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: 2 Date: 15 Mar 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 the revised version of the manuscript and the thoughtful changes that have been made. I have received a response from one reviewer and I have read the manuscript myself. We agree that the comments have been addressed well but there are some minor issues that still remain and I think would improve the manuscript. To expedite the process please include individual responses to briefly explain whether the suggested changes were made or not and indicate where the changes can be found in the updated version of the manuscript.
Thank you for the opportunity to submit a revised version of the manuscript and for taking an interest in our work. Please find our responses to the reviewer below.

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)

Jenna Wong (Reviewer 1): 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:

Thank you for the constructive comments and suggestions, and for taking an interest in this work.

Major comments:

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."

We have now revised the abstract to reflect the smoking status of the patients by changing the methods paragraph along the lines of the above suggestion:
“We applied two unsupervised learning algorithms (k-means and hierarchical clustering) in 30,961 current and former smokers diagnosed with COPD, using linked national structured electronic health records in England available through the CALIBER resource.”

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).

We agree with all of the above suggestions and have modified Figure 1 as well as the caption accordingly in page 7 of the manuscript.

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.

We appreciate that the precise definition for COPD cases was not clearly defined. We have now edited the paragraph under the header “Study Population” to precisely explain the way COPD codes were used (at least one code at any point during the study period as long as the patient was at least 35 years old).

We have now added the NICE Guidelines reference that was missing by omission in the previous revised manuscript version.

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.

We agree that more explanation is useful in the methods section and have added the following sentences in page 6 of the manuscript:
“In particular, k-means requires continuous features on comparable scales so as to not be biased towards features with large value ranges. Using orthogonal (uncorrelated) features ensures that highly correlated variables do not dominate cluster assignments.”

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.

We agree with this suggestion and have now added this explanation under the Evaluation heading, bottom of page 6.