COPD among smokers). However, I would think it is important to include non-smokers in the analysis because non-smokers with COPD likely comprise a distinct cluster of patients with different risk factors, comorbidities, and prognosis compared to smokers with COPD. Given that the objective of this study is to discover different subtypes of COPD, I suggest the authors do not exclude any types of patients with COPD. Otherwise, the analysis only generalizes to smokers, and thus the title of this paper should be modified to "Identifying clinically important COPD sub-types among smokers using data-driven approaches in primary care population based electronic health records".

Although we agree that it would be of interest to include never-smokers with COPD in the analysis we have excluded them to ensure agreement with UK clinical guidelines. In the UK, in the National Institute of Health and Care Excellent (NICE) clinical guidelines on COPD diagnoses in management in adults3, current smoking or history of smoking is explicitly mentioned as the major risk factor for COPD diagnoses. In order to reduce the rates of misclassification between asthma and COPD occurring in primary care, we excluded non-smokers from our analyses as they are likely asthma patients with an erroneous diagnosis.

We have revised our manuscript to reflect the rationale for this decision:

Page 4: Under header “COPD definition”

3. The authors categorized the three continuous covariates (BMI, FEV1% predicted and eosinophils) in the clustering algorithms. I recommend the authors do not categorize these variables because the clustering algorithms use the Euclidean space between data points to find the best separating boundaries between clusters. By categorizing these variables, the authors are drastically reducing the number of possible "separation boundaries" in the data for finding clusters. Since unsupervised learning aims to discover unknown patterns in the data, I would recommend that the authors leave the data as "raw" as possible, since naturally-occurring clusters of patients may not correspond to these categorical cutpoints, even though they are commonly used.
We agree that categorising covariates might introduce an element of bias in the algorithm, however we considered it a necessary step since our dataset mostly consisted of categorical covariates and only three continuous ones.

In order to use the Euclidean metric, we projected our data onto a 3-dimensional Euclidean space using Multiple Correspondence Analysis (MCA). Since MCA only accepts categorical variables, the only way to include numeric features in the analysis was to turn them into categories, as is usually the practice when analysing mixed datasets. Including these variables in the MCA process allowed us to account for correlations between the two types of variables, and therefore deriving components that reflect the combined effect of all input features. The categorisation was done in consultation with clinicians and in our case was relatively straightforward, given that well-accepted ranges exist for all three variables.

4. Did the authors normalize all variables in the clustering algorithms to ensure they have a similar range of values? For example, GOLD grade had possible values of 1-4, while other variables like depression and anxiety were binary (range 0-1). Since the authors are using algorithms where Euclidean space is the distance metric, it is important that all variables have a similar numerical range or else variables with larger ranges will "overpower" those with smaller ranges.

As described above (comment 3), the Euclidean distance was calculated on the basis of the MCA components, which are the eigenvalues of each datapoint’s projection onto the three principal orthogonal vectors. These eigenvalues have similar ranges and are all centred around zero. We appreciate that this is an important point to mention in the manuscript, and we have added further elaboration as follows:

Page 5-6: Under header “statistical methods”

5. It is not entirely clear to me why the authors did not consider any of the "supplementary" variables in the clustering algorithms (eg, age at diagnosis, asthma diagnosis, health utilization before diagnosis). It seems to me that these variables could also be important cluster characteristics. Perhaps the authors could add an explanation for this decision in the paper.
It is true that many of the supplementary variables could be used as input features in the analysis. We decided on their inclusion or exclusion after consideration of existing literature, review by clinicians and based on the following reasons:

Age at diagnosis and health utilization before diagnosis are both numerical features and for reasons as described in point (3), we decided that in this case it would have been non-trivial to define meaningful categories in order to include the variables in the experiment. Age at diagnosis in particular correlates with a few of the comorbidities listed, and is therefore included in the analysis in a latent way, as illustrated in the resulting cluster characteristics where clusters of distinctly younger and older patients emerge.

The mMRC score was primarily excluded due to high missingness. Asthma diagnosis is an important feature, however, the rates of asthma misdiagnosis are very high, a fact that is reflected in our dataset, as roughly 40% of the cohort has had an asthma diagnosis before or after their COPD diagnosis. We thus decided to include asthma as a supplementary variable after consultation with clinical experts in order to investigate whether clusters of more atopic patients exhibit higher asthma diagnosis rates.

6. I suspect most readers will not be familiar with the Jaccard similarity or the silhouette coefficient. I suggest the authors include more details about these statistics in the paper (eg, how they are calculated, what they represent, how to interpret their values). These statistics seem to be key to the clustering analysis, but they are not described/explained in the paper and none of their values are reported in the main tables/figures. The authors give a nice explanation for the silhouette coefficient in the Supplementary material, so I suggest moving this information and Figures S2 and S3 into the main paper.

Thank you for the suggestion, we have now moved the explanations for the silhouette coefficient and Figures S2 and S3 into the main manuscript. We have additionally included an expanded explanation of the Jaccard similarity. The changes appear in the following points in the manuscript:

Page 6: Second and third paragraph

7. It appears that the authors do not mention Figure 2 anywhere in the paper, although in my opinion, it is extremely informative - possibly more informative than all the other tables and
figures in the paper because it visually shows how well the clusters are separated in 3D/Euclidean space (which is the distance metric used by the clustering algorithms). In Figure 2, it seems that the five clusters are actually not that distinct from one another in 3D space. All the points appear as one large blob that has been segmented into five compartments, rather than five distinct blobs. I think it is very important that the authors discuss their interpretation of Figure 2 in the Results and the implications of Figure 2 on their conclusions in the Discussion.

Figure 2 (Figure 4 in the new manuscript) is mentioned in the manuscript (Page 12) but we agree that further elaboration on the figure is necessary. We have added the following text in the Results section:

Page 12: “However, given this particular low-dimensional representation using the first three MCA components, it becomes clear from Figure 3 that the clusters are not in any obvious way visually separable, and therefore k-means in combination with the Euclidean distance metric would do no more than segment the dataset attempting to minimise the distance from the assigned cluster centres.”

And the following text in the limitations section of the Discussion:

Page 17: “the graphical representation of patients (Figure 4) shows that although some clusters are clearly separable from each other, with no overlapping boundaries, for others those boundaries are not quite clear.”

Minor comments:

1. I suggest the authors explicitly describe the validated algorithms they used to identify patients with COPD and AECOPD (eg. what READ codes and within what time period). Although the authors provide references for these validated phenotyping algorithms, I recommend the authors include enough details about these algorithms to make the current paper as independent as possible.
We have now provided further details on the identification of patients with COPD as well as AECOPD episodes in the Methods section as follows:

Page 4: Under header “COPD definition”

Page 5: Under header “Clinically relevant events associated with cluster assignment”

2. In Figure 1, the number in the top box (104,143) minus the 3 exclusion criteria (1,129 - 20,258 - 41,867) do not yield the number in the second box (42,889). Is there an error in the numbers, or do they not subtract to 42,889 because the exclusion categories are not mutually exclusive? I would also suggest excluding the five boxes showing the clusters at the bottom of Figure 1, as these results are part of the analysis and not the study design. The number of patients in each cluster is also reported in Table 2.

We have edited Figure 1 (renamed to Figure 2 in the revised manuscript) to remove the Clustering results and have instead added a box with the number of patients in the training set. Finally, we corrected a typo, the number in the second box now reads “40,889” rather than “42,889”.

We have added a further note in the figure caption explaining that the numbers are indeed not mutually exclusive and therefore do not add up:

Page 7: Figure 2

3. More detail and accurate terminology are needed to describe the decision tree classifier (DTC) trained to predict the cluster labels and the Cox models used to obtain disease-specific hazard ratios. For the DTC, I suggest that the authors explicitly state: a) that the cluster labels are the "outcome" being predicted by the DTC, b) which variables were used as predictors in the DTC (I presume these are the same variables used as input for the MCA), and c) the % of individuals in the training set that were used to train the DTC. For the Cox model, the statement that the hazard ratios were "adjusted for baseline age and cluster label" is not quite accurate. Assuming that the covariate (predictor) of interest is the cluster label, the authors should write something like, "We
obtained hazard ratios for the association between cluster label and time-to-CVD and respiratory-related mortality, adjusted for baseline age."

Thank you for the suggestions, we have revised the text based on the above suggestions as follows:

Page 12: Under header “Prediction of cluster membership for unseen (test dataset) cases

4. In the Results, the authors write that they "applied the full experiment separately on test data", which was never mentioned in the Methods. I suggest the authors include this information in the Methods and explain their rationale for doing it. It would also be helpful to illustrate this process visually in Figure 1.

Thank you for the suggestion. The Methods section is now updated to include the application of the experiment in the test data as well as the rationale behind it:

Page 6: Under header “Evaluation”

We have included an additional Figure (Figure 1) with detailed information on the clustering as well as the DTC pipeline as applied to both the training and test sets.

5. Although the authors cite their Tables in the Results, I suggest they devote more space in the main text to qualitatively describe their findings and observations. For example, the information contained in the footnotes for Table 3 and Figure 4 could be integrated into the text of the Results section.

We believe that the formatting of the manuscript may be rendering the text under Figure 4 (Figure 6 in the revised manuscript) as well as Table 3 to look like footnotes, but in fact these
paragraphs starting “Cluster 1 …”, etc are meant to be part of the main text. We are happy to revise the manuscript if we have misunderstood the reviewer’s comment.

6. In the Tables, the authors write that "dark and light shading indicates higher and lower proportions respectively with regards to the entire cohort". However, I see that not all cells with proportions higher than the overall cohort are shaded dark, while and not all cells with proportions lower than the over cohort are shaded light. For some rows, the cells with the lowest/highest proportions are shaded light/dark, while for other rows, no cells are shaded at all. Perhaps the authors could be more specific about the criteria they used for the shading to make it easier to interpret (e.g., if the cluster proportion was not X% higher or lower than the overall cohort, then it was not shaded).

Thank you for the suggestion, this is an important point to raise and discuss. The dark and light shading on the tables was applied by clinical experts for two main purposes: a) to highlight information and differences in the resulting clusters that were clinically relevant to the care and management of COPD and b) to highlight the clinical features which were used by clinicians during the labelling process of the results. As a result, the shading does not always reflect an elevated proportion for a value in a particular row as that might not have been clinically-relevant.

We added this explanation in the Results section.

Page 10: Text at top of page

7. What was the reasoning behind performing the sensitivity analysis among patients without a diagnosis of asthma? I also suggest the authors write one or two sentences in the Discussion about how the results from the sensitivity analysis impact the main findings, so that readers know what to take away from this sub analysis.

As discussed above (R1, major comment 5, second paragraph), asthma misdiagnosis and asthma/COPD misclassification is very common in this cohort. The sensitivity analysis
excluding all patients with asthma diagnosis was done in order to investigate the impact of removing all patients with potential true asthma-COPD overlap.

We have updated the text in the Results section as well as the Discussion as follows:

Page 15: Under header “Sensitivity analyses”

Page 16: Second paragraph

References

