Reviewer's report

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

Version: 1 Date: 14 Mar 2019

Reviewer: Jenna Wong

Reviewer's report:

The authors have done a great job of addressing both my comments and the editor's comments. Thanks - I think the paper's clarity has greatly improved. Below, I've noted a few additional minor comments/suggestions that I think could still further improve the paper:

1. I now understand the authors' reasoning for excluding never smokers. However, I still think it is very important to mention this limitation at least somewhere in the Abstract since conceptually, their findings are not guaranteed to generalize to never-smokers. For example, they could modify the first sentence of the Methods section in the Abstract to read something like: "Using linked national structured electronic health records in England that were available through the CALIBER resource, we applied two unsupervised learning algorithms (k-means and hierarchical clustering) in 30,961 patients who were current or former smokers and diagnosed with COPD."

2. Figure 1 has improved a lot. However, I have the following additional suggestions:
   a. The authors should define the acronyms DTC and MCA in the Figure caption.
   b. I suggest switching the positions of the DTC Test set and DTC Training set boxes so that they match the order of the boxes for the Training set and Test set above it (where Training set is to the left, and Testing set is to the right).
   c. The authors should also include a line for '15 features' in the Test set box (so it matches the Training set box - since they also perform MCA on the Test set).
   d. This suggestion may just be a difference in preference, but I would suggest that the arrows connecting 'DTC model' to 'DTC Test set' and 'Test set' point in the opposite direction (away from 'DTC model'), since the DTC model was applied to these datasets. The only box with an arrow pointing into the 'DTC model' should be the 'DTC Training set', since it was used to create the model. With this change, then 5 green arrows should
come out of the 'Test set' box going to the 5 green cluster boxes (the arrows can go behind the orange boxes).

e. I find the Figure caption a little unclear. I would modify the caption to the following (please check that my interpretations for Steps 7-9 are correct):

Main experiment steps (1) Split cohort into Training and Test sets; (2) Apply multiple correspondence analysis (MCA) to the Training set using all 15 potential cluster-generating features, results in 3 components; (3) Use 3 components derived in Step 2 from MCA analysis in k-means algorithm, results in k=5 clusters; (4) Split Training set into a decision tree classifier (DTC) Training and DTC Test set to predict cluster labels obtained from k-means algorithm; (5) Train and validate DTC; (6) Apply DTC to Test set to predict cluster labels; (7) Apply MCA to Test set as in Step 2, results in 3 components; (8) Use 3 components derived in Step 7 from MCA analysis in k-means algorithm, results in k=5 clusters; (9) Compare cluster assignments in Test set from Steps 6 and 8 by calculating the Jaccard Index (% of patients overlapping in the same cluster between the two solutions).

3. COPD definition section - I appreciate that the authors have now included their READ codes in the Appendix. However, it is possible for them to include one sentence in this section about details like the number of COPD READ codes required and within what time window (eg, 1 year, 6 months?) they had to be recorded? These were details I had been thinking most about when I made my comment. Also, based on the authors' tracked comments, I think they had meant to include a reference to the UK clinical guidelines in this section, which has not been added.

4. I appreciate that the authors addressed my comment #4 and added an explanation about the MCA components in the 'Statistical Methods' section. However, I think these explanations on their own will be too technical for most readers to appreciate and connect with the requirements for a k-means algorithm. I suggest the authors include another sentence or two explaining why these characteristics (eg, continuous, orthogonal/uncorrelated, similar numerical range) are important for running a valid k-means algorithm.

5. I appreciate that the authors have added an explanation about the reasoning behind their sensitivity analysis to the manuscript (response to my minor comment #7). However, I think this information belongs in the Methods and should be added to the last paragraph under the 'Evaluation' subsection.
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 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