Using machine learning to identify patterns of coexisting conditions and outcomes in adults hospitalized with community-acquired pneumonia: A multicentre cohort study

Running Title: Multimorbidity patterns and outcomes in pneumonia

Sarah L. Malecki, MD, MSc, Department of Internal Medicine, University of Toronto, Ontario, Canada
Email: sarah.voll@mail.utoronto.ca

Hae Young Jung, MSc, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada
Email: HaeYoung.Jung@unityhealth.to

Mark A. Green, PHD, Department of Geography & Planning, University of Liverpool, Liverpool, UK
Email: Mark.Green@liverpool.ac.uk

Samir Gupta, MD, MSc, Division of Respirology, Department of Medicine, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada; Li Ka Sing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada
Email: Samir.Gupta@unityhealth.to

Derek MacFadden, MD, ScD, The Ottawa Hospital Research Institute and the University of Ottawa
Email: derek.macfadden@mail.utoronto.ca

Nick Daneman, MD, MSc, Sunnybrook Health Sciences Centre, University of Toronto
Email: Nick.Daneman@sunnybrook.ca

Ross Upshur, MD, MSc, the Division of Clinical Public Health, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada
Email: ross.upshur@utoronto.ca

Michael Fralick, MD, PhD SM Sinai Health System and the Department of Medicine, University of Toronto, Toronto, ON, Canada
Email: mike.fralick@mail.utoronto.ca

Lauren Lapointe-Shaw MD PhD Department of Medicine, University Health Network and University of Toronto
Email: lauren.lapointe-shaw@uhn.ca

Terence Tang, MD. Department of Medicine, Trillium Health Partners and University of Toronto.
Email: Terence.Tang@thp.ca

Adina Weinerman, MD, MHSc. Department of Medicine, Sunnybrook Health Sciences Centre and University of Toronto.
Email: Adina.Weinerman@sunnybrook.ca

Janice L. Kwan, MD MPH, Department of Medicine, Sinai Health System and University of Toronto
Email: Janice.Kwan@sinaihealth.ca

Jessica J. Liu, MD MSc, Department of Medicine, University Health Network and University of Toronto
Email: Jessica.liu@uhn.ca

Fahad Razak*, MD, MSc, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada; Department of Medicine, University of Toronto, Toronto, ON, Canada; Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
Email: fahad.razak@mail.utoronto.ca

Amol A. Verma*, MD, MPhil, Li Ka Shing Knowledge Institute, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada; Department of Medicine, University of Toronto, Toronto, ON, Canada; Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada
Email: amol.verma@mail.utoronto.ca

***Co-Senior authors**

Corresponding author:
Amol Verma
Li Ka Shing Knowledge Institute, St. Michael's Hospital
Room 714-2, 2 Queen St. East, Toronto, Ontario, M5C 3G7, Canada.
Email: amol.verma@mail.utoronto.ca Tel: (416) 864-6060 x76934, Fax: (416) 864-5904.

Number of references: 51
Number of tables: 3
Number of figures: 2
Appendices: 1
Manuscript word length: 3000

Abstract

Background: Little is known about patterns of coexisting conditions and their influence on clinical care or outcomes in adults hospitalized for community-acquired-pneumonia (CAP).

Methods: We studied 11,085 adults hospitalized with CAP at 7 hospitals in Ontario, Canada. Using unsupervised machine learning, we identified patient subgroups based on clustering of the comorbidities in the Charlson index. Subgroups were derived and validated in independent cohorts (derivation: 2010-15, validation: 2015-17). Differences in medications, imaging, and outcomes were described.

Results: Patients clustered into seven subgroups. The "low comorbidity" subgroup (27.5%, N=3052) had no comorbidities. The "DM-HF-Pulm" subgroup had prevalent diabetes, heart failure, and chronic lung disease (15.4%, N=1710). One disease category defined each remaining subgroup: "pulmonary" (14.6%, N=1621), "diabetes" (11.6%, N=1281), "heart failure" (12.4%, N=1370), "dementia" (9.4%, N=1038), and "cancer" (9.1%, N=1013). Corticosteroid use ranged from 11.5% to 64.9% in the "dementia" and "pulmonary" subgroups, respectively. Piperacillin-tazobactam use ranged from 9.1% to 28.0% in the "pulmonary" and "cancer" subgroups, respectively. Thoracic computed tomography ranged from 5.7% to 36.3% in the "dementia" and "cancer" subgroups, respectively. Adjusting for patient factors, in-hospital mortality was greater in the "cancer" (aOR 2.91, 95%CI: 2.20-3.86), "dementia" (aOR 1.73, 95%CI: 1.32-2.27), "heart failure" (aOR 1.66, 95%CI: 1.27-2.16), and "DM-HF-Pulm" subgroups (aOR 1.32, 95%CI: 1.02-1.71) and lower in "diabetes" (aOR 0.65, 95%CI: 0.46-0.93) compared to "low comorbidity".

Interpretation: Patients hospitalized with CAP cluster into clinically-recognizable subgroups based on coexisting conditions. Clinical care and outcomes vary among these subgroups with little evidence to guide decision-making, highlighting opportunities for research to personalize care.

Key words: Cluster analysis; Community-acquired pneumonia (CAP); Hospital medicine; Multimorbidity; Unsupervised machine learning

Introduction

Pneumonia is one of the most common reasons for hospitalization(1) and patients experience a wide range of clinical outcomes.(2, 3) The clinical care of patients with pneumonia is known to vary with respect to choice of antibiotics,(2) type of imaging use,(4) and adjunctive therapies.(5) It is not known whether patterns of coexisting conditions are associated with differences in clinical care or outcomes in patients hospitalized with pneumonia. As populations age, more people are living with multiple chronic conditions.(6) Although single coexisting diseases, such as dementia,(7) and greater comorbidity levels in general,(8-12) are known to affect clinical outcomes in patients with pneumonia, less is understood about patterns of coexisting illnesses among patients hospitalized for pneumonia. Clinical practice guidelines for pneumonia offer little guidance for how multiple coexisting conditions should affect care.(2, 13) Host phenotyping has been identified as a crucial next step in advancing the treatment of pneumonia, including calling for a focus on improving our understanding of comorbid illnesses.(14)

The objective of this study was to examine how coexisting conditions cluster in patients hospitalized with community-acquired pneumonia (CAP). We hypothesized that clinically-recognizable subgroups could be identified based on patterns of coexisting conditions, and that subgroups would differ in diagnostic imaging, medication use, and clinical outcomes. Our overall aim was to advance understanding of how multimorbidity affects CAP and to inform future research toward more personalized treatment strategies for patients hospitalized with CAP.

Methods

Design and Setting

This was a retrospective cohort study using data from seven large hospitals in Toronto and Mississauga, Ontario, Canada that were participating in the General Medicine Inpatient Initiative (GEMINI).(1) GEMINI collects administrative and clinical data from all general internal medicine (GIM) admissions. Clinical data are extracted from hospital information systems and administrative data are collected from hospitals as reported to the Canadian Institute for Health Information (CIHI) National Ambulatory Care Reporting System and Discharge Abstract Database.(15, 16) Manual validation of more than 20,000 data points within the GEMINI database demonstrated that data are highly reliable.(17) The participating hospitals serve diverse multiethnic urban and suburban populations and hospital services are publicly insured.

Study Sample

We included all GIM patients discharged between April 1, 2010 and October 31, 2017. At all participating hospitals, patients with CAP who are not admitted to the ICU are admitted almost entirely to GIM services rather than specialized respirology services. To identify patients with CAP, we included patients for whom the most responsible discharge diagnosis as reported to CIHI was "pneumonia", defined by the International and Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada (ICD-10-CA) codes J10-J18.(1, 18, 19) We also included patients for whom pneumonia was a co-morbid diagnosis with a most responsible discharge diagnosis of Chronic Obstructive Pulmonary Disease (COPD, defined by ICD-10–CA codes J41-J44).(20) These patients were included because coding convention dictates that COPD be coded as the primary diagnosis for patients with coexisting pneumonia.(16) Prior chart abstraction studies revealed that the ICD-10 code J18 alone had a sensitivity of 80% for pneumonia(21) whereas the group of ICD-10 codes J10-J18 were found to be 98% sensitive and 97% specific for pneumonia in patients 65 years and older.(18) To enhance the specificity of case identification and to separate CAP from hospital-acquired pneumonia, we only included patients who received an antibiotic with activity against respiratory pathogens(13, 22, 23) on every day for the first four days of admission or until death or hospital discharge in accordance with a standard five-day treatment regimen for CAP (13, 23) (assuming up to one day of antimicrobial administration in the emergency

department prior to admission). We excluded patients who were not admitted from the emergency department, or who were admitted to hospital in the previous 30 days, given the possibility that their pneumonia may have been related to the prior hospitalization. For patients with multiple admissions, we included only one randomly-selected admission during the study period.

Research Ethics Board Approval

This study received Research Ethics Board approval with a waiver of informed patient consent from all participating hospitals.

Measures and Outcomes

Patient characteristics

Baseline patient characteristics included age, sex, residence in a long-term care facility, transport to hospital by ambulance, overall level of comorbidity estimated using the Charlson Comorbidity Index score,(24, 25) and severity of illness, estimated using the laboratory-based acute physiology score (LAPS),(26) which is a validated predictor of in-hospital mortality based on 14 laboratory tests.(27)

Coexisting conditions

We selected comorbid conditions of interest based on one of the most widely-used comorbidity indices, the Charlson Comorbidity Index.(24) We measured conditions that are included in this index, based on patient discharge diagnoses categorized using ICD-10 codes.(25) Sensitivity and specificity for the majority of these ICD-10 codes have been reported previously and all were >95% specific while sensitivity ranged from 25% for HIV/AIDS to 83% for metastatic cancer.(28) The Charlson Comorbidity Index defines chronic lung disease as

all obstructive and restrictive diseases.(25) It also separates diabetes, liver disease, and malignancy into sub-categories based on disease severity and complications,(25) which we collapsed into single categories for each disease.

Processes of clinical care

We described the use of respiratory-acting antibiotics, other medications intended to improve respiration (glucocorticoids, inhalers and furosemide), and the use of computed tomography (CT) of the thorax.

Clinical Outcomes

The study outcomes were: in-hospital mortality, ICU admission after admission to GIM, total hospital length of stay, and readmission to GIM at any participating hospital, within 30 days of discharge.

Statistical Analysis

Cluster analysis

We performed cluster analysis to identify subgroups of patients with CAP based on the presence of 14 coexisting medical condition groupings that form the Charlson Comorbidity Index (see eMethods for full methodological details). In order to assess the stability and reproducibility of the identified patient subgroups, we performed the same cluster analysis in a derivation cohort (April 1, 2010 to March 31, 2015) and validation cohort (April 1, 2015 to October 31, 2017), similar to Seymour and colleagues(29) (eFigure 1 in the Supplement). Patients who had an admission in both the derivation and validation period were excluded from the latter so that they were only captured once. Demographics, baseline characteristics and the prevalence of coexisting conditions were reported for the two cohorts. Standardized mean differences >0.10 (10%) were used

to identify any meaningful imbalance between the cohorts.(30) After confirming a stable and reproducible clustering approach in the two cohorts, the cluster analysis was re-run using the entire study period (April 1, 2010-October, 31 2017) (eFigure 1) and all further analyses were performed on the total cohort.

We compared patient characteristics, clinical care and outcomes across subgroups using chi-square tests for categorical variables and Kruskall-Wallis tests for continuous variables. Separate logistic regression models were used to examine the effect of coexisting condition subgroups on each of: in-hospital mortality, 30-day readmission, and ICU admission. Quantile regression was used to model the non-binary outcome: median hospital length-of-stay. Models were adjusted for age, sex, hospital, and LAPS. All analyses were performed in R version 4.0.0 (R Foundation for Statistical Computing, Vienna).

Results

Overall, 11,085 patients were included in the study cohort (eFigure 1). The median age was 79 (IQR 65-87) and 52.6% were male. The mean Charlson Index score was 1.7 (SD 1.7). The five most common coexisting conditions were chronic lung disease (n=3178, 28.7%), diabetes (n=2978, 26.9%), heart failure (n=1892, 17.1%), dementia (n=1401, 12.6%) and cancer (n=1194, 10.8 %).

eTable 1 summarizes demographics, baseline characteristics and prevalence of coexisting conditions in the derivation, validation and total cohorts. There were 7066 patients in the derivation cohort and 4019 patients in the validation cohort, and the two cohorts were generally similar.

Cluster analysis

Subgroups were driven primarily by the five most common comorbidities in the cohort, and 72.5% of patients had at least one of these five conditions (Figure 1). A seven cluster solution was selected as the optimal set of

subgroups from our cluster analysis (Figure 1, also see eResults and eAppendix for more details): 1) "low comorbidity" subgroup (n=3052, 27.5%), which had none of the coexisting conditions in the Charlson Index; 2) "diabetes-heart failure-pulmonary" (DM-HF-Pulm) subgroup (n=1710, 15.4%), which was a multimorbid subgroup with high prevalence of all three of those conditions; 3) "pulmonary" subgroup (n=1621, 14.6%), which included patients with either chronic obstructive or restrictive lung diseases; 4) "diabetes" subgroup (n=1281, 11.6%); 5) "heart failure" subgroup (n=1370, 12.4%); 6) "dementia" subgroup (n=1038, 9.4%), and 7) "cancer" subgroup (n=1013, 9.1%).

Patient Characteristics

Subgroups differed significantly in age, sex and other baseline characteristics (Table 1). The "cancer" subgroup was the youngest of all the subgroups (median age 72 years) while the "dementia" subgroup was the oldest (median age 86 years). The "cancer" subgroup had the most males (60.7%), while the "dementia" subgroup had the least (45.9%). The "dementia" subgroup had the highest proportion of patients from a nursing home (37.6%) and arriving to hospital by ambulance (89.5%). The "DM-HF-Pulm" group had the highest presenting LAPS (mean=27.1, SD 18.8), whereas the "low comorbidity" subgroup had the lowest (mean=20.7, SD=15.3). The "cancer" subgroup had the highest Charlson index score (mean=3.8, SD 1.9) whereas the "low comorbidity" subgroup had the lowest (mean 0.0, SD 0.0), by definition.

Clinical care

The use of respiratory-acting antibiotic classes, glucocorticoids, inhalers, furosemide, and CT of the thorax differed significantly between subgroups (Table 2). The most notable differences were the high use of piperacillin-tazobactam (28.0%) and CT thorax (36.3%) in the "cancer" subgroup, compared to the overall population (13.3% and 18.3%, respectively). Use of fluoroquinolone antibiotics was highest in the "pulmonary" subgroup (48.2%) and CT thorax use was lowest in the dementia subgroup (5.7%). Use of glucocorticoids was

greatest in the "pulmonary" subgroup (64.9%) as was use of all inhaler types. Furosemide use was greatest in the "DM-HF-pulm" subgroup (61.6%) .

Clinical outcomes

Subgroups differed significantly in the four clinical outcomes examined (Figure 2, Table 3). Compared to the overall study population, the "low comorbidity" subgroup had fewer deaths (4.2% vs 6.9%), ICU admissions (6.7% vs. 8.9%), 30-day readmissions (7.7% vs. 10.0%), and shorter hospital length-of-stay (median 3.7 days, IQR 2.0 to 6.7 days vs. 4.7 days, IQR 2.6 to 8.5 days). Conversely, the "DM-HF-Pulm" subgroup had worse outcomes than the overall population on all measures: deaths 8.4%, ICU admissions 14.4%, 30-day readmission 13.2%, and median length-of-stay 6.2 days (IQR 3.4 to 10.7 days).

After adjusting for age, sex, hospital, and presenting LAPS, the risk of death was greater in the "cancer" (adjusted OR: 2.91, 95% CI: 2.20, 3.86), "dementia" (aOR 1.73, 95%CI: 1.32, 2.27), "heart failure" (aOR 1.66, 95%CI: 1.27, 2.16), and "DM-HF-Pulm" (aOR 1.32, 95%CI: 1.02, 1.71) subgroups compared to the "low comorbidity" subgroup. The "heart failure" and "DM-HF-Pulm" subgroups had worse outcomes on all four measures compared to the "low comorbidity" subgroup. The "diabetes" subgroup had lower risk of death (aOR 0.65, 95% CI: 0.46, 0.93) than the "low comorbidity" subgroup and had no significant differences in ICU admission, readmission, or length-of-stay. The "pulmonary" subgroup had greater ICU use (aOR 1.61, 95% CI 1.28, 2.02) and median length-of-stay (0.44 days longer, 95% CI 0.17, 0.71 days) than the "low comorbidity" subgroup but no significant difference in death. The "dementia" subgroup had greater risk of death (aOR 1.73 95% CI 1.32, 2.27), readmission (aOR 1.35, 95% CI 1.05, 1.75) and longer median length-of-stay (1.32 days longer, 95% CI 0.93, 1.71 days) but no significant difference in ICU use.

Interpretation

We used machine learning techniques to identify seven reproducible and clinically-recognizable subgroups of patients hospitalized with CAP based on patterns of coexisting conditions. Our study offers four major contributions to the literature. First, we found that five disease categories were the most prevalent coexisting conditions and drive the pattern of clustering: chronic lung diseases, diabetes mellitus, heart failure, dementia, and cancer. Second, we characterized the pattern of disease clustering. Five subgroups were dominated by a single disease category (pulmonary, diabetes, heart failure, dementia, and cancer). One subgroup represented a classically "multimorbid" phenotype with high prevalence of chronic lung disease, diabetes, heart failure, renal disease and prior myocardial infarction and one subgroup reflected patients with little comorbidity. Third, we found that use of diagnostic imaging, antibiotics, and other medications differed among these subgroups. Fourth, clinical outcomes differed among these subgroups, even after controlling for age, sex, and severity of illness at presentation. 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 (31) and lays the groundwork for decision-support tools to personalize care.(32)

Comparison to previous studies and implications for future research

Our study extends the previous literature related to comorbidity and CAP, which has focused on describing the prevalence of coexisting conditions,(33) associating single coexisting conditions with outcomes,(7, 34-37) and measuring comorbidity in general(8-12) rather than exploring patterns of disease. Diabetes mellitus has been associated with significantly increased mortality in patients hospitalized with pneumonia.(35-37) We found that 42.7% of patients with diabetes were part of the subgroup with high rates of chronic lung disease and heart failure, 8.3% had coexisting dementia, and 6.0% had coexisting cancer. All of these subgroups had significantly greater mortality risk, and poorer outcomes in general, than patients with no comorbidities. However, another 43.0% of patients with diabetes were in a subgroup without other Charlson comorbidities and these patients had significantly lower mortality and no significant differences in ICU use, readmission or length-of-stay compared

to patients with no comorbidities. This reveals that the relationship between diabetes, pneumonia, and mortality is not as simple as was previously understood. The association between diabetes and adverse outcomes in pneumonia may be driven by the degree of other organ-system involvement and highlights interesting opportunities for future research.

The "DM-HF-Pulm" subgroup had the most coexisting conditions and had poor outcomes overall, similar to prior studies of multimorbidity in pneumonia.(8-10) The specific pattern of coexisting conditions illuminates opportunities for further research in this subgroup. For example, the use of macrolide (55.6%) and fluoroquinolone (45.8%) antibiotics was not lower in this subgroup, but these drugs cause cardiac complications.(38, 39) Corticosteroids were prescribed in 51.1% of patients in this group, perhaps in part to treat concomitant COPD exacerbations, but corticosteroids may also worsen heart failure(40) and glycemic control.(41) Corticosteroid use varied across subgroups, from 64.9% in the "Pulmonary" subgroup to 11.5% of patients in the "low comorbidity" subgroup. There may be practice variation related to the controversial literature on the benefits of corticosteroids in non-severe CAP.(5, 42, 43) Further research could seek to quantify whether the risks and benefits of corticosteroids vary across subgroups, and differences in net benefits may provide opportunities for more personalized medicine.

The "pulmonary" subgroup had greater ICU use and longer hospital stays but no increased risk of mortality, consistent with prior literature.(44, 45) (9, 46) These findings correspond with the COPD GOLD guidelines,(45) which caution against therapeutic pessimism among patients hospitalized with acute exacerbations of COPD. Mortality was greater in the dementia, cancer and heart failure subgroups than in patients with no comorbidities, which is similar to previous studies.(7, 47)·(48) The dementia subgroup had less ICU and thoracic CT use overall, suggesting that clinicians and patients may be opting for less intensive approaches. The cancer subgroup had a greater use of thoracic CT scans (36.3% vs 18.3% overall) and greater use of broad-spectrum antibiotics (e.g. piperacillin-tazobactam used in 28.0% of patients vs 13.3% overall), which may be related to neutropenia or risk factors for *Pseudomonas* infection. However, there remains limited evidence about when to

select broader antibiotic therapy or advanced diagnostic imaging in patients with cancer and CAP. Further research should seek to clarify what patient factors are associated with differences in therapeutic and diagnostic choices and determine whether there are opportunities to standardize, personalize, and improve care.

Limitations and Strengths

Our study has several limitations. First, we used ICD-10 CA codes to identify medical conditions in our cohort, including CAP. Although some studies suggest these codes are highly specific, their sensitivity varies.(18, 28) This may lead to misclassification, primarily by underestimating certain conditions and overestimating the prevalence of the "low comorbidity" subgroup. We augmented ICD-10 CA codes with clinical data regarding antibiotic use to increase the specificity of our definition of CAP. Second, we used the Charlson comorbidity index to define chronic conditions, but this index is not exhaustive, leaving out some potentially important conditions including psychiatric illness. We also used the same disease groupings as in the Charlson index,(25) which do not represent single diseases (e.g. chronic lung disease, cancer, dementia, heart failure, and diabetes are all heterogeneous categories, to varying degrees). Nevertheless, the prevalence of the most common conditions in our cohort was generally similar to population-based studies of pneumonia in the United States,(33) the United Kingdom,(49) and Canada,(50) including several with prospectively-collected comorbidity data, suggesting our findings are likely generalizable. Third, coexisting conditions were measured at discharge and may not have been present on admission. However the majority of these conditions are chronic diseases and it is unlikely that the admission for CAP would represent the first occurrence of this disease. For example, the incidence of cancer after hospitalization for CAP has been reported as 1.1% within 90 days of discharge and the rate of discovery during the CAP hospitalization is likely even lower.(51) Fourth, our dataset included only patients admitted to GIM. Nearly all CAP patients are admitted to GIM at participating hospitals, with the exception of a small number of patients with complex lung diseases or acute coronary syndromes who may be cared for on dedicated respirology or cardiology units. The seven participating hospitals serve diverse multiethnic populations in two of Canada's largest cities and we used temporally-split datasets to assess the

reproducibility of clustering results. We believe our results are likely generalizable but should be externally validated.

Conclusion

In this study, unsupervised machine learning methods were able to identify stable and clinically-recognizable subgroups of patients hospitalized to GIM with CAP based on coexisting conditions. Clinical care and outcomes vary among these subgroups, despite no strong evidence about how comorbid illnesses should inform treatment decisions. This highlights opportunities for future research about whether and how hospital care for patients with CAP can be more personalized.

List of abbreviations

CAP: Community-acquired pneumonia

CIHI: Canadian Institute for Health Information

CT: Computed tomography

GEMINI: General Medicine Inpatient Initiative

GIM: General Internal Medicine

ICD-10-CA: International and Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Canada

LAPS: Laboratory-based acute physiology score

Declarations

Ethics Approval

Also stated above in methods**.** This study received Research Ethics Board approval with a waiver of informed patient consent from all participating hospitals.

Consent for Publication: Not applicable.

Availability of data

Data from this manuscript can be accessed upon request to the corresponding author, to the extent that is possible in compliance with local research ethics board requirements and data sharing agreements.

Competing interests

The authors declare that they have no competing interests.

Funding

The development of the GEMINI data platform has been supported with funding from the Canadian Cancer Society, the Canadian Frailty Network, the Canadian Institutes of Health Research, the Canadian Medical Protective Agency, Green Shield Canada Foundation, the Natural Sciences and Engineering Research Council of Canada, Ontario Health, the St. Michael's Hospital Association Innovation Fund, the University of Toronto Department of Medicine, and in-kind support from partner hospitals and Vector Institute.

Author's contributions

The study was designed by SM, AAV and FR with substantial input from all authors. SM, HYJ, MF, LL-S, TT, AW, JK, JL, FR and AAV contributed to data collection. HYJ performed data analysis. MG provided methodological support regarding cluster analysis. All authors contributed to interpretation of the results. SM wrote the first manuscript draft. All authors provided input for critical revisions, approved the final version submitted for publication, and agreed to be accountable for all parts of the study.

Acknowledgements: Not applicable.

References

1. Verma AA, Guo Y, Kwan JL, Lapointe-Shaw L, Rawal S, Tang T, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. CMAJ open. 2017;5(4):E842-e9.
2. Lim WS, Baudouin SV, George RC, Hill AT, Jamieson C, Le Jeune I, et al. BTS guidelines for the management of community acquired pneumonia in adults: update 2009. Thorax. 2009;64(Suppl 3):iii1-iii55.
3. Aliberti S, Dela Cruz CS, Amati F, Sotgiu G, Restrepo MI. Community-acquired pneumonia. The Lancet. 2021;398(10303):906-19.
4. Garg D, Johnson LB, Szpunar S, Fishbain JT. Clinical value of chest computerized tomography scans in patients admitted with pneumonia. Journal of hospital medicine. 2014;9(7):447-50.
5. Stern A, Skalsky K, Avni T, Carrara E, Leibovici L, Paul M. Corticosteroids for pneumonia. Cochrane Database of Systematic Reviews. 2017(12).
6. Pefoyo AJ, Bronskill SE, Gruneir A, Calzavara A, Thavorn K, Petrosyan Y, et al. The increasing burden and complexity of multimorbidity. BMC public health. 2015;15:415.
7. Foley NC, Affoo RH, Martin RE. A systematic review and meta-analysis examining pneumonia-associated mortality in dementia. Dement Geriatr Cogn Disord. 2015;39(1-2):52-67.
8. Weir DL, Majumdar SR, McAlister FA, Marrie TJ, Eurich DT. The impact of multimorbidity on short-term events in patients with community-acquired pneumonia: prospective cohort study. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2015;21(3):264.e7-.e13.
9. Luna CM, Palma I, Niederman MS, Membriani E, Giovini V, Wiemken TL, et al. The Impact of Age and Comorbidities on the Mortality of Patients of Different Age Groups Admitted with Community-acquired Pneumonia. Annals of the American Thoracic Society. 2016;13(9):1519-26.
10. Cilloniz C, Polverino E, Ewig S, Aliberti S, Gabarrus A, Menendez R, et al. Impact of age and comorbidity on cause and outcome in community-acquired pneumonia. Chest. 2013;144(3):999-1007.
11. Aujesky D, Fine MJ. The Pneumonia Severity Index: A Decade after the Initial Derivation and Validation, . Clinical Infectious Diseases. 2008; 47(Supplement_3,):S133–S9.
12. Vrbova L, Mamdani M, Moineddin R, Jaakimainen L, Upshur REG. Does socioeconomic status affect mortality subsequent to hospital admission for community acquired pneumonia among older persons? Journal of Negative Results in BioMedicine. 2005;4(1):4.
13. Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K, et al. Diagnosis and Treatment of Adults with Community-acquired Pneumonia. An Official Clinical Practice Guideline of the American Thoracic Society and Infectious Diseases Society of America American Journal of Respiratory and Critical Care Medicine. 2019;200(7).
14. Dela Cruz CS, Evans SE, Restrepo MI, Dean N, Torres A, Amara-Elori I, et al. Understanding the Host in the Management of Pneumonia. An Official American Thoracic Society Workshop Report. Annals of the American Thoracic Society. 2021;18(7):1087-97.
15. CIHI. National Ambulatory Care Reporting System metadata (NACRS) [Available from: https://www.cihi.ca/en/national-ambulatory-care-reporting-system-metadata-nacrs.
16. CIHI. Discharge Abstract Database metadata (DAD) [Available from: https://www.cihi.ca/en/discharge-abstract-database-metadata-dad.
17. Verma AA, Pasricha SV, Jung HY, Kushnir V, Mak DYF, Koppula R, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. J Am Med Inform Assoc. 2020.

18. Skull S, Andrews R, Byrnes G, Campbell D, Nolan T, Brown G, et al. ICD-10 codes are a valid tool for identification of pneumonia in hospitalized patients aged ⩾65 years. Epidemiology and Infection. 2008;136(2):232-40.
19. Henriksen DP, Nielsen SL, Laursen CB, Hallas J, Pedersen C, Lassen AT. How well do discharge diagnoses identify hospitalised patients with community-acquired infections?--a validation study. PloS one. 2014;9(3):e92891.
20. Quality-Based Procedures: Clinical Handbook for Chronic Obstructive Pulmonary Disease (Acute and Postacute). Health Quality Ontario & Ministry of Health and Long-Term Care; 2015.
21. Juurlink D PC, Croxford R, Chong A, Austin P, Tu J, Laupacis A. Canadian Institute for Health Information Discharge Abstract Database: A Validation Study ICES; 2006.
22. Dragan V, Wei Y, Elligsen M, Kiss A, Walker SAN, Leis JA. Prophylactic Antimicrobial Therapy for Acute Aspiration Pneumonitis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2018;67(4):513-8.
23. Mandell LA, Wunderink RG, Anzueto A, Bartlett JG, Campbell GD, Dean NC, et al. Consensus Guidelines on the Management of Community-Acquired Pneumonia in Adults. Clinical Infectious Diseases. 2007;44(Supplement_2, 1):S27–S72.
24. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987;40(5):373-83.
25. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Medical care. 2005;43(11):1130-9.
26. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. Medical care. 2008;46(3):232-9.
27. Wong J, Taljaard M, Forster AJ, Escobar GJ, van Walraven C. Derivation and validation of a model to predict daily risk of death in hospital. Medical care. 2011;49(8):734-43.
28. Quan H, Li B, Saunders LD, Parsons GA, Nilsson CI, Alibhai A, et al. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. Health services research. 2008;43(4):1424-41.
29. Seymour CW, Kennedy JN, Wang S, Chang C-CH, Elliott CF, Xu Z, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. JAMA. 2019;321 (20):2003-17.
30. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. Communications in Statistics - Simulation and Computation. 2009;38(6):1228-34.
31. Whitty CJM, MacEwen C, Goddard A, Alderson D, Marshall M, Calderwood C, et al. Rising to the challenge of multimorbidity. Bmj. 2020;368:l6964.
32. BMJ Best Practice. +Comorbidities, Treat the Whole Patient 2020 [Available from: https://bestpractice.bmj.com/info/comorbidities/.
33. Ramirez JA, Wiemken TL, Peyrani P, Arnold FW, Kelley R, Mattingly WA, et al. Adults Hospitalized With Pneumonia in the United States: Incidence, Epidemiology, and Mortality. Clinical Infectious Diseases. 2017;65(11):1806-12.
34. Di Yacovo S, Garcia-Vidal C, Viasus D, Adamuz J, Oriol I, Gili F, et al. Clinical features, etiology, and outcomes of community-acquired pneumonia in patients with diabetes mellitus. Medicine (Baltimore). 2013;92(1):42-50.
35. Falguera M, Pifarre R, Martin A, Sheikh A, Moreno A. Etiology and outcome of community-acquired pneumonia in patients with diabetes mellitus. Chest. 2005;128(5):3233-9.
36. Kornum JB, Thomsen RW, Riis A, Lervang H-H, Schønheyder HC, Sørensen HT. Type 2 Diabetes and Pneumonia Outcomes. A population-based cohort study. 2007;30(9):2251-7.
37. Martins M, Boavida JM, Raposo JF, Froes F, Nunes B, Ribeiro RT, et al. Diabetes hinders community-acquired pneumonia outcomes in hospitalized patients. BMJ Open Diabetes Res Care. 2016;4(1):e000181.

38. Juurlink DN. The cardiovascular safety of azithromycin. CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne. 2014;186(15):1127-8.
39. Bonaldo G, Andriani LA, D'Annibali O, Motola D, Vaccheri A. Cardiovascular safety of macrolide and fluoroquinolone antibiotics: An analysis of the WHO database of adverse drug reactions. Pharmacoepidemiology and Drug Safety. 2019;28(11):1457-63.
40. Celutkiene J, Balciunas M, Kablucko D, Vaitkeviciute L, Blasciuk J, Danila E. Challenges of Treating Acute Heart Failure in Patients with Chronic Obstructive Pulmonary Disease. Cardiac failure review. 2017;3(1):56-61.
41. Corticosteroids in Community-Acquired Pneumonia. JAMA. 2020;23(9):887-8.
42. Corticosteroid Therapy for Patients Hospitalized With Community-Acquired Pneumonia. Annals of Internal Medicine. 2015;163(7):519-28.
43. Lloyd M, Karahalios A, Janus E, Skinner EH, Haines T, De Silva A, et al. Effectiveness of a Bundled Intervention Including Adjunctive Corticosteroids on Outcomes of Hospitalized Patients With Community-Acquired Pneumonia: A Stepped-Wedge Randomized Clinical Trial. JAMA Intern Med. 2019;179(8):1052-60.
44. Esteban A, Anzueto A, Frutos F, Alía I, Brochard L, Stewart TE, et al. Characteristics and Outcomes in Adult Patients Receiving Mechanical VentilationA 28-Day International Study. JAMA. 2002;287(3):345-55.
45. Global Initiative for Chronic Obstructive Lung Disease (GOLD). Global Strategy for the Diagnosis, Management, andPrevention of Chronic Obstructive Pulmonary Disease. 2020.
46. de Miguel-Díez J, López-de-Andrés A, Hernández-Barrera V, Jiménez-Trujillo I, Méndez-Bailón M, de Miguel-Yanes JM, et al. Impact of COPD on outcomes in hospitalized patients with community-acquired pneumonia: Analysis of the Spanish national hospital discharge database (2004–2013). European journal of internal medicine. 2017;43:69-76.
47. Mortensen EM, Coley CM, Singer DE, Marrie TJ, Obrosky DS, Kapoor WN, et al. Causes of Death for Patients With Community-Acquired Pneumonia: Results From the Pneumonia Patient Outcomes Research Team Cohort Study. Archives of Internal Medicine. 2002;162(9):1059-64.
48. Thomsen RW, Kasatpibal N, Riis A, Nørgaard M, Sørensen HT. The impact of pre-existing heart failure on pneumonia prognosis: population-based cohort study. Journal of general internal medicine. 2008;23(9):1407-13.
49. Daniel P, Woodhead M, Welham S, Mckeever TM, Lim WS. Mortality reduction in adult community-acquired pneumonia in the UK (2009–2014): results from the British Thoracic Society audit programme. Thorax. 2016;71(11):1061-3.
50. Feagan BG, Marrie TJ, Lau CY, Wheeler SL, Wong CJ, Vandervoort MK. Treatment and outcomes of community-acquired pneumonia at Canadian hospitals. Canadian Medical Association Journal. 2000;162(10):1415-20.
51. Tang KL, Eurich DT, Minhas-Sandhu JK, Marrie TJ, Majumdar SR. Incidence, Correlates, and Chest Radiographic Yield of New Lung Cancer Diagnosis in 3398 Patients With Pneumonia. Archives of Internal Medicine. 2011;171(13):1193-8.

Figure Legends

Figure 1. Subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions. See text for details regarding cluster analysis. Number refers to number of patients. Conditions refer to coexisting conditions. Pulmonary=chronic lung disease including both obstructive and restrictive, DM= diabetes mellitus, CHF=congestive heart failure, renal=renal disease, MI=myocardial infarction, Liver= liver disease, PVD=peripheral vascular disease, Rheumatic=rheumatic disease, PUD=peptic ulcer disease. Subgroups were named by the condition(s) present in all cluster members or a large proportion if no single condition was present in 100% of the patients within a subgroup. DM-HF-Pulm= subgroup composed of a large portion of patients with diabetes, congestive heart failure and chronic lung disease.

Figure 2. Outcomes for subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions. Coloured bars are the differences in outcome (proportion or median) between the overall cohort and each subgroup. Overall cohort includes admissions that are not belonging to the subgroup being compared, e.g. the dementia subgroup admissions vs all other admissions except those in dementia subgroup. Error bars represent 95% confidence intervals (Wilson's score based interval for proportions and percentile bootstrap interval with 2000 replications for length-of-stay). In the table, the unadjusted outcomes are reported for each subgroup.

Table 1. Baseline characteristics for subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions.

	Overall	**Low Comorbidity**	**DM-HF-Pulm**	**Pulmonary**	**Diabetes**	**Heart Failure**	**Dementia**	**Cancer**	***p***
Number	11085	3052	1710	1621	1281	1370	1038	1013	
Age (median [IQR])	79 [65, 87]	75 [55, 86]	80 [71, 87]	77 [64, 85]	75 [66, 83]	83 [69,90]	86 [81, 91]	72 [62, 82]	<0.001
Male sex (%)	5832 (52.6)	1533 (50.2)	940 (55.0)	838 (51.7)	741 (57.7)	689 (50.3)	476 (45.9)	615 (60.7)	<0.001
From nursing home (%)	1224 (11.0)	213 (7.0)	237 (13.9)	113 (7.0)	102 (8.0)	139 (10.1)	390 (37.6)	30 (3.0)	<0.001
Arrived to hospital via ambulance (%)	6849 (61.8)	1672 (54.8)	1146 (67.0)	1002 (61.8)	765 (59.7)	867 (63.3)	929 (89.5)	468 (68.2)	<0.001
LAPS (mean (SD))	23.4 (16.9)	20.7 (15.3)	27.1 (18.8)	22.1 (17.0)	25.8 (16.6)	25.3 (17.7)	24.5 (16.4)	21.2 (15.8)	<0.001
Charlson score (mean (SD))	1.7 (1.7)	0.0 (0.0)	3.4 (1.4)	1.3 (0.7)	1.8 (1.0)	1.8 (1.1)	1.8 (1.0)	3.8 (2.0)	<0.001

Table 1 legend. Number=number of patients. Age is in years, IQR=interquartile range, LAPS=laboratory acute physiology score, Charlson score=calculated Charlson comorbidity index. P=2-tailed p-value for differences between subgroups, determined by chi-square test for categorical variables and Kruskall-Wallis tests for continuous variables.

Table 2. Antibiotic use, medications and imaging use among subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions.

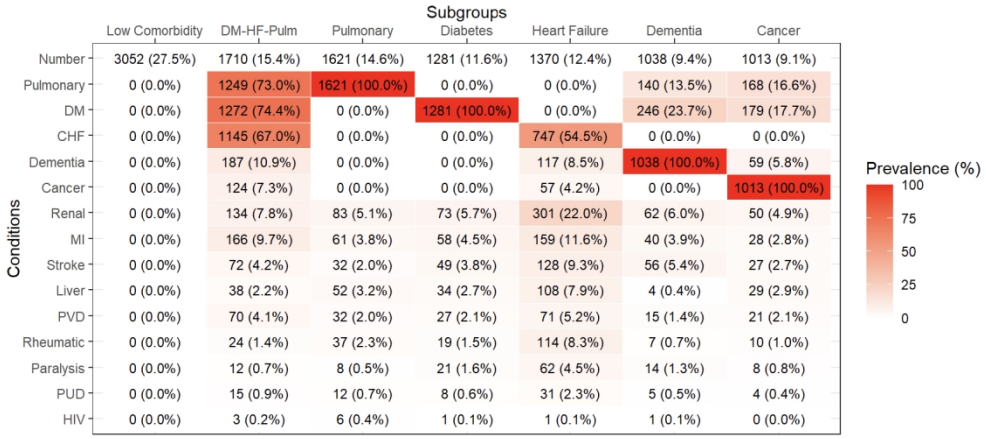
Variable	Overall	Low Comorbidity	DM-HF-Pulm	Pulmonary	Diabetes	Heart Failure	Dementia	Cancer	p
Number	11085	3052	1710	1621	1281	1370	1038	1013	
Third Gen Ceph	6696 (60.4)	1921 (62.9)	1021 (59.7)	885 (54.6)	767 (59.9)	840 (61.3)	658 (63.4)	604 (59.6)	<0.001
Macrolide	6305 (56.9)	1803 (59.1)	950 (55.6)	892 (55.0)	700 (54.6)	768 (56.1)	562 (54.1)	630 (62.2)	<0.001
Fluorquinolone	4592 (41.4)	1136 (37.2)	784 (45.8)	781 (48.2)	523 (40.8)	566 (41.3)	434 (41.8)	368 (36.3)	<0.001
Pen-derived BL	2131 (19.2)	614 (20.1)	310 (18.1)	357 (22.0)	240 (18.7)	247 (18.0)	170 (16.4)	193 (19.1)	0.006
Pip-tazo	1472 (13.3)	334 (10.9)	232 (13.6)	148 (9.1)	169 (13.2)	173 (12.6)	132 (12.7)	284 (28.0)	<0.001
Other	862 (7.8)	253 (8.3)	110 (6.4)	104 (6.4)	88 (6.9)	124 (9.1)	70 (6.7)	113 (11.2)	<0.001
MRSA	592 (5.3)	164 (5.4)	82 (4.8)	58 (3.6)	75 (5.9)	78 (5.7)	60 (5.8)	75 (7.4)	0.002
Simple penicillins	296 (2.7)	96 (3.1)	43 (2.5)	43 (2.7)	34 (2.7)	39 (2.8)	24 (2.3)	17 (1.7)	0.292
Ceftazidime	159 (1.4)	24 (0.8)	35 (2.0)	46 (2.8)	12 (0.9)	19 (1.4)	8 (0.8)	15 (1.5)	<0.001
Tetracyclines	115 (1.0)	19 (0.6)	19 (1.1)	23 (1.4)	15 (1.2)	20 (1.5)	7 (0.7)	12 (1.2)	0.07
Carbapenems (p)	106 (1.0)	16 (0.5)	15 (0.9)	11 (0.7)	17 (1.3)	12 (0.9)	10 (1.0)	25 (2.5)	<0.001
Clindamycin	59 (0.5)	11 (0.4)	11 (0.6)	6 (0.4)	10 (0.8)	8 (0.6)	8 (0.8)	5 (0.5)	0.468
Carbapenems (np)	38 (0.3)	7 (0.2)	5 (0.3)	7 (0.4)	7 (0.5)	5 (0.4)	6 (0.6)	1 (0.1)	0.352
CT thorax	2032 (18.3)	609 (20.0)	241 (14.1)	341 (21.0)	180 (14.1)	234 (17.1)	59 (5.7)	368 (36.3)	<0.001
Furosemide	3217 (29.0)	441 (14.4)	1054 (61.6)	314 (19.4)	342 (26.7)	694 (50.7)	205 (19.7)	167 (16.5)	<0.001
Glucocorticoid	3119 (28.1)	351 (11.5)	874 (51.1)	1052 (64.9)	176 (13.7)	232 (16.9)	119 (11.5)	315 (31.1)	<0.001
SABA	5179 (46.7)	976 (32.0)	1245 (72.8)	1355 (83.6)	451 (35.2)	466 (34.0)	345 (33.2)	341 (33.7)	<0.001
SAMA	3611 (32.6)	530 (17.4)	1021 (59.7)	1093 (67.4)	251 (19.6)	280 (20.4)	233 (22.4)	203 (20.0)	<0.001
LABA	100 (0.9)	8 (0.3)	24 (1.4)	48 (3.0)	3 (0.2)	7 (0.5)	2 (0.2)	8 (0.8)	<0.001
LAMA	1770 (16.0)	126 (4.1)	595 (34.8)	733 (45.2)	57 (4.4)	52 (3.8)	101 (9.7)	106 (10.5)	<0.001

Table 2 legend. Number=number of patients. Antibiotics included only those specifically mentioned in the IDSA guidelines(13, 23) or Dragen et al.(22) Cephalexin was also included since it has the same spectrum of activity to cefazolin. Third gen Ceph= Third generation cephalosporin=Ceftriaxone, cefotaxime, cefepime, cefdinir, cefditoren, cefpoxidime, ceftaroline. Macrolide=Azithromycin, clarithromycin, erythromycin. Fluoroquinolone=levofloxacin, moxifloxacin, ciprofloxacin, gemifloxicin. Tetracyclines=doxycycline . Pen-derived BL=Penicillin-derived beta-lactamases=amoxicillin-clavulinic acid, ampicillin-sulbactam, ticarcillin-clavulanate. Pip-tazo=piperacillin-tazobactam. Carbapenems (p)= carbapenems (pseudomonas coverage)= meropenem, imipenem, impenem+cilastatin. Carbapenem (np)= carbapenems (no pseudomonas coverage)=ertapenem. MRSA coverage=vancomycin, linezolid. Simple penicillins=Penicillin G, amoxicillin, ticarcillin, flucloxacillin, ampicillin, pipracillin. Other=aztreonam, streptomycin, colistin, gentamicin, Septra (trimethoprim-sulfamethoxazole), 1st+2nd gen cephalosporins (cefazolin, cefprozil, cefuroxime, cephalexin). CT thorax= CT thorax performed in first 4 days of admission. SABA=short-acting beta agonist. SAMA=short-acting muscarinic antagonist. LABA=long-acting beta agonist. LAMA=long-acting muscarinic antagonist. P=2-tailed p-value for differences between subgroups, determined by chi-square test.

Table 3. Association of patient subgroup based on coexisting conditions with clinical outcomes after multivariable adjustment.

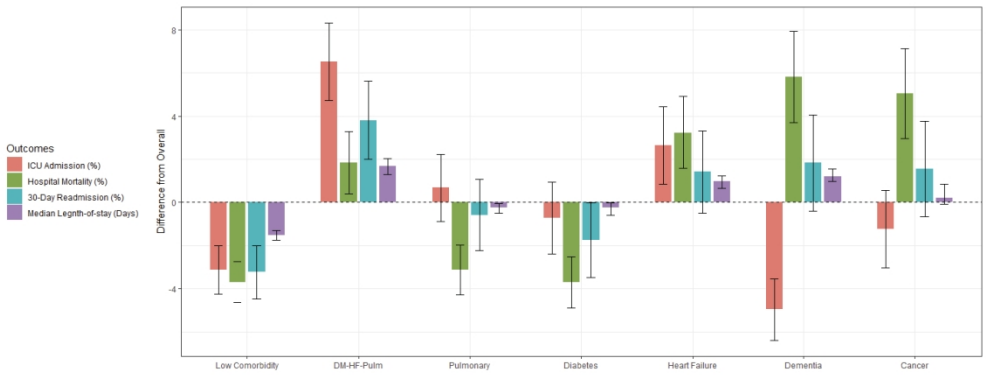
	Mortality		ICU Admission		30-day readmission		Median Length of Stay	
Subgroup	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**	**Coeff (95% CI)**	**p-value**
Low comorbidity	Reference							
DM-HF-Pulm	1.32 (1.02-1.71)	0.036	2.50 (2.02-3.10)	<0.001	1.58 (1.28-1.95)	<0.001	1.68 (1.35-2.01)	<0.001
Pulmonary	0.83 (0.61-1.12)	0.230	1.61 (1.28-2.02)	<0.001	1.20 (0.96-1.49)	0.112	0.44 (0.17-0.71)	<0.001
Diabetes	0.65 (0.46-0.93)	0.017	1.27 (0.98-1.65)	0.065	1.04 (0.81-1.33)	0.774	0.22 (-0.05-0.49)	0.112
Heart failure	1.66 (1.27-2.16)	<0.001	1.84 (1.45-2.32)	<0.001	1.32 (1.05-1.66)	0.017	1.37 (1.03-1.71)	<0.001
Dementia	1.73 (1.32-2.27)	<0.001	0.90 (0.64-1.28)	0.571	1.35 (1.05-1.75)	0.019	1.32 (0.93-1.71)	<0.001
Cancer	2.91 (2.20-3.86)	<0.001	1.32 (0.98-1.76)	0.063	1.38 (1.06-1.78)	0.016	1.15 (0.78-1.52)	<0.001

Table 3 legend. Results for mortality, ICU admission and 30-day readmission are from binary Logistic Regression analysis. Results for length of stay are from Quantile Regression. Each subgroup was defined as a binary variable and compared to the "low comorbidity" subgroup as a reference. Models were adjusted for patient age, sex, hospital, and laboratory-based acute physiology score. OR=odds ratio. Coeff=coefficient in quantile regression. CI=confidence interval.



Subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions. See text for details regarding cluster analysis. Number refers to number of patients. Conditions refer to coexisting conditions. Pulmonary=chronic lung disease including both obstructive and restrictive, DM= diabetes mellitus, CHF=congestive heart failure, renal=renal disease, MI=myocardial infarction, Liver= liver disease, PVD=peripheral vascular disease, Rheumatic=rheumatic disease, PUD=peptic ulcer disease. Subgroups were named by the condition(s) present in all cluster members or a large proportion if no single condition was present in 100% of the patients within a subgroup. DM-HF-Pulm= subgroup composed of a large portion of patients with diabetes, congestive heart failure and chronic lung disease.

190x85mm (230 x 230 DPI)



	Overall	Low Comorbidity	DM-HF-Pulm	Pulmonary	Diabetes	Heart Failure	Dementia	Cancer
Number	11085	3052	1710	1621	1281	1370	1038	1013
ICU Admission - n (%)	989 (8.9)	203 (6.7)	247 (14.4)	154 (9.5)	106 (8.3)	154 (11.2)	46 (4.4)	79 (7.8)
Hospital Mortality - n (%)	761 (6.9)	128 (4.2)	144 (8.4)	68 (4.2)	46 (3.6)	133 (9.7)	126 (12.1)	116 (11.5)
30-Day Readmission - n (%)	1022 (10.0)	221 (7.7)	206 (13.2)	146 (9.5)	103 (8.4)	138 (11.2)	106 (11.6)	102 (11.4)
Legnth-of-stay - median [IQR]	4.7 [2.6, 8.5]	3.7 [2.0, 6.7]	6.2 [3.4, 10.7]	4.5 [2.6, 7.6]	4.5 [2.6, 7.6]	5.6 [3.2, 9.6]	5.8 [3.5, 10.6]	4.9 [2.6, 9.0]

Outcomes for subgroups of patients with community acquired pneumonia admitted to General Internal Medicine (2010-2017) identified by cluster analysis according to coexisting conditions. Coloured bars are the differences in outcome (proportion or median) between the overall cohort and each subgroup. Overall cohort includes admissions that are not belonging to the subgroup being compared, e.g. the dementia subgroup admissions vs all other admissions except those in dementia subgroup. Error bars represent 95% confidence intervals (Wilson's score based interval for proportions and percentile bootstrap interval with 2000 replications for length-of-stay). In the table, the unadjusted outcomes are reported for each subgroup.

190x118mm (193 x 193 DPI)

Supporting information for "Using machine learning to identify patterns of coexisting conditions and outcomes in adults hospitalized with community-acquired pneumonia: A multicentre cohort study "

Sarah L. Malecki, Hae Young Jung, Mark Green, Samir Gupta, Derek MacFadden, Nick Daneman, Ross Upshur, Michael Fralick, Lauren Lapointe-Shaw, Terence Tang, Adina Weinerman, Janice L. Kwan, Jessica J. Liu, Fahad Razak, Amol A. Verma

eMethods

Cluster analysis

We used a consensus cluster analysis approach to derive clusters in the derivation cohort, similar to that employed in recent studies describing sepsis phenotypes[1] and ICU subgroups[2]. Using a baseline clustering algorithm, this approach consists of performing *x* algorithm replications to form a consensus matrix between pairs of observations. A hierarchical clustering algorithm is then run on the consensus values to obtain the final clustering solution.

We compared the performance of three different baseline clustering alorithms (K-modes,[3] partitioning around mediods [PAM][4] and hierarchical aggglomerative clustering [HAC]).[4] These three algorithms were selected because they could each deal with binary data (an important limitation of common methods). There is no common agreement in the literature over the 'best' clustering algorithms. Comparing the solutions derived from three models allowed us to evaluate their performance, minimise any bias introduced by relying on a single method and selected the approach that performed best with our data. Models were run using a modified version of the R package "ConsensusClusterPlus". We performed 100 replications of K-modes, PAM and HAC using 80% resampling of the cohort with each iteration to obtain three final consensus clustering solutions.

Unsupervised cluster analysis methods require defining the number of clusters within a model (with the algorithm then iteratively refining the allocation of cases into the selected number of groups). We did not have *a priori* justification of what types of clusters to expect. We took an exploratory data-driven approach to select the number of clusters that best summarised our data. Because there is no single metric to define optimal clustering, we examined numerous measures and visualizations to select the best-fitting cluster solution across *k*=2-10 clusters. We did not consider more than 10 clusters as we wanted to find the parsimonious solution. Similar to Seymour et al. 2019,[1] best fit was determined by examination of characteristics of consensus cumulative distribution function plots[1] and consensus matrix heat maps to select a solution that maximized separation of clusters.[1] We ensured that pairwise consensus values between cluster members was >0.8.[1] We also calculated and plotted the silhouette width[5] (looking for the maximum value), expected versus observed cluster size and total sum of squares (looking for the minimum value) across k=2-10 clusters.

Finally, given the subjective nature of interpreting cluster analyses, we examined the clusters for identifiable clinical patterns. Study co-investigators with clinical expertise in general internal medicine and respirology were asked to provide feedback on the interpretation of cluster characteristics and if they made sense clinically, to come to a final consensus on the optimal clustering solution.

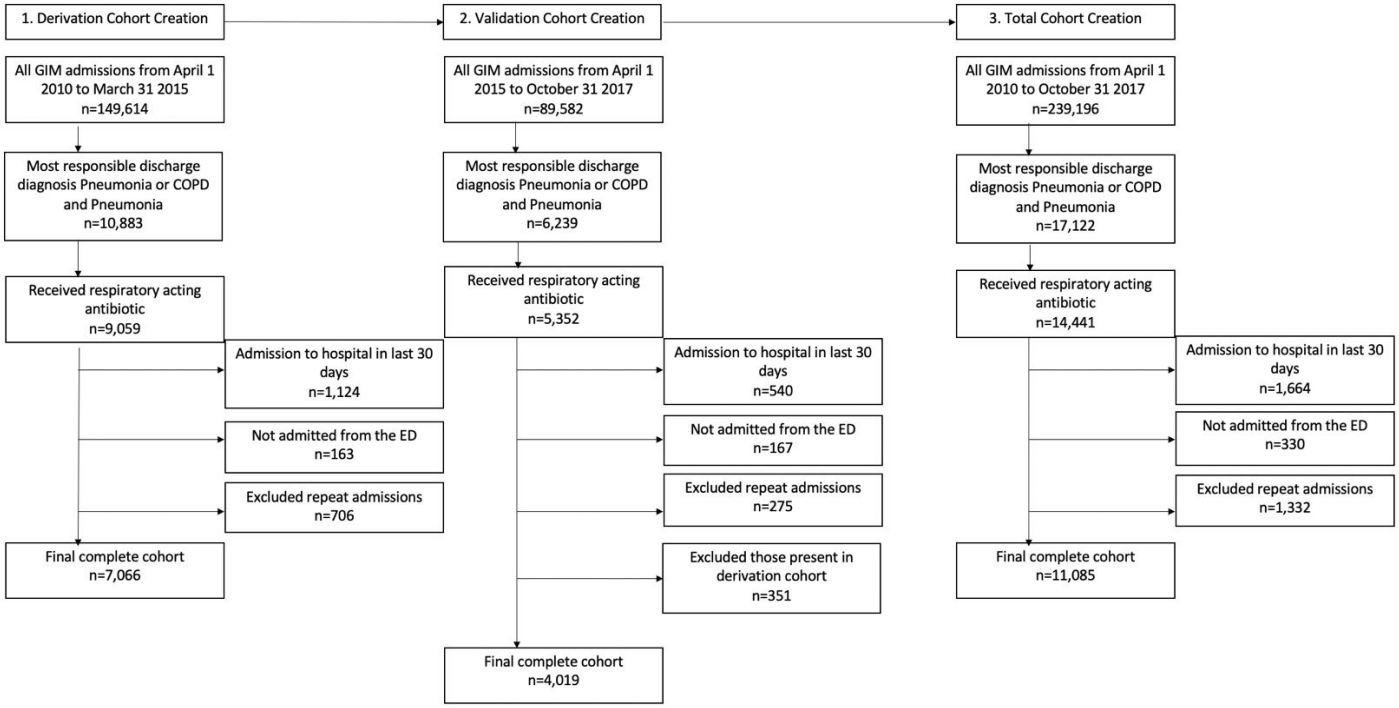
eResults

Cluster Analysis

Based on examination of the consensus plots and additional indices to evaluate different clustering solutions (see Appendix), PAM was selected as the method of choice because it yielded a better clustering solution than K-modes or HAC, regardless of the number of clusters chosen.

For the derivation cohort, clustering solution PAM k=7 was the best solution based on objective indices selected (Appendix). Other candidate clustering solutions, including PAM k=6 and k=8 were presented to the coauthors. Qualitatively, the clusters produced by PAM k=5,6,7 and 8 solutions were similar and PAM7 was selected as not only the best on the objective indices but also balancing clinically meaningful results with simplicity. PAM k=7 was reproducible in the validation cohort (eTable 2 and 3). Therefore, k=7 clusters was selected as the most clinically relevant and reproducible clustering solution.

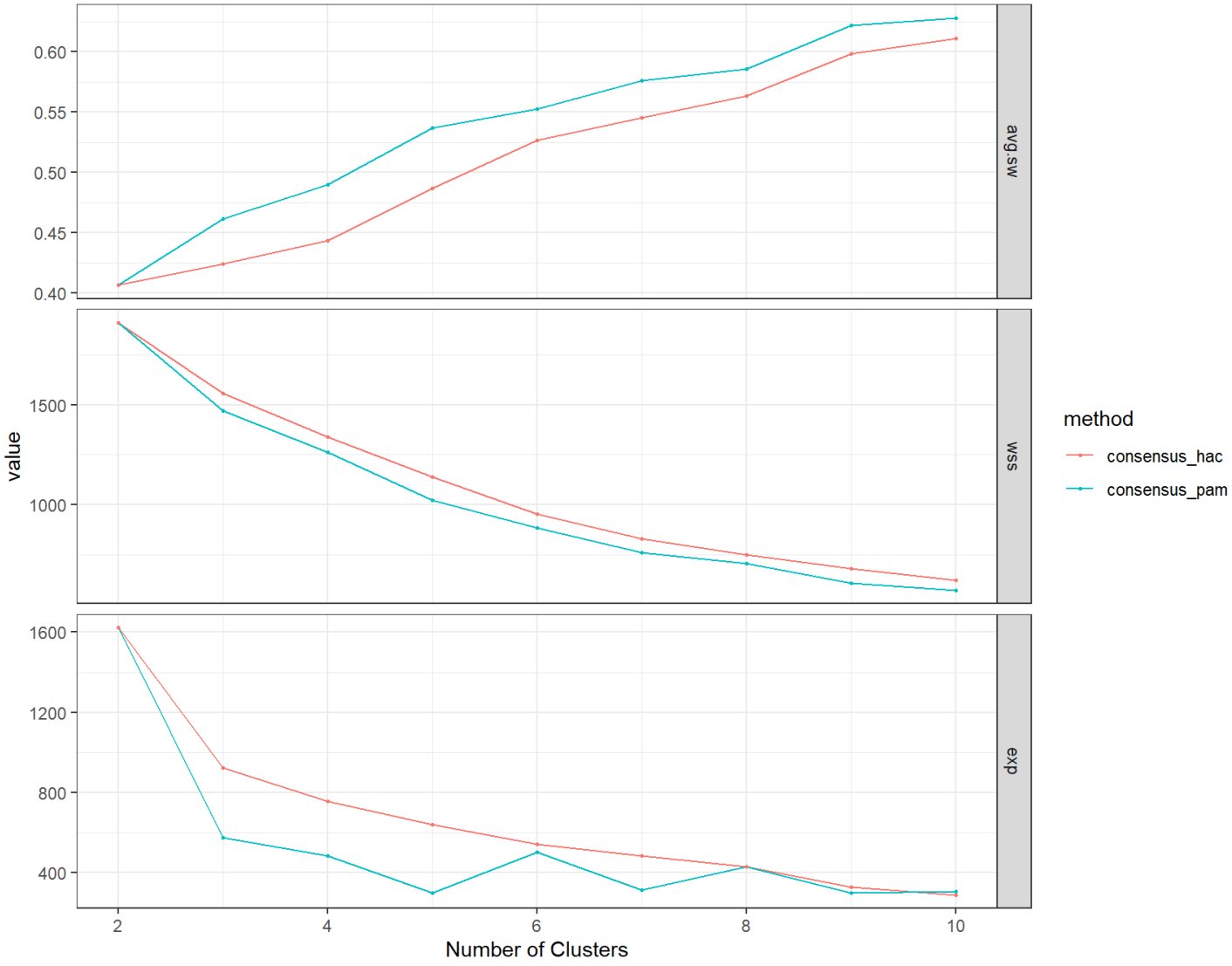
eFigure 1. Cohort creation.



eAppendix : Cluster analysis figures

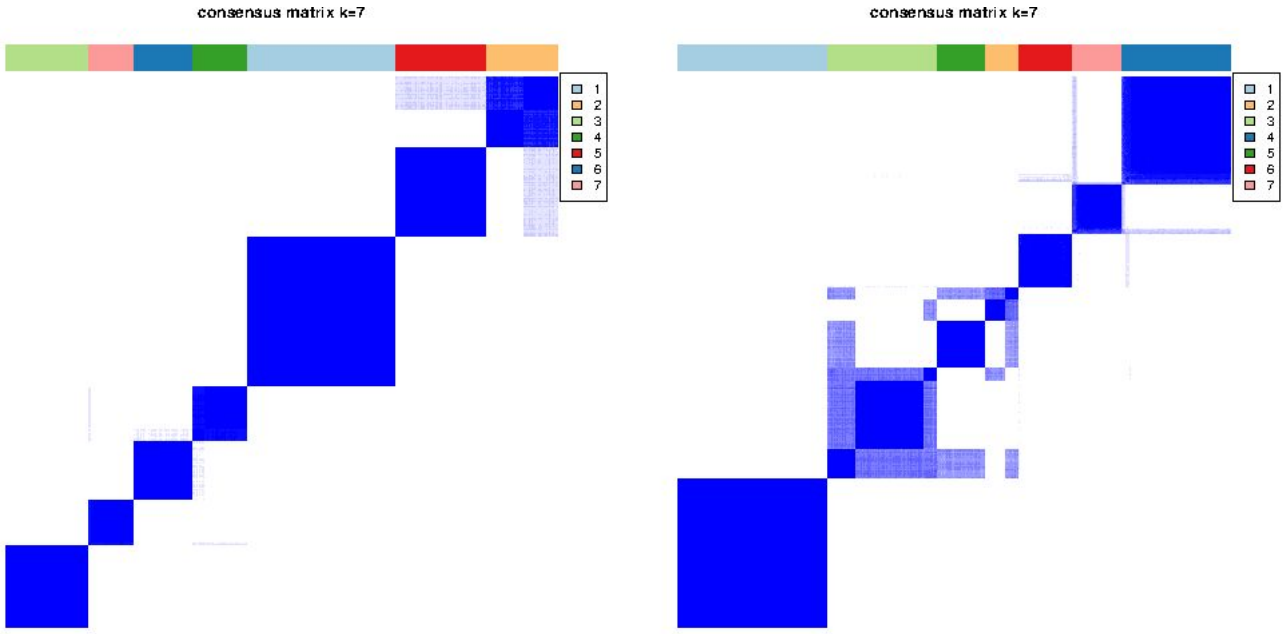
Examples of plots used when selecting the best clustering solution. For all but the first plot, the ConsensusClusterPlus package was used in R to generate plots.[6]

A. Comparison of different baseline clustering algorithms (HAC, PAM) in the derivation cohort.

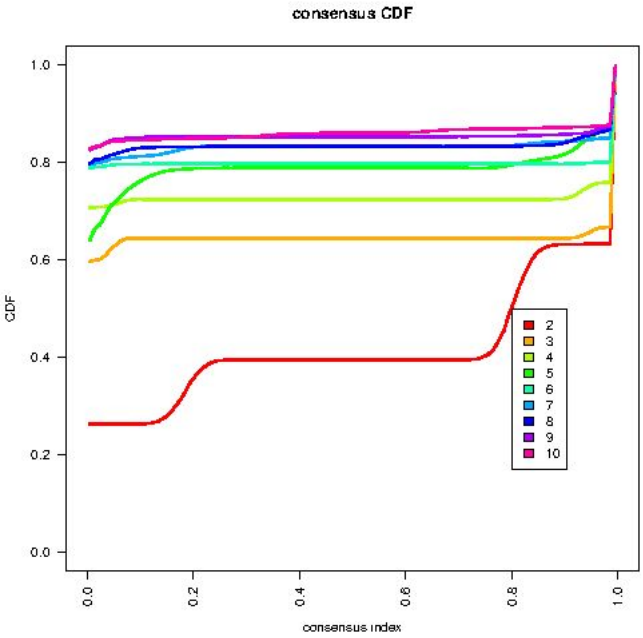


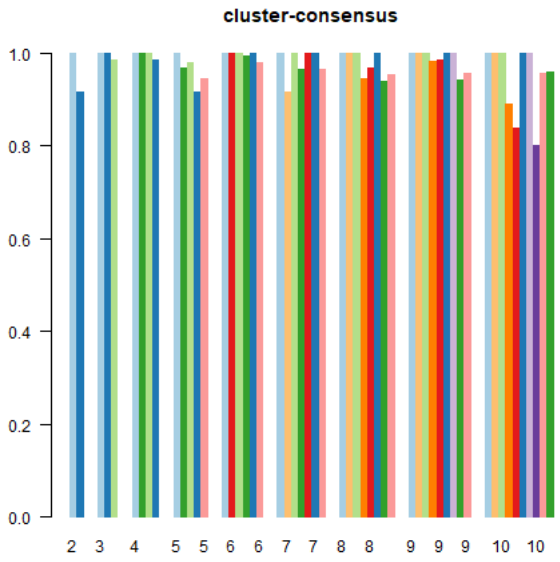
Different calculated indices used to compare algorithms for k=2-10 clusters. K-modes performed poorly overall and is not pictured here. Avg.sw=sillouette width looking to maximize, exp=expected vs observed cluster size looking to minimize, and wss=within-cluster sum of squares looking to minimize.

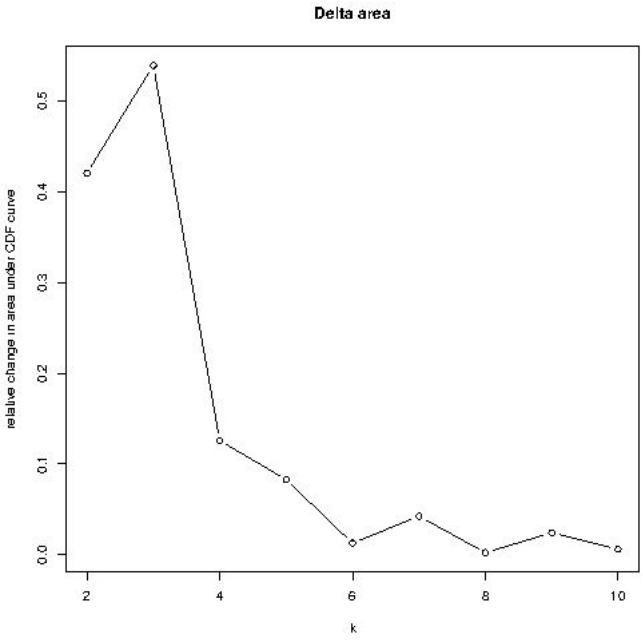
Consensus matrix heat maps for k=7 clusters in the derivation cohort. PAM left, HAC on the right.



B. Comparing k=2-10 clusters for PAM in derivation cohort

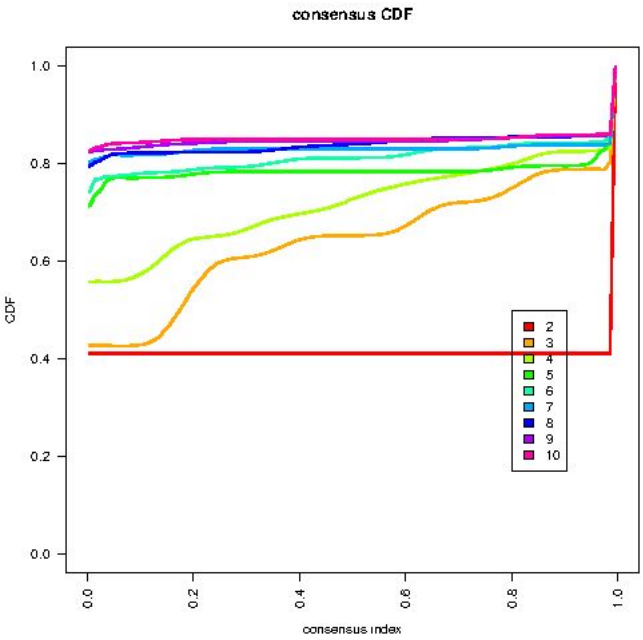


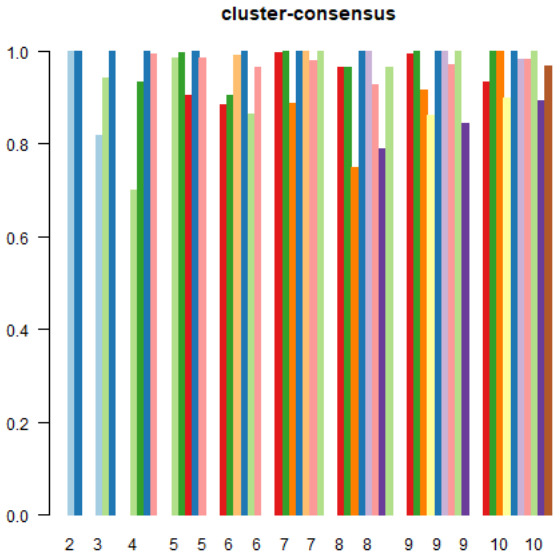


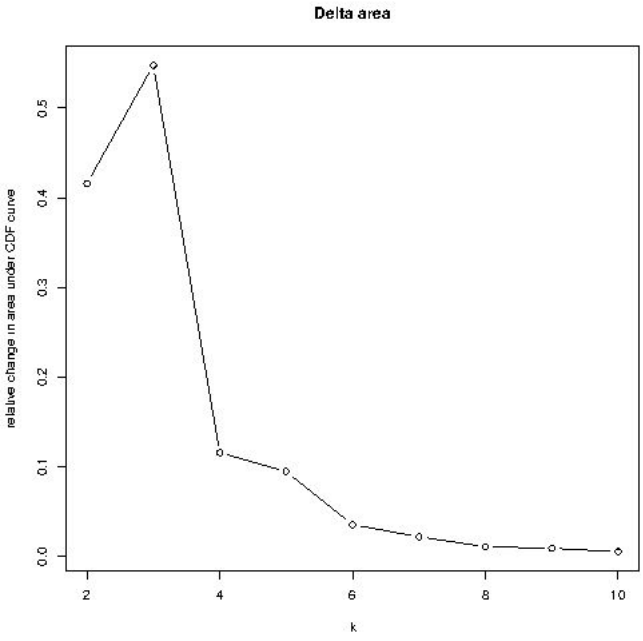


Top panel left to right: cumulative distribution function (CDF) plot looking for the number of clusters maximizing the CDF, and pairwise consensus values between clusters, looking for at least 0.8. Bottom panel: delta area for the CDF function curve, looking for the solution with the biggest change.

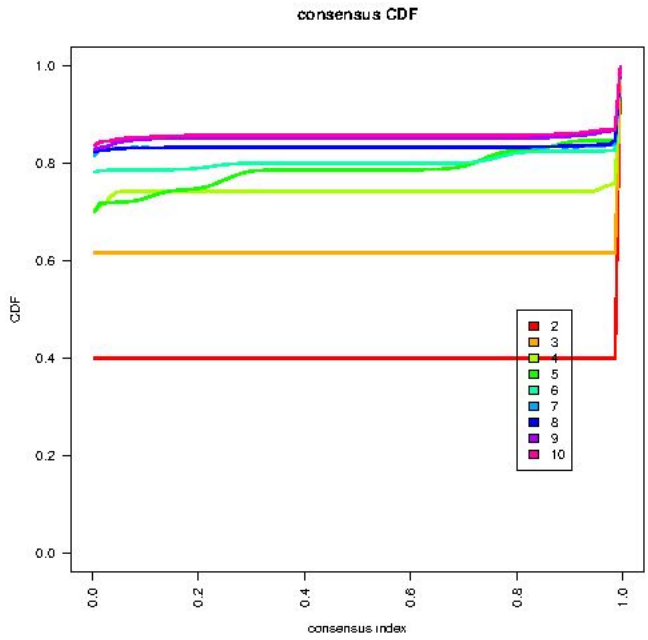
C. Comparing k=2-10 clusters for PAM in validation cohort

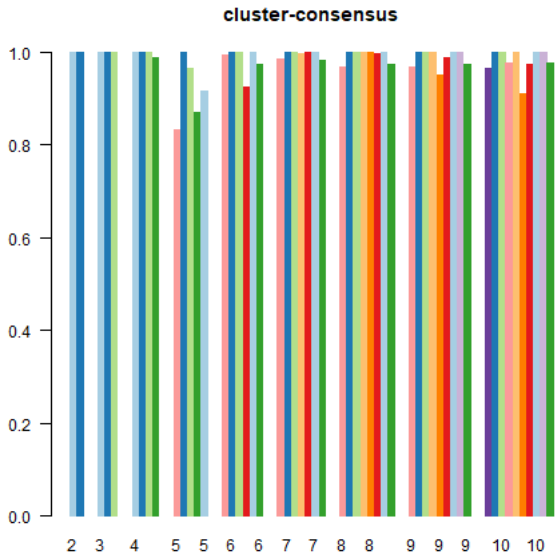


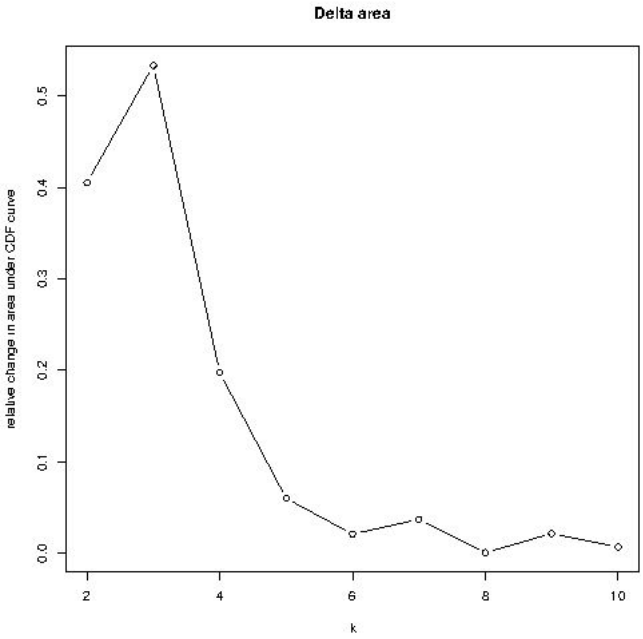




D. Comparing k=2-10 clusters for PAM in total cohort







eTable 1. Baseline characteristics and coexisting conditions for patients with community acquired pneumonia admitted to general internal medicine (2010-2017)

Baseline characteristic or coexisting condition	Overall Cohort	Derivation Cohort	Validation Cohort	Standardized Mean Difference
Number	11085	7066	4019	
Age (years, median [IQR])	79.0 [65.0, 87.0]	79.0 [66.0, 87.0]	78.0 [65.0, 87.0]	0.03
Male sex (%)	5832 (52.6)	3737 (52.9)	2095 (52.1)	0.02
From nursing home (%)	1224 (11.0)	856 (12.1)	365 (9.1)	0.1
Transport via ambulance (%)	6849 (61.8)	4433 (62.7)	2396 (59.6)	0.06
LAPS (mean (SD))	23.4 (16.9)	24.0 (17.2)	22.4 (16.3)	0.09
Charlson index (mean (SD))	1.7 (1.7)	1.7 (1.7)	1.6 (1.7)	0.03
Pulmonary (%)	3178 (28.7)	2119 (30.0)	1046 (26.0)	0.09
DM (%)	2978 (26.9)	1862 (26.4)	1113 (27.7)	0.03
CHF (%)	1892 (17.1)	1261 (17.8)	603 (15.0)	0.08
Dementia (%)	1401 (12.6)	931 (13.2)	459 (11.4)	0.05
Cancer (%)	1194 (10.8)	690 (9.8)	496 (12.3)	0.08
Renal (%)	703 (6.3)	464 (6.6)	229 (5.7)	0.04
MI (%)	512 (4.6)	367 (5.2)	133 (3.3)	0.09
Stroke (%)	364 (3.3)	285 (4.0)	76 (1.9)	0.13
Liver (%)	265 (2.4)	159 (2.3)	103 (2.6)	0.02
PVD (%)	236 (2.1)	165 (2.3)	73 (1.8)	0.04
Rheumatic (%)	211 (1.9)	143 (2.0)	71 (1.8)	0.02
Paralysis (%)	125 (1.1)	86 (1.2)	40 (1.0)	0.02
PUD (%)	75 (0.7)	48 (0.7)	23 (0.6)	0.01
HIV (%)	12 (0.1)	9 (0.1)	4 (0.1)	0.01

eTable 1 legend. Coexisting conditions were defined based on a previously published coding algorithm to define charlson comorbidities based on ICD-10 codes (see text). N refers to number of patients. Age is in years. LAPS=laboratory-based acute physiology score. Charlson score=calculated Charlson comorbidity index. Pulmonary=chronic lung disease including both obstructive and restrictive, DM= diabetes mellitus, CHF=congestive heart failure, renal=renal disease, MI=myocardial infarction, Liver= liver disease, PVD=peripheral vascular disease, Rheumatic=rheumatic disease, PUD=peptic ulcer disease.

eTable 2. Clustering solution for PAM with k=7 clusters (Derivation cohort)

	Low Comorbidity	DM-HF-Pulm	Pulmonary	Diabetes	Heart Failure	Dementia	Cancer	p
Number	1910	1149	1060	758	918	693	578	
Age (years, median [IQR])	75.0 [54.0, 86.0]	80.0 [72.0, 87.0]	77.0 [64.0, 84.0]	75.0 [66.0, 83.0]	82.0 [68.2, 89.0]	86.0 [81.0, 90.0]	72.0 [61.2, 83.0]	<0.001
Male sex (%)	958 (50.2)	626 (54.5)	571 (53.9)	437 (57.7)	469 (51.1)	328 (47.3)	348 (60.2)	<0.001
From nursing home (%)	149 (7.8)	160 (13.9)	88 (8.3)	58 (7.7)	105 (11.4)	272 (39.2)	24 (4.2)	<0.001
Transport by ambulance (%)	1051 (55.0)	768 (66.8)	655 (61.8)	454 (59.9)	589 (64.2)	627 (90.5)	289 (50.0)	<0.001
LAPS (mean (SD))	21.2 (15.5)	27.7 (19.1)	22.7 (17.4)	26.2 (16.7)	25.2 (17.8)	24.5 (16.6)	22.6 (16.3)	<0.001
Charlson score (mean (SD))	0.0 (0.0)	3.4 (1.4)	1.3 (0.8)	1.8 (1.0)	1.8 (1.1)	1.8 (1.1)	3.8 (2.0)	<0.001
Pulmonary (%)	0 (0.0)	858 (74.7)	1060 (100.0)	0 (0.0)	0 (0.0)	105 (15.2)	96 (16.6)	<0.001
DM (%)	0 (0.0)	849 (73.9)	0 (0.0)	758 (100.0)	0 (0.0)	157 (22.7)	98 (17.0)	<0.001
CHF (%)	0 (0.0)	771 (67.1)	0 (0.0)	0 (0.0)	490 (53.4)	0 (0.0)	0 (0.0)	<0.001
Dementia (%)	0 (0.0)	127 (11.1)	0 (0.0)	0 (0.0)	76 (8.3)	693 (100.0)	35 (6.1)	<0.001
Cancer (%)	0 (0.0)	73 (6.4)	0 (0.0)	0 (0.0)	39 (4.2)	0 (0.0)	578 (100.0)	<0.001
Renal (%)	0 (0.0)	97 (8.4)	57 (5.4)	52 (6.9)	183 (19.9)	49 (7.1)	26 (4.5)	<0.001
MI (%)	0 (0.0)	119 (10.4)	41 (3.9)	36 (4.7)	121 (13.2)	32 (4.6)	18 (3.1)	<0.001
Stroke (%)	0 (0.0)	56 (4.9)	24 (2.3)	38 (5.0)	98 (10.7)	50 (7.2)	19 (3.3)	<0.001
Liver (%)	0 (0.0)	23 (2.0)	34 (3.2)	19 (2.5)	70 (7.6)	4 (0.6)	9 (1.6)	<0.001
PVD (%)	0 (0.0)	48 (4.2)	27 (2.5)	17 (2.2)	50 (5.4)	11 (1.6)	12 (2.1)	<0.001
Rheumatic (%)	0 (0.0)	20 (1.7)	23 (2.2)	8 (1.1)	79 (8.6)	6 (0.9)	7 (1.2)	<0.001
Paralysis (%)	0 (0.0)	6 (0.5)	5 (0.5)	12 (1.6)	45 (4.9)	12 (1.7)	6 (1.0)	<0.001
PUD (%)	0 (0.0)	9 (0.8)	8 (0.8)	4 (0.5)	22 (2.4)	3 (0.4)	2 (0.3)	<0.001
HIV (%)	0 (0.0)	3 (0.3)	4 (0.4)	0 (0.0)	1 (0.1)	1 (0.1)	0 (0.0)	0.088

eTable 2 legend. See text for details regarding cluster analysis. Number refers to number of patients. LAPS=laboratory-based acute physiology score. Charlson score=calculated Charlson comorbidity index. Pulmonary=chronic lung disease including both obstructive and restrictive, DM= diabetes mellitus, CHF=congestive heart failure, renal=renal disease, MI=myocardial infarction, Liver= liver disease, PVD=peripheral vascular disease, Rheumatic=rheumatic disease, PUD=peptic ulcer disease. Subgroups were named by the condition(s) present in all cluster members or a large proportion if no single condition was present in 100% of the patients within a subgroup. DM-HF-Pulm= subgroup composed of a large portion of patients with diabetes, congestive heart failure and chronic lung disease. P=2-tailed p-value for differences between subgroups, determined by chi-square test for categorical variables and Kruskall-Wallis tests for continuous variables.

eTable 3. Clustering solution for PAM with k=7 clusters (Validation Cohort)

	Low Comorbidity	DM-HF-Pulm	Pulmonary	Diabetes	Heart Failure	Dementia	Cancer	p
Number	1156	533	567	532	456	343	432	
Age (years, median [IQR])	75.0 [57.0, 86.0]	80.0 [69.0, 86.0]	77.0 [63.5, 85.0]	75.0 [65.0, 83.0]	83.0 [69.0, 90.0]	87.0 [82.0, 91.0]	71.0 [62.0, 80.0]	<0.001
Male sex (%)	585 (50.6)	293 (55.0)	278 (49.0)	310 (58.3)	224 (49.1)	143 (41.7)	262 (60.6)	<0.001
From nursing home (%)	68 (5.9)	75 (14.1)	24 (4.2)	42 (7.9)	35 (7.7)	113 (32.9)	8 (1.9)	<0.001
Transport by ambulance (%)	638 (55.2)	347 (65.1)	342 (60.3)	320 (60.2)	274 (60.1)	301 (87.8)	174 (40.3)	<0.001
LAPS (mean (SD))	20.0 (14.8)	25.7 (17.8)	20.5 (16.4)	25.2 (16.1)	25.7 (17.7)	24.0 (15.4)	19.5 (15.4)	<0.001
Charlson score (mean (SD))	0.0 (0.0)	3.3 (1.4)	1.2 (0.7)	1.8 (0.9)	1.8 (1.1)	1.6 (0.9)	3.8 (2.0)	<0.001
DM (%)	0 (0.0)	412 (77.3)	0 (0.0)	532 (100.0)	0 (0.0)	89 (25.9)	80 (18.5)	<0.001
Pulmonary (%)	0 (0.0)	371 (69.6)	567 (100.0)	0 (0.0)	0 (0.0)	37 (10.8)	71 (16.4)	<0.001
CHF (%)	0 (0.0)	345 (64.7)	0 (0.0)	0 (0.0)	258 (56.6)	0 (0.0)	0 (0.0)	<0.001
Cancer (%)	0 (0.0)	45 (8.4)	0 (0.0)	0 (0.0)	19 (4.2)	0 (0.0)	432 (100.0)	<0.001
Dementia (%)	0 (0.0)	57 (10.7)	0 (0.0)	0 (0.0)	37 (8.1)	343 (100.0)	22 (5.1)	<0.001
Renal (%)	0 (0.0)	33 (6.2)	26 (4.6)	21 (3.9)	113 (24.8)	13 (3.8)	23 (5.3)	<0.001
MI (%)	0 (0.0)	40 (7.5)	16 (2.8)	20 (3.8)	39 (8.6)	9 (2.6)	9 (2.1)	<0.001
Stroke (%)	0 (0.0)	15 (2.8)	10 (1.8)	12 (2.3)	29 (6.4)	4 (1.2)	6 (1.4)	<0.001
Liver (%)	0 (0.0)	14 (2.6)	19 (3.4)	14 (2.6)	38 (8.3)	2 (0.6)	16 (3.7)	<0.001
PVD (%)	0 (0.0)	22 (4.1)	4 (0.7)	10 (1.9)	21 (4.6)	6 (1.7)	10 (2.3)	<0.001
Rheumatic (%)	0 (0.0)	4 (0.8)	10 (1.8)	11 (2.1)	38 (8.3)	2 (0.6)	6 (1.4)	<0.001
Paralysis (%)	0 (0.0)	5 (0.9)	3 (0.5)	10 (1.9)	18 (3.9)	3 (0.9)	1 (0.2)	<0.001
PUD (%)	0 (0.0)	2 (0.4)	4 (0.7)	4 (0.8)	9 (2.0)	2 (0.6)	2 (0.5)	0.001
HIV (%)	0 (0.0)	0 (0.0)	3 (0.5)	1 (0.2)	0 (0.0)	0 (0.0)	0 (0.0)	0.031

eTable 3 legend. See text for details regarding cluster analysis. Number refers to number of patients. LAPS=laboratory-based acute physiology score. Charlson score=calculated Charlson comorbidity index. Pulmonary=chronic lung disease including both obstructive and restrictive, DM= diabetes mellitus, CHF=congestive heart failure, renal=renal disease, MI=myocardial infarction, Liver= liver disease, PVD=peripheral vascular disease, Rheumatic=rheumatic disease, PUD=peptic ulcer disease. Subgroups were named by the condition(s) present in all cluster members or a large proportion if no single condition was present in 100% of the patients within a subgroup. DM-HF-Pulm= subgroup composed of a large portion of patients with diabetes, congestive heart failure and chronic lung disease. P=2-tailed p-value for differences between subgroups, determined by chi-square test for categorical variables and Kruskall-Wallis tests for continuous variables.

References

1. Seymour CW, Kennedy JN, Wang S, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA.* 2019;321 (20):2003-2017.
2. Vranas KC, Jopling JK, Sweeney TE, et al. Identifying Distinct Subgroups of ICU Patients: A Machine Learning Approach. *Crit Care Med.* 2017;45(10):1607-1615.
3. Huang Z. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery.* 1998;2(3):283–304.
4. Thrun MC. Approaches to Cluster Analysis. In: *Projection-Based Clustering through Self-Organization and Swarm Intelligence.*2018.
5. J.Rousseeuw P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics.* 1987;20:53-65.
6. Wilkerson MD. ConsensusClusterPlus (Tutorial). https://bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf. Published 2020. Accessed.