

# Supporting information for “Identifying clusters of coexisting conditions and outcomes among adults admitted to hospital with community-acquired pneumonia: A multicentre cohort study”

Sarah L. Malecki, Hae Young Jung, Anne Loffler, 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, p. 2**

**eResults, p. 2**

**eFigure 1. Cohort Creation, p. 4**

**eAppendix. Cluster analysis figures, p. 5**

**eTable 1. Baseline characteristics and coexisting conditions for patients with community acquired pneumonia admitted to general internal medicine (2010-2017), p. 10**

**eTable 2. Baseline characteristics and coexisting conditions for patients with community acquired pneumonia admitted to general internal medicine by hospital (2010-2017), p. 11**

**eTable 3. Clustering solution for PAM with k=7 clusters (Derivation cohort), p. 12**

**eTable 4. Clustering solution for PAM with k=7 clusters (Replication Cohort), p. 13**

**eTable 5. Clustering solution for HAC with k=7 clusters (Derivation cohort), p. 14**

**eTable 6. Clustering solution for k-modes with k=7 clusters (Derivation cohort), p. 15**

**eTable 7. Sensitivity analysis: Association of patient subgroup based on coexisting conditions with clinical outcomes after multivariable adjustment, including adjustment for secondary comorbidities, p. 16**

**References, p. 17**

## eMethods

### Cluster analysis

Machine learning is the concept of applying computer-based algorithms to large amounts of data to understand and reformulate underlying data structure into useful output for the end user. It has many useful applications to healthcare in the era of increasing amounts of data available through electronic information systems. Cluster analysis is the concept of separating unorganized data into categories based on similarities and differences in different characteristics.<sup>1</sup>

There are several approaches to cluster analysis for clinical data, each with their own set of strengths and limitations.<sup>2</sup> Unsupervised approaches are useful for exploring patterns in the data without making any underlying assumptions about data structure.<sup>3</sup> Partitioning approaches and hierarchical clustering approaches are common unsupervised approaches that share the strength of being easy to implement and interpret, but are limited by their sensitivity to outliers.<sup>2</sup> One way to address this limitation is to use a consensus cluster analysis approach, which has been employed to describe sepsis phenotypes<sup>4</sup> and ICU subgroups<sup>3</sup> in recent studies. 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 therefore used a consensus cluster analysis approach to derive clusters in our derivation cohort. We compared the performance of three different baseline unsupervised clustering algorithms (K-modes,<sup>5</sup> partitioning around medoids [PAM]<sup>6</sup> and hierarchical agglomerative clustering [HAC]).<sup>6</sup> These three algorithms were selected because they could each be implemented with binary data<sup>7</sup> (an important limitation of other 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 and qualitative reproducibility (with each cluster defined by the condition(s) shared by the majority of patients, or the highest prevalence condition(s) in the cluster, following the convention of assessing for broad similarity given that exact quantitative matches are near impossible<sup>8</sup>), minimise any bias introduced by relying on a single method and select the approach that performed best with our data. Models were run using a modified version of the R package “ConsensusClusterPlus”.<sup>9</sup> For the K-modes algorithm, we optimized it for our asymmetric binary data by changing the simple distance measure to the Jaccard distance.<sup>7</sup> 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,<sup>10</sup> we examined numerous measures and visualizations to select 1) the best-performing clustering algorithm overall, and 2) 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,<sup>4</sup> best fit was determined by examination of characteristics of consensus cumulative distribution function plots<sup>4</sup> and consensus matrix heat maps to select a solution that maximized separation of clusters.<sup>4</sup> We ensured that pairwise consensus values between cluster members was  $>0.8$ .<sup>4</sup> We also calculated and plotted eight common indices used to assess cluster analysis performance. These included silhouette width, pearsons’s gamma, The Dunn index, VI index, generalized Calinski and Harabasz index, within-between cluster ratio, within-cluster sum of squares, and expected versus observed cluster size. All indices were calculated using the flexible procedures for clustering package in R<sup>11</sup>, except for expected versus observed cluster size<sup>8</sup> which was calculated manually, assuming equal sized clusters in attempts to avoid small clusters which may represent outliers.

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 respiratory 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. Specifically, clinicians were asked to reflect on whether the patterns of comorbidity in the clustering solutions were recognizable based on the patients they see in clinical practice and whether the patterns of comorbidity in different clusters might affect clinical decision-making with respect to the investigation and treatment of patients with CAP.

## eResults

### Cluster Analysis

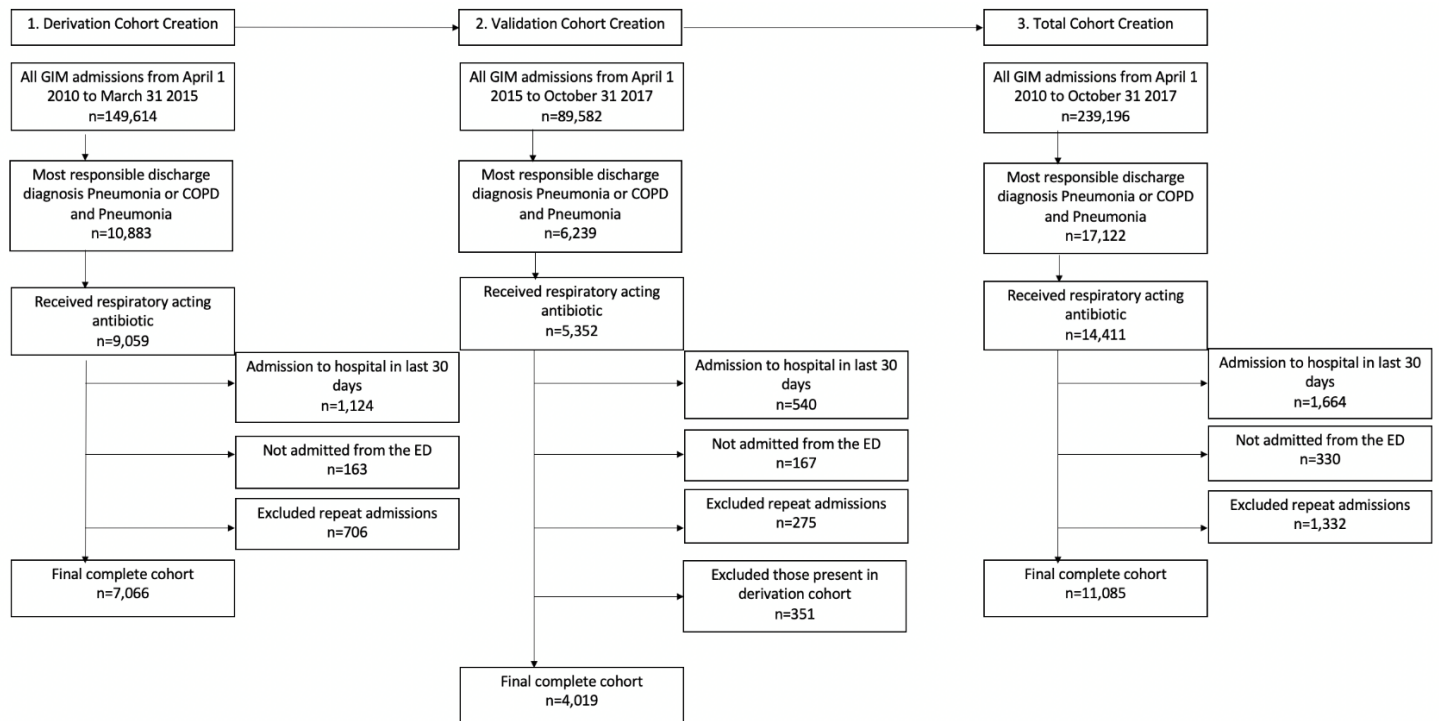
Based on examination of the consensus plots and additional indices to evaluate different clustering solutions (see eAppendix), PAM was selected as the method of choice because it yielded a better clustering solution than K-modes or HAC, regardless of the number of clusters chosen. HAC performed second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

For the derivation cohort, clustering solution PAM  $k=7$  was selected as the best overall solution based on objective indices, reproducibility and clinical relevance. Candidate clustering solutions with reasonable performance on objective indices, including

PAM k=6, k=7 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 also qualitatively reproducible in the validation cohort (eTable 3 and 4). Therefore, k=7 clusters was selected as the most clinically relevant and reproducible clustering solution.

Qualitatively, the clusters obtained in the PAM k=7 solution were reproduced in the HAC k=7 solution (eTable 5), with the exception that the highest prevalence condition in the multi-morbid group was ~30% and that this group also included a relatively high prevalence of renal disease and MI. Five of the PAM k=7 clusters (HF, Pulm, DM, dementia and cancer) were reproduced in the k-modes algorithm (eTable 6). Similar to other approaches, there was a group with relatively few comorbidities and one where no single condition predominated, but the separation between groups was not as clean and the highest prevalence conditions in the disproportionately smaller multi-morbid group composed of 39 patients were renal disease, stroke and MI.

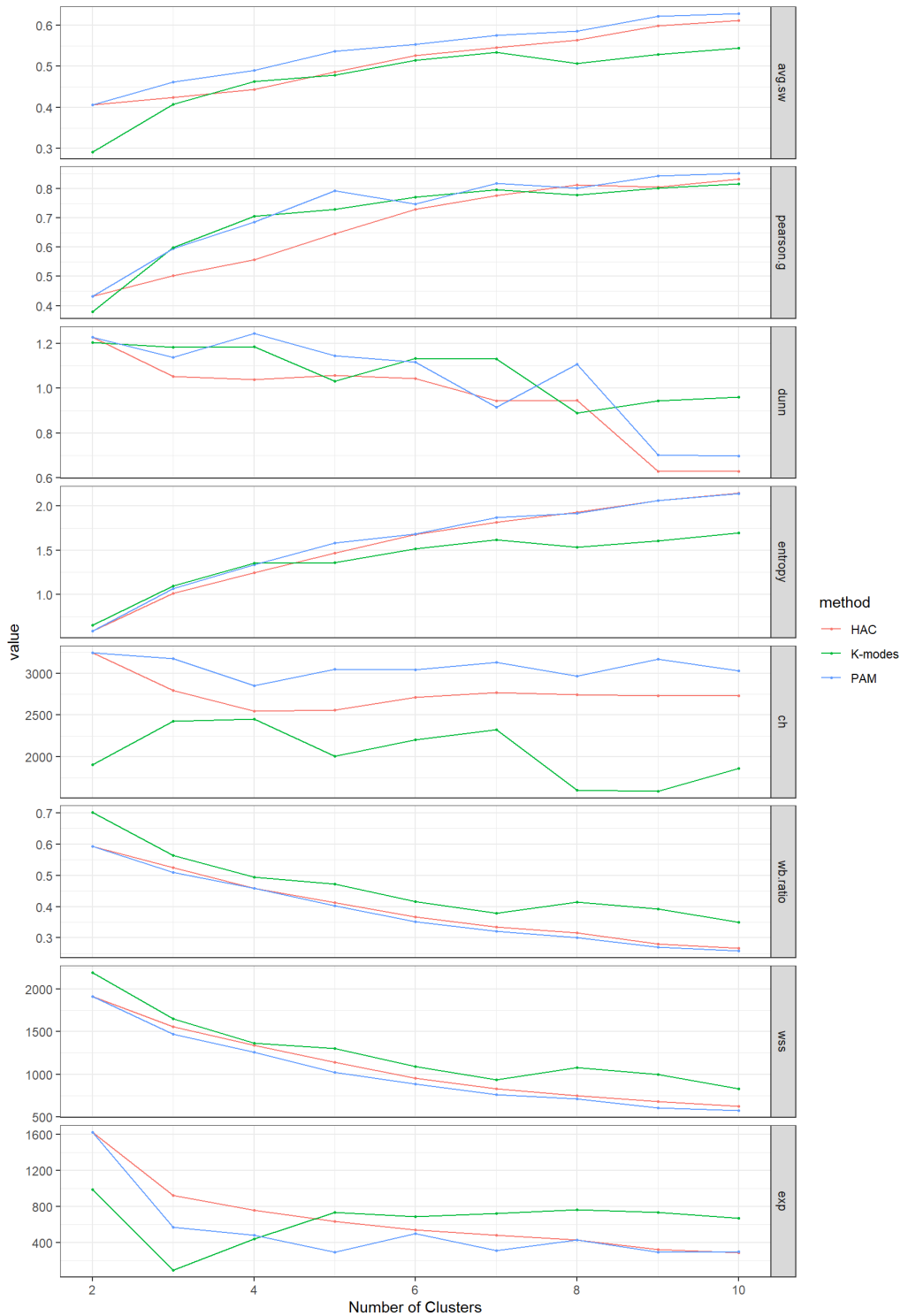
**eFigure 1. Cohort creation.**



## eAppendix : Cluster analysis figures

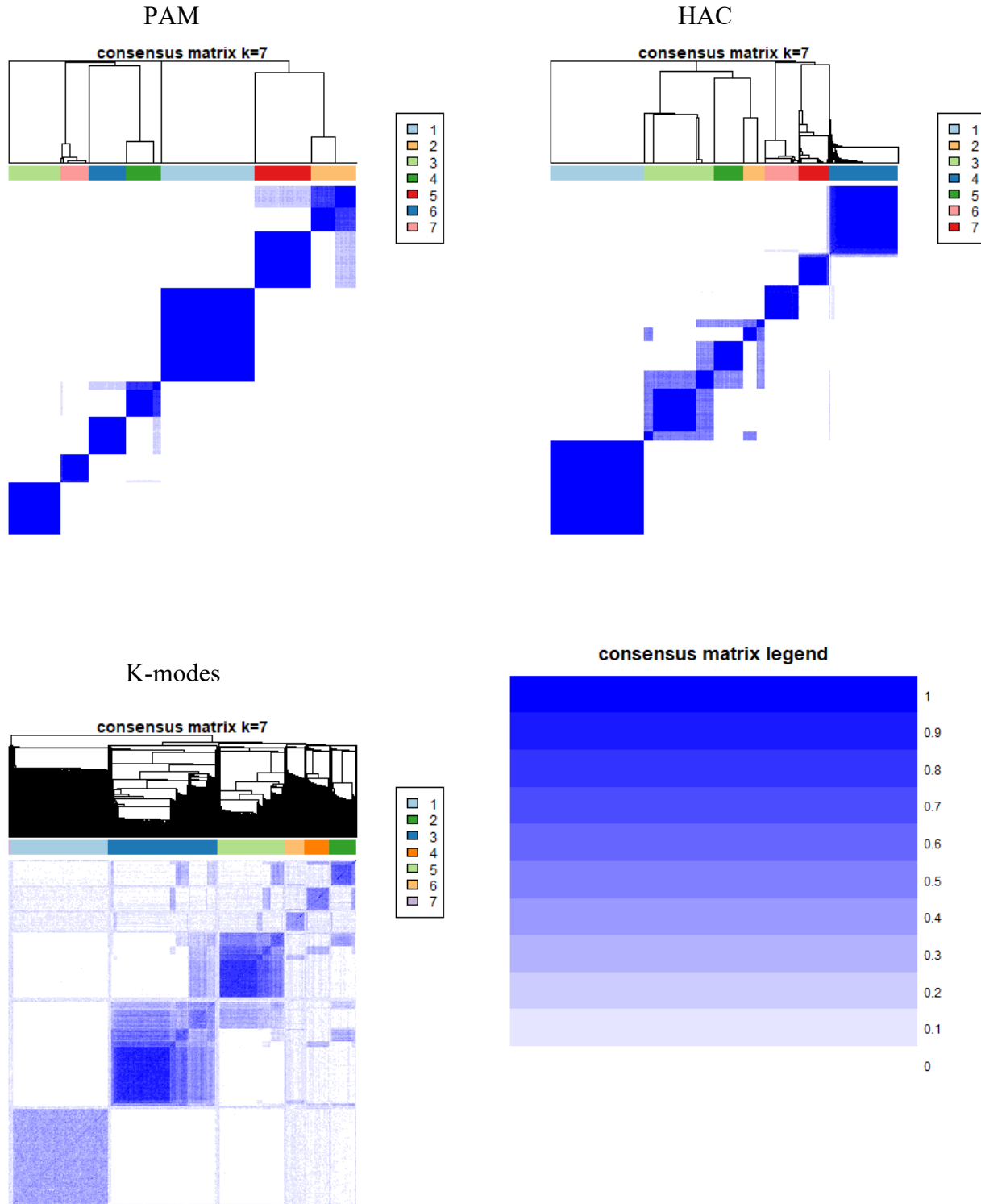
Examples of plots used when selecting the best clustering solution. For all but the first plot, the ConsensusClusterPlus package was used in R to generate plots.<sup>12</sup>

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

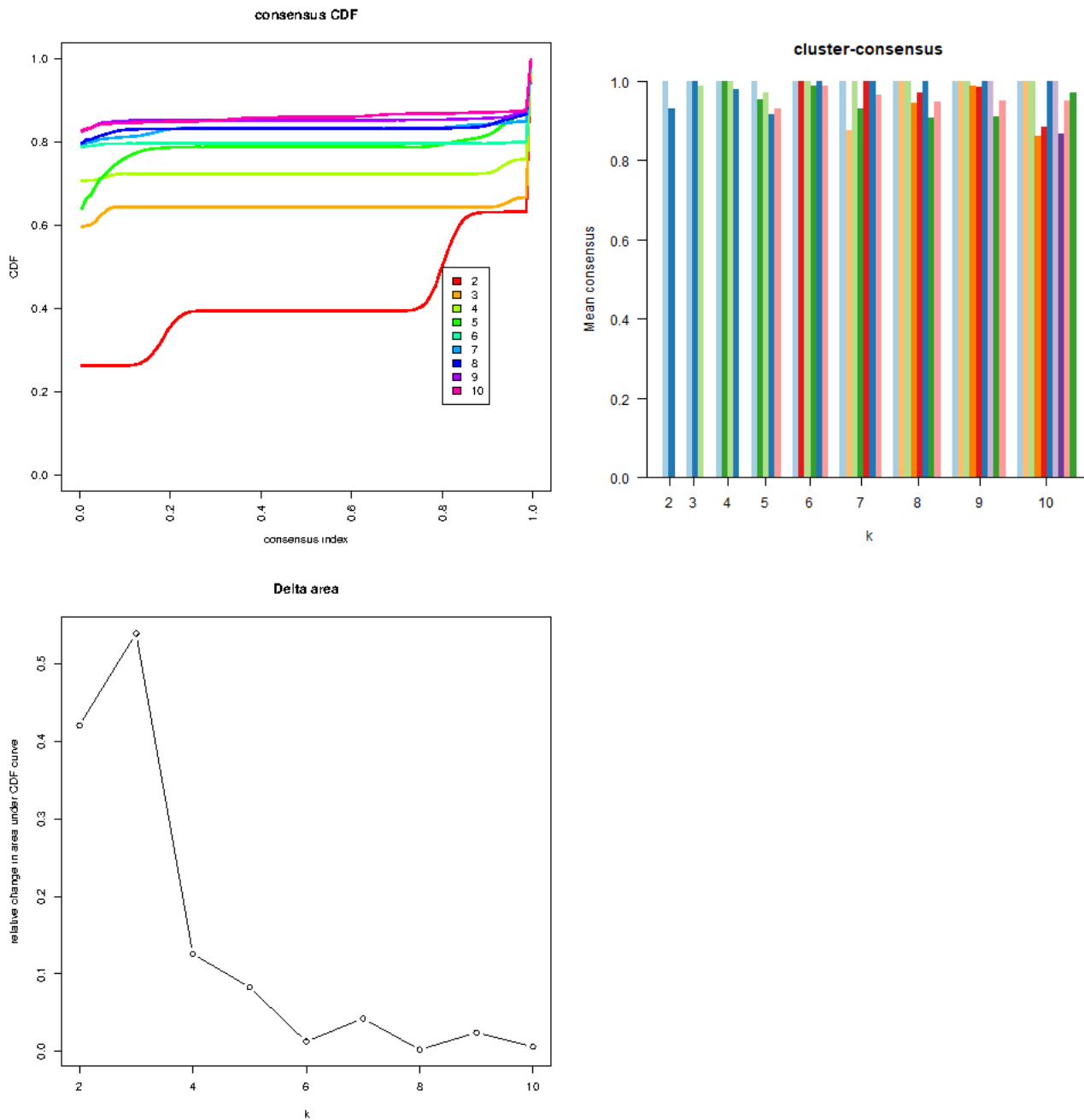


Different calculated indices used to compare algorithms for k=2-10 clusters. Avg.sw=silhouette width looking to maximize, pearson.g=pearson gamma, looking to maximize, dunn2=Dunn index, looking to maximize, entropy=VI index, looking to maximize, ch=Generalized Calinski and Harabasz index, looking to maximize, wb.ratio=within-between cluster ratio, looking to minimize, wss=within-cluster sum of squares looking to minimize, exp=expected vs observed cluster size looking to minimize. All indices were calculated using the flexible procedures for clustering package in R<sup>11</sup>, except for the exp index, which was calculated manually.

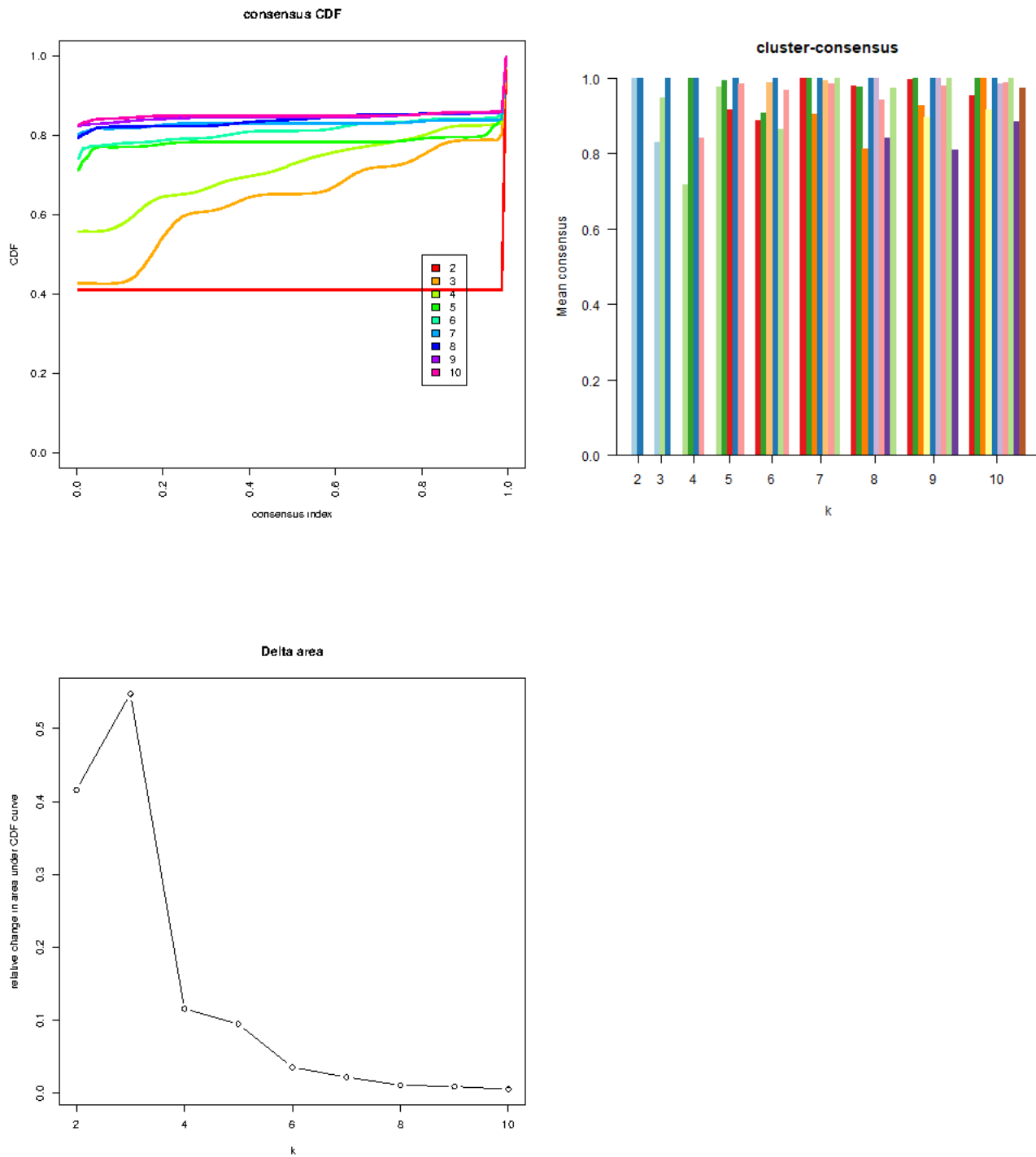
B. Consensus matrix heat maps for k=7 clusters in the derivation cohort. PAM top left, HAC on the right, K-modes bottom. The rows and columns in each heatmap refer to a given patient. The consensus values indicate the proportion of cluster iterations in which each pair of patients was grouped into the same cluster. Consensus values range from 0 (white; never clustered together) to 1 (dark blue; always clustered together). The matrices are ordered by the consensus clustering, which is shown as a dendrogram above each heatmap.



**C.** 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.

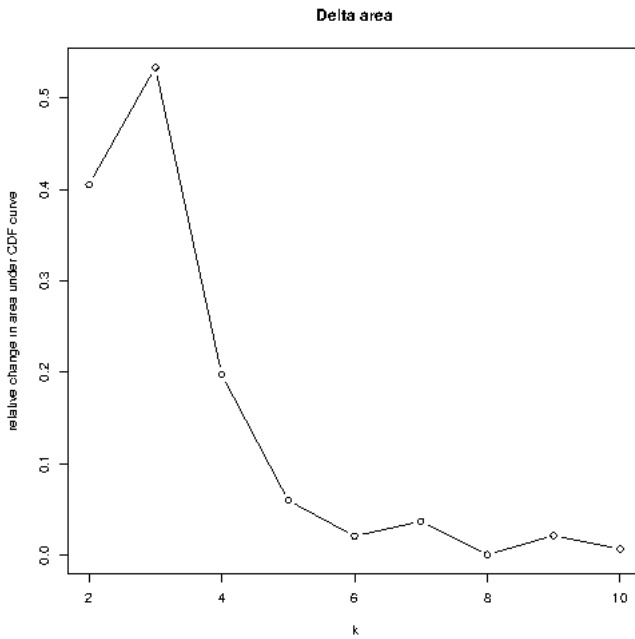
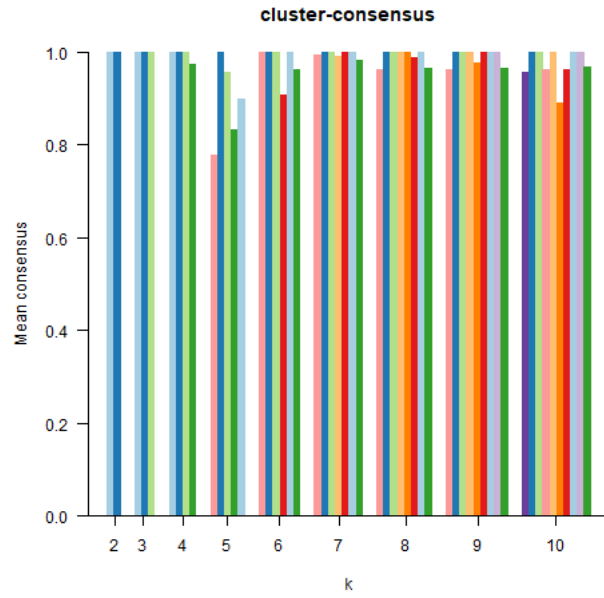
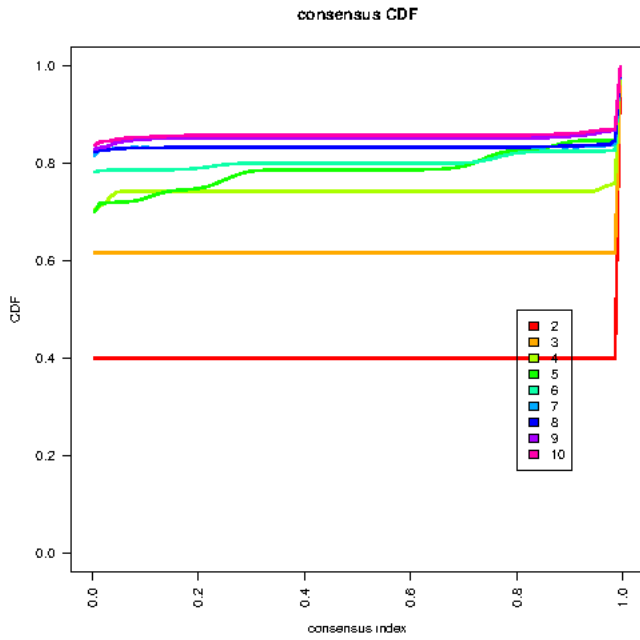


#### D. Comparing k=2-10 clusters for PAM in replication cohort





### E. 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
Charlson index categories (%)				
0	3052 (27.5)	1910 (27.0)	1156 (28.8)	0.04
1	3209 (28.9)	2087 (29.5)	1138 (28.3)	0.03
2	2191 (19.8)	1353 (19.1)	837 (20.8)	0.04
3	1266 (11.4)	817 (11.6)	439 (10.9)	0.02
4+	1367 (12.3)	899 (12.7)	449 (11.2)	0.05
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. Baseline characteristics and coexisting conditions for patients with community acquired pneumonia admitted to general internal medicine (2010-2017), by hospital

	<b>Overall</b>	<b>Hospital A</b>	<b>Hospital B</b>	<b>Hospital C</b>	<b>Hospital D</b>	<b>Hospital E</b>	<b>Hospital F</b>	<b>Hospital G</b>
Number of patients	11085	1782	1514	1613	1717	2047	1181	1231
Charlson Comorbidity Index (mean [SD])	1.7 (1.7)	2.3 (2.0)	1.7 (1.6)	2.0 (1.7)	1.6 (1.5)	1.3 (1.5)	1.4 (1.6)	1.3 (1.4)
Charlson Comorbidity Index, categories (%)								
0	3052 (27.5)	318 (17.8)	360 (23.8)	318 (19.7)	466 (27.1)	703 (34.3)	422 (35.7)	465 (37.8)
1	3209 (28.9)	409 (23.0)	457 (30.2)	464 (28.8)	541 (31.5)	634 (31.0)	359 (30.4)	345 (28.0)
2	2191 (19.8)	442 (24.8)	340 (22.5)	317 (19.7)	330 (19.2)	342 (16.7)	210 (17.8)	210 (17.1)
3	1266 (11.4)	230 (12.9)	190 (12.5)	245 (15.2)	203 (11.8)	205 (10.0)	73 ( 6.2)	120 ( 9.7)
4+	1367 (12.3)	383 (21.5)	167 (11.0)	269 (16.7)	177 (10.3)	163 ( 8.0)	117 ( 9.9)	91 ( 7.4)
Pulmonary (%)	3178 (28.7)	503 (28.2)	385 (25.4)	595 (36.9)	520 (30.3)	499 (24.4)	317 (26.8)	359 (29.2)
DM (%)	2978 (26.9)	444 (24.9)	535 (35.3)	469 (29.1)	501 (29.2)	475 (23.2)	245 (20.7)	309 (25.1)
CHF (%)	1892 (17.1)	283 (15.9)	272 (18.0)	359 (22.3)	290 (16.9)	384 (18.8)	150 (12.7)	154 (12.5)
Dementia (%)	1401 (12.6)	187 (10.5)	247 (16.3)	303 (18.8)	215 (12.5)	246 (12.0)	143 (12.1)	60 ( 4.9)
Cancer (%)	1194 (10.8)	502 (28.2)	126 ( 8.3)	144 ( 8.9)	147 ( 8.6)	124 ( 6.1)	101 ( 8.6)	50 ( 4.1)
Renal (%)	703 ( 6.3)	160 ( 9.0)	89 ( 5.9)	76 ( 4.7)	121 ( 7.0)	110 ( 5.4)	58 ( 4.9)	89 ( 7.2)
MI (%)	512 ( 4.6)	114 ( 6.4)	52 ( 3.4)	130 ( 8.1)	64 ( 3.7)	81 ( 4.0)	31 ( 2.6)	40 ( 3.2)
Stroke (%)	364 ( 3.3)	89 ( 5.0)	25 ( 1.7)	147 ( 9.1)	32 ( 1.9)	26 ( 1.3)	29 ( 2.5)	16 ( 1.3)
Liver (%)	265 ( 2.4)	71 ( 4.0)	14 ( 0.9)	82 ( 5.1)	27 ( 1.6)	23 ( 1.1)	18 ( 1.5)	30 ( 2.4)
PVD (%)	236 ( 2.1)	73 ( 4.1)	7 ( 0.5)	67 ( 4.2)	25 ( 1.5)	48 ( 2.3)	8 ( 0.7)	8 ( 0.6)
Rheumatic (%)	211 ( 1.9)	50 ( 2.8)	24 ( 1.6)	48 ( 3.0)	29 ( 1.7)	28 ( 1.4)	18 ( 1.5)	14 ( 1.1)
Paralysis (%)	125 ( 1.1)	18 ( 1.0)	17 ( 1.1)	34 ( 2.1)	13 ( 0.8)	21 ( 1.0)	10 ( 0.8)	12 ( 1.0)
PUD (%)	75 ( 0.7)	12 ( 0.7)	4 ( 0.3)	32 ( 2.0)	4 ( 0.2)	14 ( 0.7)	1 ( 0.1)	8 ( 0.6)
HIV (%)	12 ( 0.1)	6 ( 0.3)	0 ( 0.0)	0 ( 0.0)	1 ( 0.1)	0 ( 0.0)	2 ( 0.2)	3 ( 0.2)

eTable 2 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 3. Clustering solution for PAM with k=7 clusters (Derivation cohort)

	<b>Low Comorbidity</b>	<b>DM-HF-Pulm</b>	<b>Pulmonary</b>	<b>Diabetes</b>	<b>Heart Failure</b>	<b>Dementia</b>	<b>Cancer</b>	<b>p</b>
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 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 Kruskal-Wallis tests for continuous variables.

eTable 4. Clustering solution for PAM with k=7 clusters (Replication Cohort)

	<b>Low Comorbidity</b>	<b>DM-HF-Pulm</b>	<b>Pulmonary</b>	<b>Diabetes</b>	<b>Heart Failure</b>	<b>Dementia</b>	<b>Cancer</b>	<b>p</b>
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 4 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 Kruskal-Wallis tests for continuous variables.

eTable 5. Clustering solution for HAC with k=7 clusters (Derivation cohort)

	<b>Overall</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
n	7066	1910	435	1415	1395	604	683	624
MI (%)	367 ( 5.2)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	343 (24.6)	0 ( 0.0)	0 ( 0.0)	24 ( 3.8)
CHF (%)	1261 (17.8)	0 ( 0.0)	435 (100.0)	287 ( 20.3)	363 (26.0)	0 ( 0.0)	109 ( 16.0)	67 ( 10.7)
PVD (%)	165 ( 2.3)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	162 (11.6)	0 ( 0.0)	0 ( 0.0)	3 ( 0.5)
Stroke (%)	285 ( 4.0)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	282 (20.2)	0 ( 0.0)	0 ( 0.0)	3 ( 0.5)
Dementia (%)	931 (13.2)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	205 (14.7)	0 ( 0.0)	683 (100.0)	43 ( 6.9)
Pulmonary (%)	2119 (30.0)	0 ( 0.0)	0 ( 0.0)	1415 (100.0)	438 (31.4)	0 ( 0.0)	134 ( 19.6)	132 ( 21.2)
Rheumatic (%)	143 ( 2.0)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	143 (10.3)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)
PUD (%)	48 ( 0.7)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	46 ( 3.3)	0 ( 0.0)	1 ( 0.1)	1 ( 0.2)
DM (%)	1862 (26.4)	0 ( 0.0)	167 ( 38.4)	372 ( 26.3)	411 (29.5)	604 (100.0)	174 ( 25.5)	134 ( 21.5)
Paralysis (%)	86 ( 1.2)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	86 ( 6.2)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)
Renal (%)	464 ( 6.6)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	434 (31.1)	0 ( 0.0)	0 ( 0.0)	30 ( 4.8)
Cancer (%)	690 ( 9.8)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	66 ( 4.7)	0 ( 0.0)	0 ( 0.0)	624 (100.0)
Liver (%)	159 ( 2.3)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	159 (11.4)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)
HIV (%)	9 ( 0.1)	0 ( 0.0)	0 ( 0.0)	2 ( 0.1)	6 ( 0.4)	0 ( 0.0)	0 ( 0.0)	1 ( 0.2)
age65 (%)	5394 (76.3)	1205 (63.1)	391 ( 89.9)	1115 ( 78.8)	1104 (79.1)	462 ( 76.5)	679 ( 99.4)	438 ( 70.2)
gender_male (%)	3737 (52.9)	958 (50.2)	205 ( 47.1)	772 ( 54.6)	788 (56.5)	336 ( 55.6)	304 ( 44.5)	374 ( 59.9)

eTable 5 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. 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.

e Table 6. Clustering solution for k-modes with k=7 clusters (Derivation cohort)

	<b>Overall</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>
n	7066	1976	557	2220	497	1376	401	39
MI (%)	367 ( 5.2)	15 ( 0.8)	60 (10.8)	148 ( 6.7)	19 ( 3.8)	99 ( 7.2)	18 ( 4.5)	8 (20.5)
CHF (%)	1261 (17.8)	0 ( 0.0)	490 (88.0)	480 (21.6)	0 ( 0.0)	291 (21.1)	0 ( 0.0)	0 ( 0.0)
PVD (%)	165 ( 2.3)	8 ( 0.4)	25 ( 4.5)	76 ( 3.4)	10 ( 2.0)	38 ( 2.8)	7 ( 1.7)	1 ( 2.6)
Stroke (%)	285 ( 4.0)	19 ( 1.0)	46 ( 8.3)	91 ( 4.1)	33 ( 6.6)	77 ( 5.6)	11 ( 2.7)	8 (20.5)
Dementia (%)	931 (13.2)	0 ( 0.0)	76 (13.6)	197 ( 8.9)	459 (92.4)	199 (14.5)	0 ( 0.0)	0 ( 0.0)
Pulmonary (%)	2119 (30.0)	0 ( 0.0)	0 ( 0.0)	2119 (95.5)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)	0 ( 0.0)
Rheumatic (%)	143 ( 2.0)	4 ( 0.2)	19 ( 3.4)	68 ( 3.1)	11 ( 2.2)	24 ( 1.7)	14 ( 3.5)	3 ( 7.7)
PUD (%)	48 ( 0.7)	3 ( 0.2)	8 ( 1.4)	20 ( 0.9)	1 ( 0.2)	9 ( 0.7)	5 ( 1.2)	2 ( 5.1)
DM (%)	1862 (26.4)	0 ( 0.0)	0 ( 0.0)	558 (25.1)	0 ( 0.0)	1304 (94.8)	0 ( 0.0)	0 ( 0.0)
Paralysis (%)	86 ( 1.2)	7 ( 0.4)	8 ( 1.4)	16 ( 0.7)	21 ( 4.2)	26 ( 1.9)	7 ( 1.7)	1 ( 2.6)
Renal (%)	464 ( 6.6)	20 ( 1.0)	79 (14.2)	150 ( 6.8)	42 ( 8.5)	131 ( 9.5)	26 ( 6.5)	16 (41.0)
Cancer (%)	690 ( 9.8)	0 ( 0.0)	39 ( 7.0)	156 ( 7.0)	28 ( 5.6)	111 ( 8.1)	356 (88.8)	0 ( 0.0)
Liver (%)	159 ( 2.3)	3 ( 0.2)	16 ( 2.9)	78 ( 3.5)	8 ( 1.6)	38 ( 2.8)	12 ( 3.0)	4 (10.3)
HIV (%)	9 ( 0.1)	0 ( 0.0)	0 ( 0.0)	8 ( 0.4)	0 ( 0.0)	1 ( 0.1)	0 ( 0.0)	0 ( 0.0)
age65 (%)	5394 (76.3)	1256 (63.6)	483 (86.7)	1774 (79.9)	476 (95.8)	1127 (81.9)	243 (60.6)	35 (89.7)
gender_male (%)	3737 (52.9)	995 (50.4)	266 (47.8)	1208 (54.4)	233 (46.9)	767 (55.7)	238 (59.4)	30 (76.9)

eTable 6 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. 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 7.** Sensitivity analysis: Association of patient subgroup based on coexisting conditions with clinical outcomes after multivariable adjustment, including adjustment for secondary comorbidities.

Subgroup	Mortality		ICU Admission		30-day readmission		Median Length of Stay	
	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 (1.12-1.55)	0.001	2.11 (1.67-2.66)	<0.001	1.56 (1.11-2.19)	0.011	1.61 (1.29-1.92)	<0.001
Pulmonary	0.84 (0.64-1.11)	0.228	1.40 (1.14-1.70)	0.001	1.18 (0.80-1.75)	0.399	0.39 (0.06-0.71)	0.02
Diabetes	0.65 (0.50-0.85)	0.002	1.09 (0.95-1.26)	0.203	1.01 (0.70-1.46)	0.969	0.19 (-0.09-0.46)	0.183
Heart failure	1.63 (1.33-2.00)	<0.001	1.75 (1.34-2.30)	<0.001	1.26 (0.93-1.70)	0.140	1.14 (0.70-1.57)	<0.001
Dementia	1.55 (1.05-2.29)	0.027	0.85 (0.67-1.08)	0.175	1.27 (0.96-1.69)	0.091	1.23 (0.80-1.66)	<0.001
Cancer	3.09 (2.42-3.95)	<0.001	1.18 (0.75-1.86)	0.467	1.39 (1.14-1.69)	0.001	1.17 (0.74-1.60)	<0.001

eTable 7 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, arrival to hospital from nursing home or by ambulance, laboratory-based acute physiology score, and all comorbidities that were not main drivers of the clusters (renal disease, MI, stroke, liver disease, peripheral vascular disease, rheumatic disease, paralysis, peptic ulcer disease, and HIV). Age and LAPS were modeled using non-linear splines. OR=odds ratio. Coeff=coefficient in quantile regression. CI=confidence interval.



## References

1. Nayyar A, Gadhavi L, Zaman N. Chapter 2 - Machine learning in healthcare: review, opportunities and challenges. In: Singh KK, Elhoseny M, Singh A, Elngar AA, eds. *Machine Learning and the Internet of Medical Things in Healthcare*. Academic Press; 2021:23-45.
2. Loftus TJ, Shickel B, Balch JA, et al. Phenotype clustering in health care: A narrative review for clinicians. *Front Artif Intell*. 2022;5:842306.
3. 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.
4. 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.
5. 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.
6. Thrun MC. Approaches to Cluster Analysis. In: *Projection-Based Clustering through Self-Organization and Swarm Intelligence*. 2018.
7. Coombes CE, Liu X, Abrams ZB, Coombes KR, Brock G. Simulation-derived best practices for clustering clinical data. *Journal of Biomedical Informatics*. 2021;118:103788.
8. Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis / Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl King's College London, UK*. 5th edition ed. Chichester: Wiley; 2010.
9. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572-1573.
10. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*. 1985;50(2):159-179.
11. Hennig C. Package 'fpc' Flexible Procedures for Clustering in R. <https://cran.r-project.org/web/packages/fpc/fpc.pdf>. Published 2022. Accessed.
12. Wilkerson MD. ConsensusClusterPlus (Tutorial). <https://bioconductor.org/packages/release/bioc/vignettes/ConsensusClusterPlus/inst/doc/ConsensusClusterPlus.pdf>. Published 2020. Accessed.