Reviewer's report

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

Version: 0  Date: 04 Feb 2019

Reviewer: Jenna Wong

Reviewer's report:

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.

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.

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".

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.

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.

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.

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.

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.

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.

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.

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."

4. In the Results, the authors write that they "applied the full analysis pipeline 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.
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.

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 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 (eg, if the cluster proportion was not X% higher or lower than the overall cohort, then it was not shaded).

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.

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.

Yes

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 recommend additional statistical review
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