Article details: 2022-0193
Title: Identifying clusters of coexisting conditions and outcomes in adults hospitalized with community-acquired pneumonia: a multicentre cohort study
Authors: Sarah L. Malecki MD MSc, Hae Young Jung MSc, Anne Loffler PhD, Mark A. Green PHD, Samir Gupta MD MSc, Derek MacFadden MD ScD, Nick Daneman MD MSc, Ross Upshur MD MSc, Michael Fralick MD PhD, Lauren Lapointe-Shaw MD PhD, Terence Tang MD, Adina Weinerman MD MHSc, Janice L. Kwan MD MPH, Jessica J. Liu MD MSc, Fahad Razak MD MSc, Amol A. Verma MD MPhil

Reviewer 1: Jenna Wong
Institution: Harvard Medical School
General comments (author response in bold)

1. The authors explored the use of 3 different clustering algorithms (K-modes, PAM, and HAC), all of which appeared to produce different results. Only the final clustering solution is presented in the main paper, which was selected through a combination of examining diagnostic plots, inspecting clinical characteristics of the resulting clusters, and incorporating feedback from clinical experts regarding the plausibility of the cluster characteristics. These details are not mentioned in the main paper (likely due to space concerns), but I do think they are important to mention, even if only briefly. Otherwise, the paper gives the impression that the final clustering solution is more robust (to the choice of clustering algorithm) than it may really be. Do the different findings across the 3 clustering methods affect the authors' confidence in the validity of the final clustering solution they have presented? This question would be worth addressing in the Interpretation section.
We have now described our methods more fully in the Supplemental material, and have added some details to the manuscript text.
We did find that the final selected clustering solution was reproducible qualitatively between the PAM and HAC methods (see supplement), and have now noted this strength in the interpretation section. (Supplemental eMethods and eResults, manuscript p.9, and 16.)

2. Relatedly, in the eMethods, one of the reasons given for comparing multiple clustering algorithms is to "minimize any bias introduced by relying on a single method". In the end though, it seems the paper did rely on a single method (PAM). To increase transparency of the study findings and help readers better understand how discrepant the different clustering methods were, it would be good to include results from the other clustering methods in the eAppendix. For example, plot A could include results for K-modes to support the authors' conclusion in the footnotes that this method performed poorly. The authors also presented consensus matrix heat maps for k=7 clusters under PAM and HAC, but the clinical characteristics of the resulting clusters under HAC when k=7 is not reported. I would be interested to know if the prevalence of the Charlson comorbidities under HAC when k=7 was similar to PAM (this would give a sense of how robust the clinical descriptions of the 7 clusters were to the choice of clustering method).
We thank the reviewer for these helpful suggestions.
We have now included results for K-modes in plot A and B of the eAppendix, as well as the clinical characteristics of the resulting clusters for HAC and k modes when k=7 in the supplemental material.
Of note, we do find that HAC when k=7 resulted in qualitatively similar groupings of patients, with some differences that may also be related to the inferior performance of this approach by objective indices, lending support to the

robustness of our approach. We have discussed these findings in the eResults and as above, made a note in the interpretation section as well. (Supplement, eAppendix A, B, eTable 5 and 6, eResults, manuscript p.16.)

3. The authors wrote that a limitation of their study was that they could have missed some potentially important chronic conditions, such as psychiatric illness, that were not included in the Charlson index. Another popular comorbidity index, the Elixhauser, includes more conditions (n=30), including categories for psychoses and depression. Given that studies have shown the Elixhauser index to have slightly superior discrimination over Charlson for predicting inhospital mortality, did the authors consider using Elixhauser instead of Charlson, or combining the two indexes for a more comprehensive set of candidate comorbidities for the clustering algorithm? I'm not necessarily saying the authors should expand or re-do their analysis (I realize this would be a lot of work), but at the very least, this point is worth addressing in the paper, especially given that the Elixhauser system also has associated ICD-10 coding algorithms that were developed and validated by Quan et al.
As per our response to Editorial Comment 5:
We chose the Charlson index as it is among the most commonly used comorbidity indices, ICD-10 codes for its constituent components have been validated, and it can be readily applied in our dataset. In defining our disease clusters, we did not consider the weights in the index. Ultimately, there is no right answer about which list of comorbidities to include, however, we selected Charlson over Elixhauser because the smaller number of comorbidities helps with identifying more parsimonious and clinically-recognizable clusters. We have added the following to our Discussion:
"we used the Charlson comorbidity index to define chronic conditions due to its widespread use, simplicity, and because ICD-10 codes for its components have been validated, but this index is not exhaustive, leaving out some potentially important conditions including psychiatric illness. Other indices such as the Elixhauser index (25) could be considered in future work to further validate our findings." (Manuscript p. 15.)

4. To what extent could there have been measurement error in the LAPS due to lack of available lab test results at baseline, and is it possible this missingness was differential across patient subgroups? Since missing lab test results are assigned 0 LAPS points (corresponding to a normal test result), this would underestimate severity of illness and potentially introduce unmeasured confounding in the association between the clusters and clinical outcomes. Were there more clinical variables in the hospital database that could have been used to measure severity of illness at admission, in addition to LAPS?
There is no missingness in lab data related to data that were not captured in the Gemini system since we excluded patients who were not admitted from the emergency department. The absence of data results from tests not being performed, which occurs due to clinical reasoning. It is possible that test ordering would be differential across groups (ie. doctors might order fewer blood tests in some groups) but the LAPS score was originally developed and validated with the approach of assigning normal values to unmeasured tests, and remains a strong predictor of in-hospital mortality, suggesting that this assumption is generally valid (see ref 27,28 and new reference 29 in manuscript - Escobar et al. 2008, Wong et al 2011, and Van Walraven et al 2011).
Unfortunately, during portions of the study period, many participating hospitals used paper charting for patient vital signs and other clinical variables. We

included all of the clinical variables available to us for risk adjustment. We note that laboratory variables are known to provide excellent additional contribution to risk adjustment beyond administrative data (see Walker et al 2017, The Lancet, Volume 390, Issue 10089, 62 - 72 DOI:https://doi.org/10.1016/S0140-6736(17)30782-1)
We have added to the Limitations:
"We were also unable to include patient vital signs or other clinical markers of illness severity because much of that documentation occurred in paper charts during the study period." (Manuscript p. 7, p. 16.)

5. In addition to creating temporally split datasets to assess reproducibility of the clustering results, did the authors consider splitting the data by hospital to see if any one hospital had exceptional influence on the clustering results? For example, a cross-validation like procedure where the clustering algorithm is derived in n-1 hospitals and validated in the held-out hospital, or derived in 6 hospitals and validated in the largest hospital? Alternatively, a table showing the number of patients and distribution of comorbidities by hospital would give readers a sense of how similar their patient populations were.
Please see our response to editorial comment 1 above.
We have now provided a table showing the distribution of comorbidities by hospital in the supplemental material, and have discussed the main findings and implications in the manuscript. (Supplement eTable2, manuscript p. 16)

6. In the heart failure group, the prevalence of renal disease was notable (22%) and considerably higher than in all the other comorbidity groups (0%-7.8%). Is renal disease also worth highlighting in the clinical description of this patient subgroup?
We agree with the reviewer's suggestion and have added this to the clinical description of this subgroup. We now describe this as:
""heart failure" subgroup (n=1370, 12.4%) with a relatively high prevalence of renal disease"
Notably, adjusting for renal disease and the other comorbidities that were less prevalent did not substantially affect our findings (please see response to Editorial Comment 8). (Manuscript, p. 10-11)

7. A very minor suggestion – I would recommend adding the word "unsupervised" to the title of the manuscript (i.e., "Using unsupervised machine learning to identify…").
We have slightly changed the title of the manuscript, as per the statistical reviewer's suggestion. Therefore, instead of referring to machine learning, we simply highlight cluster identification.

Reviewer 2: Dr. Bridget Ryan
General comments (author response in bold)

Methods:
4. It would be helpful to provide meaning of Charlson and LAPS scoring including range of possible scores and whether higher equals sicker.
We have added this information to the methods section of the manuscript.
Charlson max score is 24.
LAPS max score 256. (Manuscript, p. 7.)

5. In regression, how did you decide to control for age, sex hospital and LAP but not Charlson, LTC, and ambulance?
We excluded overall charlson score since we were already accounting for individual comorbidities.
We have now adjusted for arrival to hospital from nursing home and by ambulance, since these baseline characteristics could theoretically influence outcomes. (Table 3)

6. Patients were nested in seven hospitals; should hospital therefore be included as a second level in a multi-level model rather than controlled at the individual level? Can you add rationale for not doing so?
We adjusted for hospital as a fixed effect because a sample size of 7 hospitals is too small to account for hospitals as a random effect. Random effects models with small samples can be prone to bias and Type 1 error (https://qualitysafety.bmj.com/content/28/12/1032). To account for the fact that patients were clustered within hospitals, we reported cluster-robust standard errors. We have now provided clarification in the manuscript methods.
(Manuscript, p. 9.)

7. The term machine learning is not used in the prose until the Interpretation section. It would be good to link machine learning with cluster analysis in the Methods and provide a reference and short description for those unfamiliar with this technique. If the journal permits the space, I would like to see the first and last paragraphs from eMethods included in the Methods to provide an overview of cluster analysis, with reference there to eMethods for further detail.
We have now added a description of machine learning, reference, and more explanation re: cluster analysis to the Methods text of the supplement and main manuscript. We have also removed the term machine learning from the title.
(Supplemental emethods, manuscript p. 8-9)

8. As well, in the Conclusion, you use for the first time term unsupervised machine learning. Again it would be helpful to provide some definitions in Methods.
We agree with the reviewer's suggestion and have modified the manuscript accordingly. (See prior comments.)

Results:

9. I found the Results well-presented and easy to understand. I found the finding about diabetes having a different effect on outcomes depending on other morbidities a particularly helpful illustration of why it is important to understand the complexity of multimorbidity. Perhaps those with diabetes, but without other significant morbidities, are generally healthier and/or with perhaps better controlled disease. While this cannot be tested directly in these kinds of HA data, using a cluster analysis allowed you to delve into some of this complexity. Further research that expands the list of included chronic conditions would be valuable.
We thank the reviewer for their comments and agree.

10. In Tables 1 and 2, was the chi-square testing an overall significant difference across all clusters? Some clusters were quite similar in their descriptive.

The reviewer is correct that in tables 1 and 2, the chi-square testing is for overall significance across all clusters. We have added this to the table legends to clarify. (Table 1 & 2 Table legends)

Discussion:

11. You indicate the diversity of patients within the seven hospitals indicating the potential generalizability; however, I wonder if the generalizability might be more limited because the hospitals were large urban hospitals. Would you find the same results in smaller rural or remote hospitals, perhaps with different resources? I think a short discussion of this would be helpful.
We agree with the reviewer's comment. We see two aspects of generalizability between urban/rural settings:1) the patient population/ comorbidity pattern (it's not obvious that patients attending rural hospitals would necessarily have different comorbidity patterns) and 2)type of care received and outcomes. The latter may not generalize (e.g. availability of CT scans differs for example).
We have added the following to the limitations:
"Processes of care, such as advanced imaging use, may be less generalizable to smaller hospitals depending on availability of resources" (Manuscript p. 16.)

12. Is there any cohort/temporal effect that needs to be considered with respect to; for example, any differences in treatments from 2010 to 2017?
There may be trends over time. However this analysis would add substantial complexity and is beyond the scope of this paper. Having now defined reproducible clusters over time, this is exactly the type of research question we hope that our work would facilitate in future research.

13. You discuss that the use of steroids may not be appropriate in some clusters and that future research needs to more firmly establish this. Having established the appropriateness of certain drugs for certain clusters of conditions, you indicate this is an opportunity for personalized medicine if there is net benefit for patients in some clusters. I think you handle this uncertainty well and are measured in your discussion.
I do however have a concern that machine learning can lead to a stance that all treatments can be decided through an elaboration of the patient's clinical presentation. While I agree being able to better target treatment holds great promise for those with multimorbidity, I think this may also signal the need to engage with patients to weigh the benefits versus the risks of these treatments. I think that it is important that the use of AI and machine learning be embedded not solely in a biomedical model but that we continue to use a patient-centred approach that also considers the entirety of the person and not only their clinical presentation. I would appreciate a brief elaboration of this in the Discussion where you discuss future research.
We agree with the reviewer about the limitations of AI models to make treatment decisions and have added this as a point of discussion to the manuscript:
"Examining patterns of coexisting conditions, rather than single comorbidities, offers novel insights that align with a proposed paradigm shift from single disease treatment toward "cluster medicine" for patients with multimorbidity (34) and lays the groundwork for decision-support tools(35) to incorporate with patient preferences and other factors to personalize care." (Manuscript p. 13.)

14. Further to item #13, you indicated that the seven hospitals serve diverse multiethnic populations; however, I think it should be stated as a limitation that you were not able to

identify personal characteristics other than age and sex (I am assuming the hospitals did not have data available such as gender, race, and other social determinants of health). Inclusion of these kinds of data going forward in machine learning will be important and I think should be acknowledged as a limitation.

We agree with the reviewer's comment and have added this as a limitation to the manuscript. (Manuscript p.16.)