

Explainable hierarchical clustering for patient subtyping and risk prediction

Enrico Werner^{1*}, Jeffrey N Clark^{1*}, Alexander Hepburn¹, Ranjeet S Bhamber¹, Michael Ambler², Christopher P Bourdeaux³, Christopher J McWilliams⁴ and Raul Santos-Rodriguez⁵

¹University of Bristol, Bristol BS1 5DD, UK; ²University of Bristol, Bristol BS8 1TD, UK; ³University Hospitals Bristol NHS Foundation Trust, Bristol BS2 8HW, UK; ⁴University of Bristol, Bristol BS8 1TW, UK; ⁵University of Bristol, Bristol BS8 1QU, UK

*These authors contributed equally to this paper.

Corresponding author: Enrico Werner. Email: enrico.werner@bristol.ac.uk

Impact Statement

With the introduction of electronic health records, hospitals are increasingly using automatic processing of real-time patient data to flag potential issues to clinicians. While the National Early Warning Score 2 (NEWS2) is effective at identifying patients with high clinical acuity, it is less reliable at predicting the clinical outcomes of patients with diverse presentations. We argue that including patient subtypes in such scores would be beneficial to increase their accuracy. We demonstrate a pipeline to hierarchically subtype patients in an explainable way, resulting in improved outcome predictions for many of the identified subtypes. Using routinely collected data, this work could be widely applied following further validation across multiple hospital sites. Evaluation is conducted using automatic techniques as well as evaluation by clinicians, furthering the field by increasing trust in the assigned subtypes and utilizing the mutually beneficial relationship between clinicians and machine learning techniques.

Abstract

We present a pipeline in which machine learning techniques are used to automatically identify and evaluate subtypes of hospital patients admitted between 2017 and 2021 in a large UK teaching hospital. Patient clusters are determined using routinely collected hospital data, such as those used in the UK's National Early Warning Score 2 (NEWS2). An iterative, hierarchical clustering process was used to identify the minimum set of relevant features for cluster separation. With the use of state-of-the-art explainability techniques, the identified subtypes are interpreted and assigned clinical meaning, illustrating their robustness. In parallel, clinicians assessed intracluster similarities and intercluster differences of the identified patient subtypes within the context of their clinical knowledge. For each cluster, outcome prediction models were trained and their forecasting ability was illustrated against the NEWS2 of the unclustered patient cohort. These preliminary results suggest that subtype models can outperform the established NEWS2 method, providing improved prediction of patient deterioration. By considering both the computational outputs and clinician-based explanations in patient subtyping, we aim to highlight the mutual benefit of combining machine learning techniques with clinical expertise.

Keywords: Hierarchical clustering, clinical evaluation, explainability, patient subtypes, mortality prediction, early warning score

Experimental Biology and Medicine 2023; 248: 2547–2559. DOI: 10.1177/15353702231214253

Introduction

Admission to hospital can result in vastly different outcomes depending on the level of illness severity, morbidities and response to treatments.¹ Therefore, selecting the right treatment is challenging even when patients are initially diagnosed with the same conditions. Physicians utilize a huge amount of data to determine diagnosis and treatment options, including but not limited to a patient's medical history and physicians' expertise and intuition.¹ The introduction of electronic health records (EHRs) mean that more information is available to physicians than ever before and it is accessed with ease. The transition phase from paper

records to EHR leads to challenges when analyzing the data and attempting to learn from the heterogeneous population of hospital admissions.² As intensive care units (ICUs) are the most data-rich hospital department and the importance of having quick responses to patient deterioration is large, machine learning approaches have mostly focused on these environments.^{1,3–5} Recent progress has also been made for general wards.^{6–9}

Two areas of high clinical importance which could benefit from developments with EHRs are outcome prediction and risk scoring. Several risk scoring methods have been developed and deployed, for example, Rothman index,¹⁰

MEWS,¹¹ APACHE IV,¹² and SOFA.¹³ These scoring systems typically aim to quantify patient risk in relation to outcomes such as in-hospital mortality, admission to ICU, or cardiac arrest.¹⁴ Supervised machine learning has been applied to directly predict these outcomes^{15,16} in addition to emergency admissions¹⁷ and readiness for discharge.⁵ The NEWS2¹⁸ is a score manually designed by the Royal College of Physicians that assigns weights to clinical observations to detect patient deterioration. NEWS2 is increasingly used in UK hospitals¹⁹ and has good predictive ability in patients with infections and sepsis.²⁰ However, for respiratory diseases like COVID-19, the results are conflicting.^{6,21,22} This indicates that to improve the generalization of scores, patients' similarities must be considered which raises the question: could risk prediction be improved by subtyping patients?

We argue that the predictive ability of scoring systems could be further improved by incorporating patient subtyping that physicians can understand and provide feedback on. Historically, patients were grouped based on their level of sickness, that is, the creation of ICUs. The reorganization presented an innovation, as expertise in caring for the critically ill could be focused on a single location.²³ Instead of focusing on severity of sickness, patients could be further grouped based on their clinical needs.³ While the categorization of patients into risk levels, like the ICU, and diagnosis groups is useful, there could be subtypes of patients shared between these categories that share characteristics, something which physicians would benefit from knowing.

In a pilot study, non-ICU patients were physically grouped based on similar patient characteristics rather than diagnoses, leading to a reduced admittance of low-risk patients to ICU from 42% to 22%. In addition, the average ICU length of stay was reduced from 4.6 to 4.1 days.²⁴

Automatic patient subtyping aims to assign patients to clinically meaningful groups using factors such as their disease progression, medical history, EHR, and ultimately paves the path to precision or personalized medicine by tailoring diagnostic and therapeutic strategies to the patient's needs.^{2,25} Subtyping can be framed as an unsupervised machine learning task, using clustering methods to identify distinct high-density regions separated by sparse regions within a data set.²⁶ These clusters represent patients who are in some sense similar according to the data, where the similarity is not always immediately obvious to the practitioner. Clustering algorithms such as *k*-means and hierarchical clustering have recently been applied to identify clusters in a general ICU population,³ cardiovascular clusters in sepsis patients,²⁷ and corticosteroid response in patients with severe asthma.²⁸

However, clustering alone is insufficient to provide practical support to determine treatment options. The interpretation of the resulting clusters must be validated, by physicians and the cluster assignments predictive abilities.

Before these models can be widely deployed in hospitals, the final users must "trust" the models. This requires an in-depth understanding of the models' behavior and confidence in individual predictions.²⁹ Model-agnostic explainability approaches such as LIME and variants²⁹ can be used for explaining the predictions of clustered data.³⁰ From these

Table 1. Clinical characterization of the full data set.

Number of patients	64,238
Number of hospital admissions	101,670
Gender (% male)	49.15
Age (years, SD)	60.05 (\pm 20.47)
Length of stay (h, median)	14.77 (6.35, 43.22)
Mortality rate (%)	2.91
NEWS2 (SD)	1.39 (\pm 1.73)
Temperature ($^{\circ}$ C, SD)	36.83 (\pm 0.57)
Systolic blood pressure (mmHg, SD)	128.93 (\pm 22.22)
Heart rate (bpm, SD)	79.98 (\pm 16.64)
SATS (%), (SD)	96.66 (\pm 2.32)
Respiratory rate (bpm, SD)	17.23 (\pm 2.73)
Limited level of consciousness (%)	0.526
Number of ICD-10 codes (SD)	8.81 (\pm 5.29)

NEWS2: National Early Warning Score 2; SATS: hemoglobin saturation with oxygen; ICD-10: Tenth Revision of International Classification of Diseases. Value format is mean (SD) or median (25th, 75th percentile).

methods, we can gain an understanding into which patient features contribute to each cluster assignment. Utilizing the assigned subtypes must also be validated, as our hypothesis is that patients within a cluster will present similar hospital stays, that is, length of stay and patient outcome. As such, evaluating the predictive power of the cluster assignments is a must. Building upon the proof-of-concept explainable subtyping process in Werner *et al.*,³¹ this article presents a pipeline from subtyping to outcome prediction in which we:

- Demonstrate the use of unsupervised machine learning techniques to identify patient subtypes on admission for a data set of hospital patients from a large UK teaching hospital.
- Implement a combination of explainability techniques and statistical properties of the clusters to evaluate and assign clinical meaning to the identified subtypes.
- In parallel and independently, hospital clinicians derive the main clinical properties of the identified subtypes using additional records, a key and necessary step in developing human in-the-loop machine learning systems in medical settings.
- Assess the predictive power of the identified patient subtypes for in-hospital mortality and admission to high-dependency hospital units, in comparison to the unclustered existing NEWS2 scoring system.

Materials and methods

In what follows, subjects are patients who were admitted to the Bristol Royal Infirmary, a large teaching hospital covering most medical and surgical specialties. The clinical characteristics of this historical data source are summarized in Table 1. Only patients were considered for which all considered features were available, that is, six vitals (temperature, systolic blood pressure, heart rate, hemoglobin saturation with oxygen [SATS], respiratory rate, level of consciousness), age at hospital admission, gender and number of attributed *Tenth Revision of International Classification of Diseases* (ICD-10) codes at hospital admission. Only patients with vitals taken

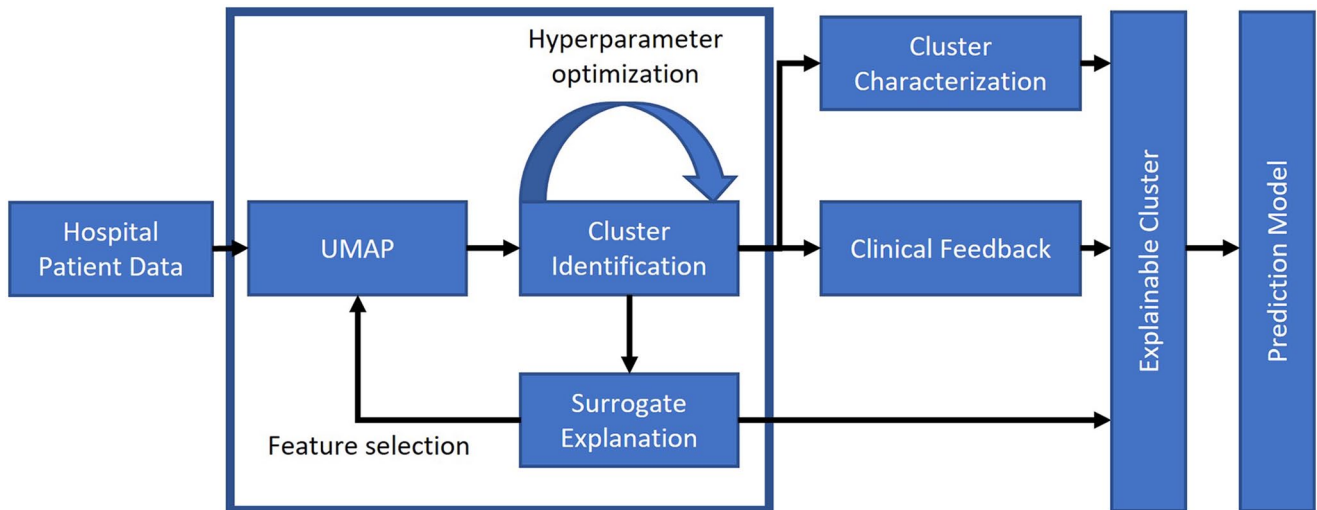


Figure 1. Pipeline overview, from data set import to generation of explainable clusters and clinical outcome predictions. The blue box denotes the iterative clustering process.

within the first 24h after hospital admission were considered, and only their first set of vitals was utilized in this study. Patient visits lasting less than 2h were considered as routine appointments and omitted. Some patients were admitted several times and each admission is considered as an independent event. Patients with restricted or limited level of consciousness are described as “unconscious.”

Clustering process

Cluster identification and feature selection followed an iterative and hierarchical process. First, the entire population was analyzed, characterized, features selected and explained, and then the same processing steps were repeated for the identified clusters (where number of patients ≥ 1000) individually.

Therefore, each stage followed the same subsequent steps. To simplify the analysis and aid interpretability, dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP)³² based on the six vitals (temperature, systolic blood pressure, heart rate, SATS, respiratory rate, and level of consciousness), age at hospital admission, gender, and number of ICD-10 codes at hospital admission. The first three vitals were scaled, the latter features were transformed with the logit function. After dimensionality reduction, HDBSCAN³³ was applied to the embedding to identify clusters. The hyperparameters `min_samples` (range = 10–100 in steps of 1) and `min_cluster_size` (range = 20 in 100 steps of 10) for HDBSCAN were selected based on the fast approximation of the density-based cluster validity (DBSV) score³⁴ which in return also indicated the optimal number of clusters. Next, surrogate explanations were used to identify the features that did not contribute to the cluster separation and could therefore be excluded in the next iteration. This process was repeated until only contributing features remained. This pipeline is visualized in Figure 1.

Subclustering results are only presented for the clusters containing the lowest and highest numbers of subclusters.

Clustering explanations

Understanding how and why patients were clustered is a fundamental requirement to establish clinical trust and ultimately reduce the risk of unintended harm. Each feature’s contribution for the cluster assignment of each patient was determined using a surrogate explainer to estimate the local decision boundary between the assigned cluster and the rest of the data. A total of 25,000 samples were generated around a query point and using the TabularBlimeyTree decision tree explainer³⁵ within FAT Forensics (v0.1.1): an open source toolbox,³⁶ a surrogate model was fit in order to predict the cluster assignment of the samples. Once fitted, feature importance can be determined from the weights within the surrogate model. Input features were the scaled features. The probabilistic argument was set to false. Default arguments and settings were otherwise used. All generated samples were visualized in the embedding space to ensure that they approximately followed the distribution of the underlying data.

Outcome prediction

Clustering identifies patient subtypes that are more homogeneous and could result in improved outcome predictions. In this preliminary analysis, the same nine features employed for initial clustering were applied for outcome prediction and with the same feature scaling regime. Gender was one-hot encoded. Predictions were made using logistic regression classification models for two targets: in-hospital mortality and admission to higher care units within the hospital stay, comprising general ICU, cardiac ICU, and the critical care unit. Data were split 80:20 for training and testing with stratified sampling for the two targets. `Class_weight` was “balanced,” all other hyperparameters were as default, and hence, no validation set was required. All outcome prediction analysis was in Python 3.9.11, with models and performance metrics implemented using scikit-learn 1.0.2.

For comparison purposes, the predictive power of NEWS2 was assessed. For cases during the study window for which

NEWS2 was recorded ($n=77,731$), values were normalized by dividing by the maximum NEWS2 score. The optimum threshold for binarizing the prediction for each target was computed by maximizing the F1 score in the training set for all patient stays and this threshold was used to compute performance in the test set.

Clinical evaluation

Clinical validation was conducted by providing two intensive care clinicians with the cluster characterization and occurrence of ICD-10 codes for each cluster (Figures 3 to 5). The clinicians assessed and evaluated intracluster similarities and intercluster differences according to both the data and their clinical knowledge (Figure 1). Blinded to the surrogate clustering explanations, clinicians independently assigned a clinically meaningful name and description to each cluster, highlighting which input features resulted in each cluster’s unique characteristics. Thereafter, the features described by the clinicians were compared against those computed automatically.

Results

Cluster characterization

The data extracted between November 2017 and March 2021 comprised 116,004 cases (70,452 patients). Of these, 101,670 cases (64,238 patients) had all vitals taken within the first 24h of their ≥ 2 h hospital stay and were included in the study.

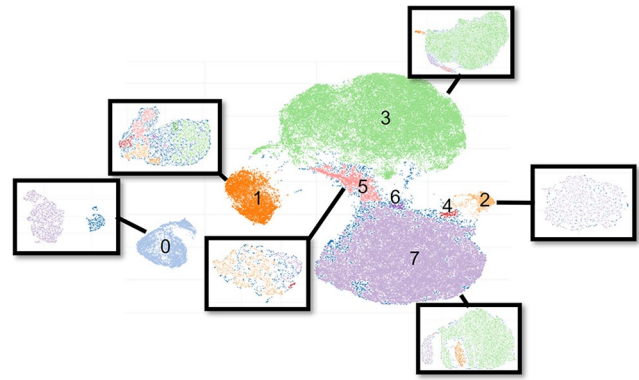


Figure 2. Patients mapped onto the two-dimensional embedding space after dimensionality reduction and clustering. Clusters inside black boxes depict the subclustering results. Subclustering was not performed for clusters 4 and 6 as both contained less than 1000 patients. Unclusterable patients are shown in dark blue, often at the edges of clusters.

Dimensionality reduction and clustering of the entire population revealed eight clusters and a group of unclusterable patients (Figure 2, summarized in Tables 2 and 3). Clusters 0 and 1 are separated from the other clusters, while clusters 2–7 are closer together, joined via patients our clustering algorithm deemed as unclusterable patients. Subclustering was performed on all clusters except for clusters 4 and 6 as both contained less than 1000 patients (Table 3).

Using surrogate explanations, features contributing to cluster separation were identified and irrelevant features

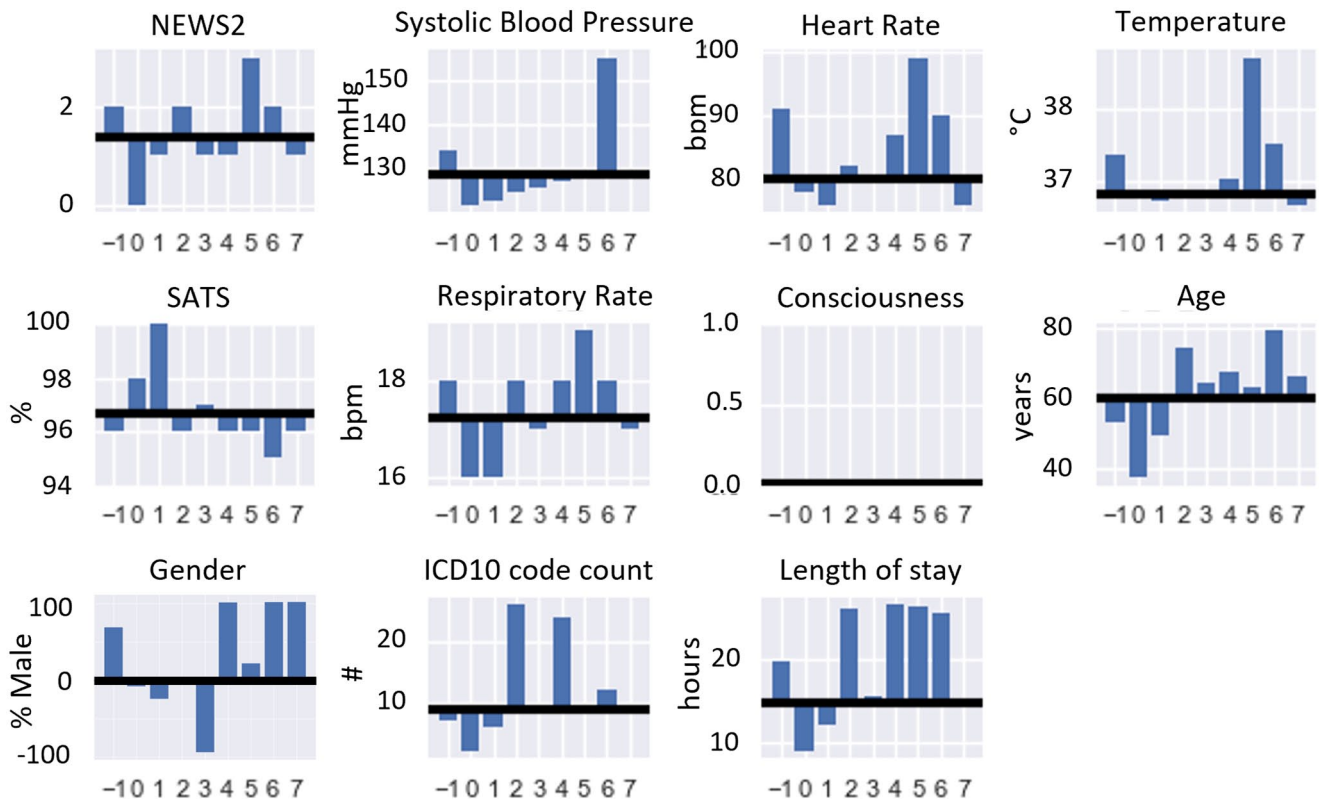


Figure 3. NEWS, vitals, age, gender, ICD-10 code count, and length of stay for individual clusters from clustering the entire population. The mean value of each cluster is compared to the mean or median value depending on the feature (black line) of the whole population.

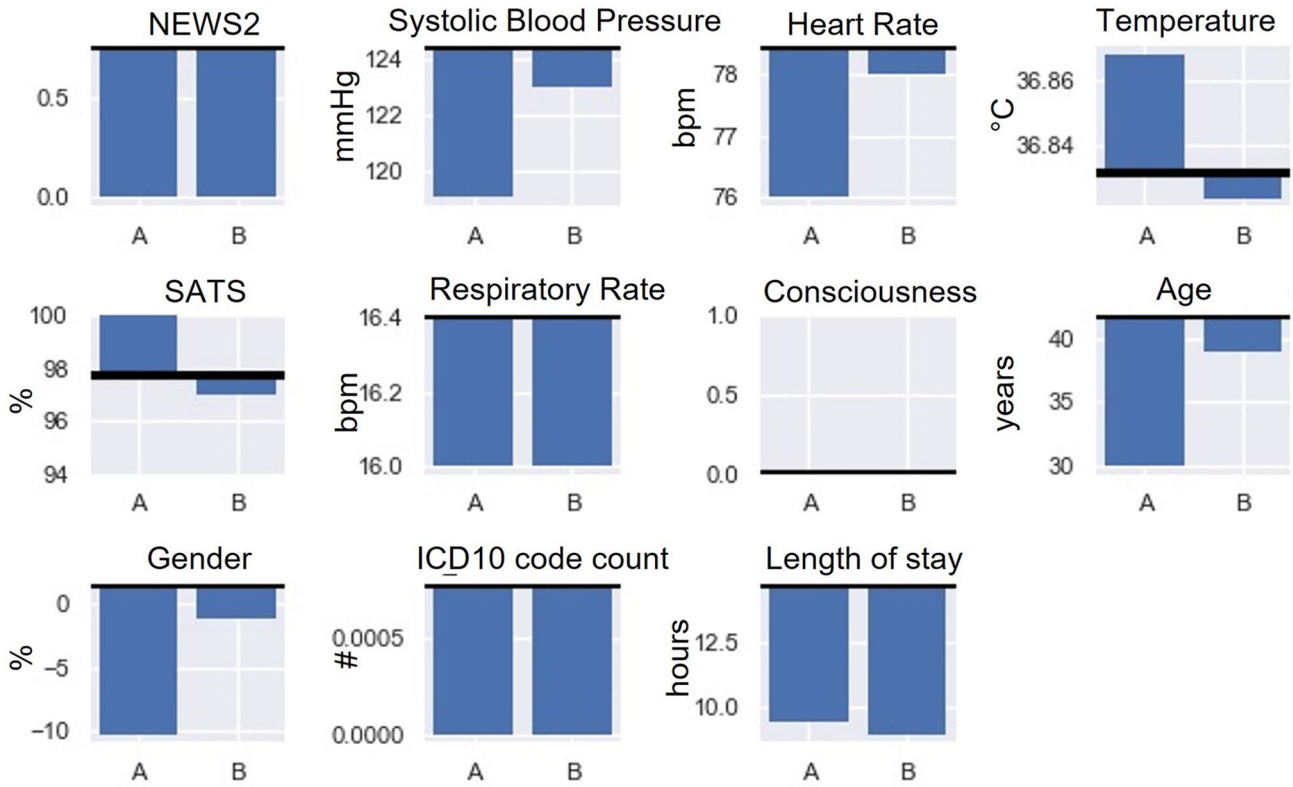


Figure 4. NEWS, vitals, age, gender, ICD-10 code count, and length of stay show the subclustering results for cluster 0. The mean value of each cluster is compared to the mean/median value (black line) of the parent cluster.

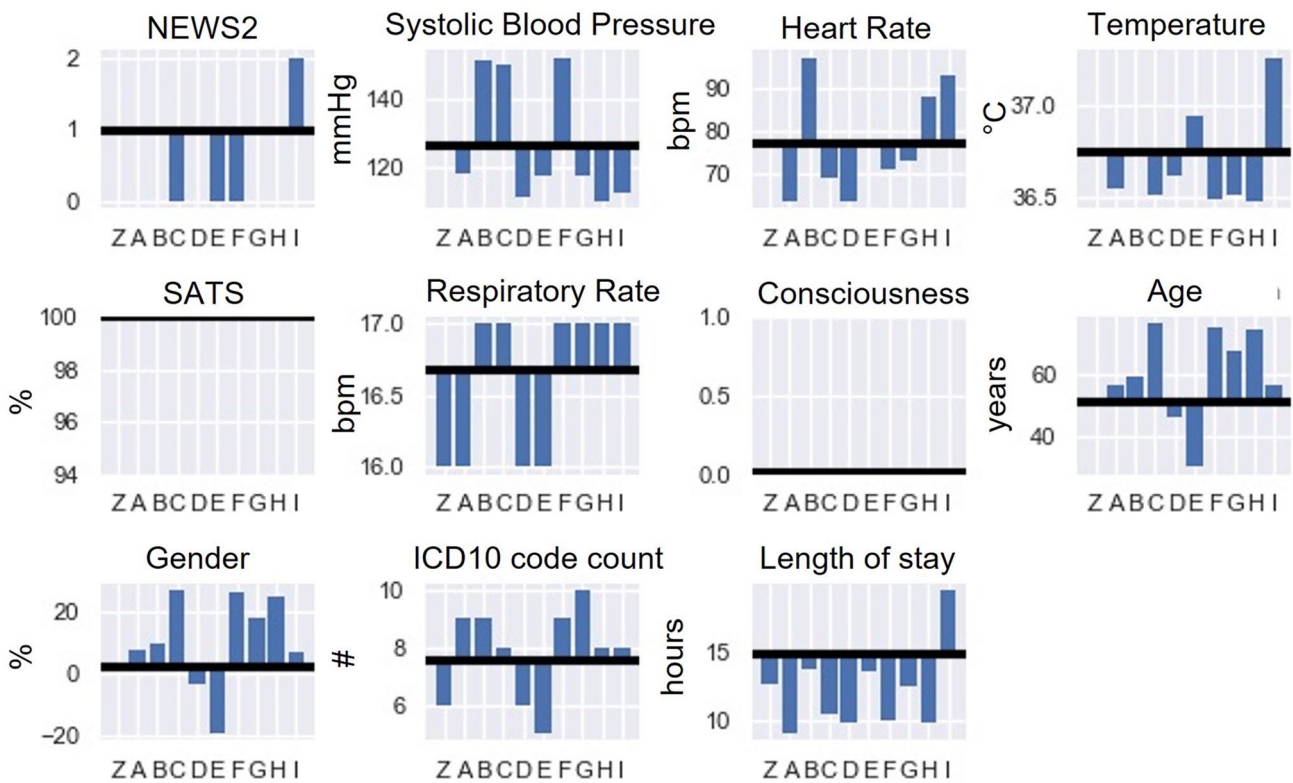


Figure 5. NEWS, vitals, age, gender, ICD-10 code count, and length of stay show the subclustering results for cluster 1. The mean value of each cluster is compared to the mean/median value (black line) of the parent cluster.

Table 2. NEWS and features of identified clusters of the full data set.

Cluster	NEWS (median)	Temperature (°C, mean)	Blood pressure (mmHg, mean)	Heart rate (bpm, mean)	SATS (% , median)	Respiratory rate (bpm, mean)	Consciousness (median)
-1	2.00 (1.00, 3.00)	37.36 (±0.74)	137.56 (±24.89)	90.73 (±20.17)	96.00 (95.00, 98.00)	18.08 (±3.53)	0.00 (0.00, 0.00)
0	0.00 (0.00, 1.00)	36.83 (±0.48)	124.46 (±17.32)	78.43 (±14.56)	98.00 (97.00, 99.00)	16.41 (±2.04)	0.00 (0.00, 0.00)
1	1.00 (0.00, 2.00)	36.75 (±0.50)	125.96 (±21.11)	77.14 (±15.37)	100 (100.00, 100.00)	16.68 (±2.38)	0.00 (0.00, 0.00)
2	2.00 (1.00, 3.00)	36.84 (±2.33)	127.95 (±24.33)	83.12 (±17.01)	96.0 (94.00, 97.00)	18.34 (±3.33)	0.00 (0.00, 0.00)
3	1.0 (0.00, 2.00)	36.85 (±0.49)	128.99 (±23.83)	82.02 (±17.01)	97.0 (96.00, 98.00)	17.23 (±2.76)	0.0 (0.00, 0.00)
4	1.0 (0.00, 2.00)	37.04 (±0.39)	127.70 (±14.01)	87.07 (±8.78)	96.0 (95.75, 98.00)	18.41 (±3.67)	0.0 (0.00, 0.00)
5	1.0 (0.00, 2.00)	38.69 (±0.51)	129.92 (±21.65)	99.20 (±16.32)	96.0 (95.75, 98.00)	19.59 (±4.20)	0.0 (0.00, 0.00)
6	3.0 (2.00, 5.00)	37.53 (±0.48)	155.65 (±15.27)	91.35 (±16.83)	96.0 (95.00, 97.00)	19.09 (±3.65)	0.0 (0.00, 0.00)
7	1.0 (0.00, 2.00)	36.68 (±0.48)	129.64 (±20.85)	76.77 (±15.26)	96.00 (96.00, 98.00)	17.24 (±2.57)	0.0 (0.00, 0.00)

NEWS: National Early Warning Score; SATS: hemoglobin saturation with oxygen.

Value format is mean (SD). ICD-10 (Tenth Revision of International Classification of Diseases) specifies codes for diseases and diagnoses, where for each cluster, the most frequent code corresponds to the following; clusters -1 (unclusterable patients), 4, 5=sepsis; cluster 0=acute tonsillitis, unspecified; cluster 1=lower abdominal pain; cluster 3=chronic obstructive pulmonary disease with (acute) lower respiratory infection; cluster 6=chronic ischemic heart disease.

Bold values are the highest value per feature.

Table 3. Cluster characterization of the full data set.

Cluster	Count	Mortality	Number of ICD-10 codes	ICD-10 (most frequent)	Length of stay	Gender (% male)	Features (subclustering)
-1	1917	4.3297	7.00 (4.00, 13.00)	A419	56 h 49 min	83.15	–
0	6457	0.0155	2.00 (2.00, 2.00)	J039	20 h 10 min	45.33	BP, HR, SATS, age
1	7569	1.7043	6.00 (4.00, 10.00)	R103	39 h 7 min	37.30	BP, temp, age, HR, gender, ICD-10 codes
2	1839	0.7069	26.00 (25.00, 26.00)	I500	126 h 48 min	48.02	BP, temp, SATS, ICD-10 codes, AVPU, gender, HR, resp, age
3	42,423	0.0942	8.00 (5.00, 12.00)	J440	47 h 04 min	3.06	BP, HR, temp, age, ICD-10 codes
4	116	1.7241	24.00 (23.00, 24.00)	A419	124 h 44 min	99.14	–
5	2148	4.3762	24.00 (23.00, 24.00)	A419	65 h 33 min	60.01	BP, HR, temp, age, gender, ICD-10 codes
6	147	5.4422	9.00 (6.00, 13.00)	J189	68 h 26 min	99.32	–
7	39,054	2.6758	8.00 (6.00, 12.00)	I251	46 h 31 min	99.40	HR, temp, age, ICD-10 codes

BP: blood pressure; HR: heart rate; SATS: hemoglobin saturation with oxygen.

Value format is mean (SD). ICD-10 (Tenth Revision of International Classification of Diseases) specifies codes for diseases and diagnoses, where for each cluster, the most frequent code corresponds to the following: cluster -1 (unclusterable patients), 4, 5=sepsis; cluster 0=acute tonsillitis, unspecified; cluster 1=lower abdominal pain; cluster 3=chronic obstructive pulmonary disease with (acute) lower respiratory infection; cluster 6=chronic ischemic heart disease.

Bold values are the highest value per feature.

excluded. Within three iterations of this process, the final set of features was established (Figure 9). When repeating this process for subclusters, subclusters 1C and 1E required only one and two iterations, respectively.

Subclustering disclosed multiple clusters, ranging from two new clusters (Figure 2, cluster 0) to eight new clusters (Figure 2, cluster 1). Subclustering of cluster 2, however, did not reveal any new clusters and only distinguished clustered patients from unclusterable patients. The degree of unclusterable patients varied from 0% in cluster 0 to 38.6% in cluster 1. The variations in the degree of unclusterable patients are also reflected in the DBSV score. While clustering the entire population revealed a DBSV = 0.5442, subclustering affected the DBSV score positively (0 = 0.8742, 3 = 0.7895, 5 = 0.6976) and negatively (1 = 0.4310, 2 = 0.4165, 4 = 0.4248). Below, clusters 0 and 1 were chosen as examples as they represent the extremes of the lowest and highest number of unclustered patients and subclusters.

Cluster visualization and vitals

A goal of analyzing the individual clusters is to identify their unique characteristics relative to each other and the overall population. Figure 3 shows the difference in measurements

for each cluster relative to the overall population. Figures 4 and 5 show the same but for the identified subclusters of clusters 0 and 1, respectively.

The vast majority of patients in the population (Figure 3) are male in clusters 2 and 6, whereas cluster 0 is predominantly female. The average length of stay is the longest for the unclusterable patients and in clusters 6, 2, 5, and 4 (in ascending order). Although patients in cluster 5 are close to the population average age, this cluster contains patients with the highest temperature, heart rate, respiratory rate which is also reflected in the highest NEWS2. In contrast, cluster 0 has the youngest population, with the lowest respiratory rate, the lowest blood pressure, and the lowest NEWS2. Consequently, the median length of stay is also the shortest. Patients are the oldest in cluster 6, with by far the highest blood pressure and the lowest SATS. Unclusterable patients (“-1”) are mostly male and have elevated temperature, systolic blood pressure, heart rate, and respiratory rate, hence the increased NEWS2.

Cluster 0 (Figure 4) consists of two subclusters in which all patients could be assigned to a cluster; therefore, no unclusterable patients were identified. While the median NEWS2, mean respiratory rate, and count of ICD-10 codes are equal for both clusters, cluster 0A contains much younger

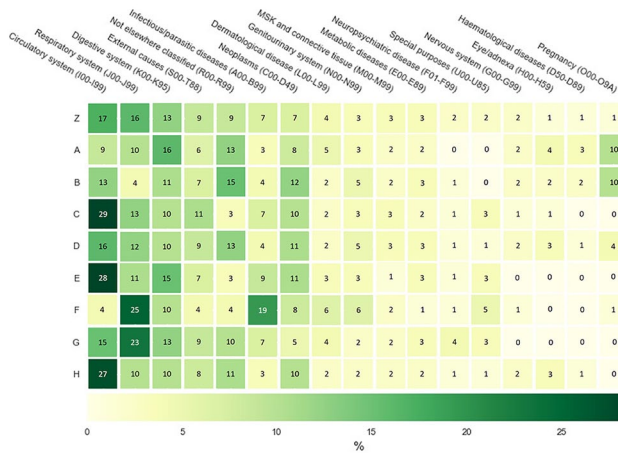


Figure 6. Heatmap of primary ICD-10 codes of full-population clustering and collated by top-level grouping. For display purposes, only ICD-10 codes with ≥2% incidence for at least one cluster are displayed. Since only a subset of ICD-10 codes are visualized, each row does not add up to 100. MSK: musculoskeletal.

patients than 0B and approximately 34% more likely to be female. In addition, 0A has a 4.5% longer median hospital stay, and approximately 90% increase in a limited level of consciousness and a lower systolic blood pressure.

Nine subclusters and a group of unclusterable patients were identified in cluster 1 (Figure 5). All identified subclusters have the same median SATS of 100%, and subclusters 1B, 1C, 1F, 1G, 1H, and 1I have an increase of approximately 6% in respiratory rate. Cluster 1C has the oldest population and is mostly female. Compared to the rest of the subpopulation, the temperature is marginally lower, and the number of ICD-10 codes increased by ~32%. Patients have the longest hospital stay in cluster 1I and the highest median NEWS and temperature. Yet, they are also most likely to be conscious.

ICD-10 codes

Figure 6 shows the frequency of identified primary ICD-10 code groupings per cluster. The highest incidence ICD-10 code group is “Circulatory system (I00–I99)” with 29%, 28%, and 27.0% in clusters 2, 4 and 7, respectively.

Diseases of the “respiratory system (J00–J99)” are common in clusters 5 (25%) and 6 (23%), with approximately twice times higher occurrence than in the other clusters. “Infectious/parasitic diseases (A00–B99)” are mostly in cluster 5 (19%) with the second closes being in cluster 4 with 9%. “Pregnancy-related (O00–O9A)” patients are almost exclusively found in clusters 0 and 1. In contrast to the other ICD-10 code groups, “digestive system (K00–K95)” is relatively equally split between all clusters, ranging between 10% and 16%.

The ICD-10 codes for the identified subclusters of clusters 0 and 1 are distributed differently. For cluster 0’s subclusters (Figure 7), the most prominent ICD-10 code is “Pregnancy-related (O00–O9A)” with 22% in cluster 0A. This is followed by “Digestive system (K00–K95)” with 15% and “Not elsewhere classified (R00–R99)” with 13%. These two are also the most relevant ICD-10 codes in cluster 0B with 16% and 12%, respectively.

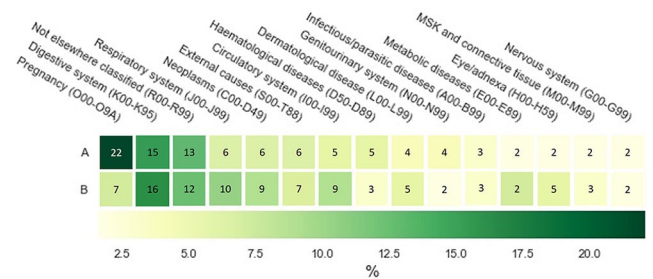


Figure 7. Heatmap of primary ICD-10 codes of the subclusters of cluster 0 as recorded by clinicians at the time of patient admission and collated by top-level grouping. For display purposes, only ICD-10 codes with ≥2% incidence for at least one cluster are displayed. Since only a subset of ICD-10 codes are visualized, each row does not add up to 100. MSK: musculoskeletal.

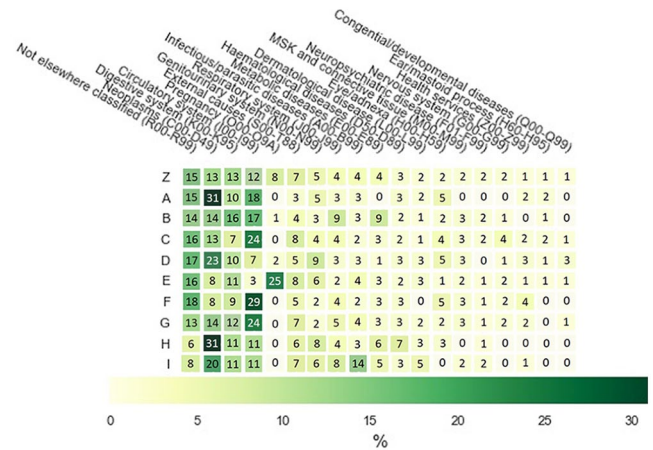


Figure 8. Heatmap of primary ICD-10 codes of different subclusters as recorded by clinicians at the time of patient admission and collated by top-level grouping. For display purposes, only ICD-10 codes with ≥2% incidence for at least one cluster are displayed. Since only a subset of ICD-10 codes are visualized, each row does not add up to 100. MSK: musculoskeletal.

For cluster 1 (Figure 8), “Not elsewhere classified (R00–R99)” is common occurring in 15% of unclustered patients (1Z), followed by “Neoplasms (C00–D49),” “Digestive system (K00–K95),” and “Circulatory system (I00–I99)” with 13%, 13%, and 12%, respectively. “Neoplasms (C00–D49)” is overall the most common ICD-10 code and has a value of 31% in clusters 1A and 1H. “Pregnancy (O00–O9A)”-related cases appear almost exclusively in 1E with 25%. “Circulatory system (I00–I99)” patients are mostly common in clusters 1F, 1G, and 1C with 29%, 21% and 21%, respectively. Whereas “Infectious/parasitic diseases (A00–B99)” range mostly between 2% and 4%, cluster 1I is an exception with 14%.

The distribution of ICD-10 codes allows us to see that the clusters are not just comprised of patients with one type of disease and in fact that clusters usually contain a mixture of different diagnoses. This reinforces the need for additional patient clustering as it is not enough to consider only diagnoses codes when assigning patient subtypes.

Surrogate explanations

Using surrogate explanations, features not relevant for cluster separation could be excluded. Consequently, different

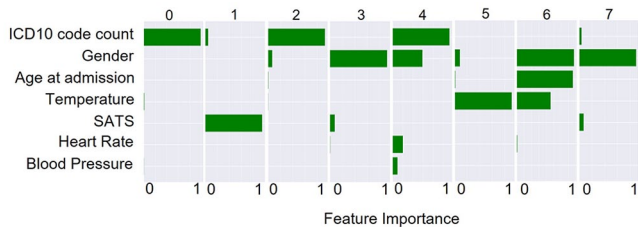


Figure 9. Surrogate explanations for the contribution of each vital in determining the assignment of patients into each cluster from clustering of the entire population. SATS: hemoglobin saturation with oxygen.

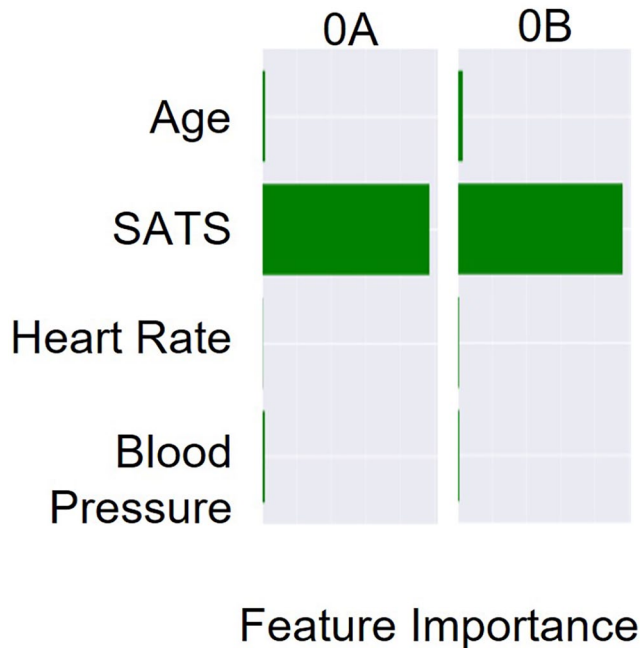


Figure 10. Surrogate explanations for the contribution of each vital in determining the assignment of patients into each subcluster of cluster 0. SATS: hemoglobin saturation with oxygen.

features remained for clustering the entire population and subclustering clusters 0 and 1.

For the entire population, “ICD-10 code count” is the most important feature for assigning patients to clusters 0, 2, and 4. Whereas “Gender” is most important for clusters 3, 6, and 7. “SATS” and “Temperature” are the most important factors for clusters 1 and 5, respectively. “Heart Rate” and “Blood Pressure” have only a minor role, mostly in cluster 4 (Figure 9).

Clustering 0 identified only two subclusters with “SATS” being by far the most dominant feature with subcluster 0A containing patients with a median SATS level of 100% and 97% saturation for subcluster 0B at their first measurement point (Figure 10). The other features, that is, “Age,” “Heart Rate,” and “Blood Pressure,” have only a minor impact (Figure 10).

In total, nine subclusters were determined for cluster 1. “Age” is the most important factor in five clusters, that is, 1A, 1C, 1D, 1E, and 1I, and has still a high impact on clusters 1F, 1G, and 1H. Other features with high impact are “Gender,” “Heart Rate,” and “Blood Pressure.” Whereas

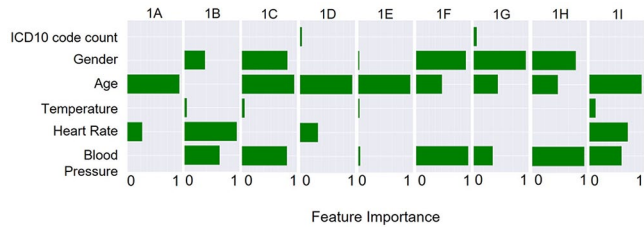


Figure 11. Surrogate explanations for the contribution of each vital in determining the assignment of patients into each subcluster of cluster 1. SATS: hemoglobin saturation with oxygen.

“ICD-10 code count” and “Temperature” have only a minor influence on cluster separation (Figure 11).

Clinical evaluation

The clinicians were able to detect intercluster differences and intracluster similarities, enabling the generation of labels and descriptions for each cluster (Table 4, full descriptions in the Supplementary section). For example, cluster 0 was defined by having young healthy patients with few comorbidities, an even spread of primary diagnoses, brief stay, median NEWS2 of 0, and low mortality. All cluster descriptions are available in the Supplementary Materials. The two clinicians generally gave similar descriptions for each cluster, although there were some differences in which features they used to define each cluster. Of the 21 features mentioned by clinician A and 30 features by clinician B to describe membership to clusters 0–7, 15 were in common. Including in common omission of irrelevant features (36), in total out of 72 opportunities to comment on features, there was an overlap of 51 features (71%).

Features found to contribute to cluster assignment through the automated clustering process (Figure 6) were largely identified by the two clinicians when independently assessing the characteristics of each cluster (Table 4). Of the 26 features which contributed to level 1 clustering as identified automatically, 17 were identified by at least one clinician, and 12 were identified by both clinicians. Including in common omission of irrelevant features for the artificial intelligence (AI) system and both clinicians (27), in total there was an agreement in features of 54% between all three agents, and 74% between the AI system and at least one of the two clinicians.

Outcome prediction

Relative performance in predicting mortality and admission to higher care varied by cluster, and performance metric, when compared to the entire cohort (“All”) and NEWS2 (Table 5). For mortality, clusters –1 (unclusterable patients), 1, 2, 3, and 7 performed particularly well, and this was reflected in the receiver operating characteristic (ROC) and precision recall curves (Figure 12(a) and (b)). For admission to higher care, clusters –1 (unclusterable patients), 2, and 6 performed particularly well compared to the NEWS2 scoring system. Logistic regression predictive models were generated for subclusters within clusters 0 and 1 (Supplementary Figures 1 and 2 and Supplementary Tables 1 and 2). Several

Table 4. Comparison of feature importance in characterizing each cluster between clinicians and surrogate explanations.

Cluster	0	1	2	3	4	5	6	7
Label	Young, low acuity with oxygen saturations < 100%	Young, low acuity with oxygen saturations 100%	Old, multimorbid, presenting with cardiovascular disease, high mortality	Short stay, average women	Men with cardiovascular disease, high risk	Inflammatory/infective	Older men with respiratory disease and hypertension	Men with cardiovascular disease, low risk
Observer	A	B	A	A	A	B	A	A
ICD-10 code count	x	x	x	x	x	x	x	x
Gender	x	x	x	x	x	x	x	x
Age at admission	x	x	x	x	x	x	x	x
Temperature	x	x	x	x	x	x	x	x
SATS	x	x	x	x	x	x	x	x
Heart rate	x	x	x	x	x	x	x	x
Blood pressure	x	x	x	x	x	x	x	x
Respiratory rate	x	x	x	x	x	x	x	x
Consciousness	x	x	x	x	x	x	x	x

SATS: hemoglobin saturation with oxygen; ICD-10: Tenth Revision of International Classification of Diseases; AI: artificial intelligence. Comparison of feature importance in characterizing each cluster between clinicians A and B and compared against the surrogate explanations for the AI clustering algorithm as shown in Figure 8. Displayed labels were provided by clinician A. Cluster -1 (unclusterable patients) is not included. Full labels and cluster descriptions provided by each clinician are available in the Supplementary section.

Table 5. Cluster outcome prediction performance.

Cluster	Positive cases (%)	Brier loss	Accuracy	AUROC	AUPRC	Balanced accuracy	F1	Sensitivity	Specificity
(a) Mortality									
-1	4.38	0.136	0.807	0.835	0.256	0.731	0.229	0.647	0.815
1	1.78	0.132	0.814	0.932	0.155	0.887	0.156	0.963	0.812
2	22.68	0.222	0.641	0.701	0.414	0.649	0.455	0.663	0.635
3	2.63	0.145	0.787	0.885	0.213	0.801	0.167	0.816	0.786
4	16.38	0.343	0.583	0.325	0.120	0.550	0.286	0.500	0.600
5	4.38	0.190	0.714	0.687	0.088	0.624	0.140	0.526	0.723
6	5.44	0.188	0.667	0.482	0.062	0.357	0.000	0.000	0.714
7	2.77	0.147	0.792	0.865	0.214	0.790	0.174	0.788	0.792
All	2.91	0.138	0.804	0.886	0.195	0.813	0.196	0.822	0.803
NEWS2	—	0.033	0.925	0.739	0.127	0.626	0.201	0.307	0.945
(b) Admission to higher care									
-1	11.22	0.180	0.781	0.879	0.446	0.806	0.462	0.837	0.774
0	1.08	0.217	0.642	0.785	0.041	0.713	0.045	0.786	0.641
1	4.44	0.200	0.691	0.754	0.174	0.682	0.161	0.672	0.692
2	37.25	0.210	0.668	0.737	0.563	0.672	0.606	0.686	0.658
3	7.56	0.221	0.645	0.687	0.158	0.625	0.204	0.603	0.648
4	40.52	0.279	0.458	0.421	0.434	0.436	0.316	0.300	0.571
5	7.45	0.197	0.693	0.755	0.224	0.662	0.233	0.625	0.698
6	11.56	0.263	0.533	0.667	0.695	0.593	0.222	0.667	0.519
7	14.49	0.224	0.638	0.692	0.284	0.635	0.335	0.630	0.639
All	10.23	0.209	0.670	0.739	0.246	0.669	0.293	0.668	0.670
NEWS2	—	0.106	0.426	0.536	0.131	0.516	0.199	0.633	0.400

NEWS: National Early Warning Score; AUROC: area under receiver operating characteristic curve; AUPRC: area under precision recall curve. Performance metrics in predicting (a) mortality and (b) admission to higher care (general ICU, cardiac ICU, and critical care unit) by implementing a logistic regression model for each cluster and the unclustered "All" patient cohort. Models are compared against predictive performance of the existing NEWS2 risk scoring system by thresholding at NEWS2 ≥ 5 for mortality and NEWS2 ≥ 2 for admission to higher care. Mortality was not predicted for cluster 0 since only one positive case occurred. Models with better performance than NEWS2 are highlighted in bold.

of these subclusters showed improved performance compared to their parent cluster's performance: for example, subclusters 1D and 1F for admission to higher care units.

Logistic regression mortality prediction models were not generated for cluster 0, nor any of its subclusters, since there was only one positive example in this cluster. Likewise, within cluster 1, no mortality predictions were made for subclusters 1A, 1B, 1D, and 1H nor subclusters 1H and 1I for admission to higher care due to insufficient positive examples. Thresholds for NEWS2's predictive power were optimized by maximizing the F1 score and were found to be NEWS2 ≥ 5 for mortality and NEWS2 ≥ 2 for admission to higher care.

Discussion

This study presents a pipeline in which explainable hierarchical clustering is used for patient subtyping and risk prediction. Patient subtyping by way of clustering could be the first step toward a personalized scoring system, improving the predictive success of currently deployed risk scoring metrics.^{10–13,19}

The clusters identified in this study were based on six vitals from the first set of readings taken during a hospital stay in combination with age, gender, and number of ICD-10 codes at admission. Using these few features and the focus on hospital departments outside of intensive care are in contrast to previous studies. Castela Forte *et al.*¹ and Vranas *et al.*³ included 76 and 23 clinical features, respectively, resulting

in the identification of six subtypes of ICU patients. In this study, the first clustering iteration of the entire population revealed eight clusters and a group of unclusterable patients. The clinicians identified "unclusterable patients" as average patients without any distinguishable features. In the second clustering iteration, that is, subclustering, identified a total of 23 subclusters. Cluster 4 and 6 were excluded from subclustering as they contained less than 1000 patients. The subclusters enabled the clinicians to assign descriptions with more precise clinical meaning. This could be the result of patient subtypes becoming more homogeneous which would also be supported by the increase in the DBSV score.

The feature contributions identified by the surrogate explanations varied between clusters and were found to be largely in agreement with the clinicians. This encourages tailored feature selection for individual patient subtypes and will, in the future, be expanded upon so that some additional features will be available for some patient subtypes, for example, additional blood tests for some patient subtypes.

Here, cluster 6 has the highest mortality rate with 5.44% which is also reflected in the highest median NEWS2 of 3.0. Surprisingly, this does not correlate with the maximum average hospital length of stay. While cluster 6 patients stay on average in hospital for 68 h 26 min, patients in cluster 2 and 4 stay longer with an average length of stay of 126 h 48 min and 124 h 44 min, respectively. Cluster 2 also has the highest ICD-10 code count with 26, followed by cluster 4 and 5 with 24.

To provide further insights into the clinical meaning of the patient subtypes, the most frequent ICD-10 codes were

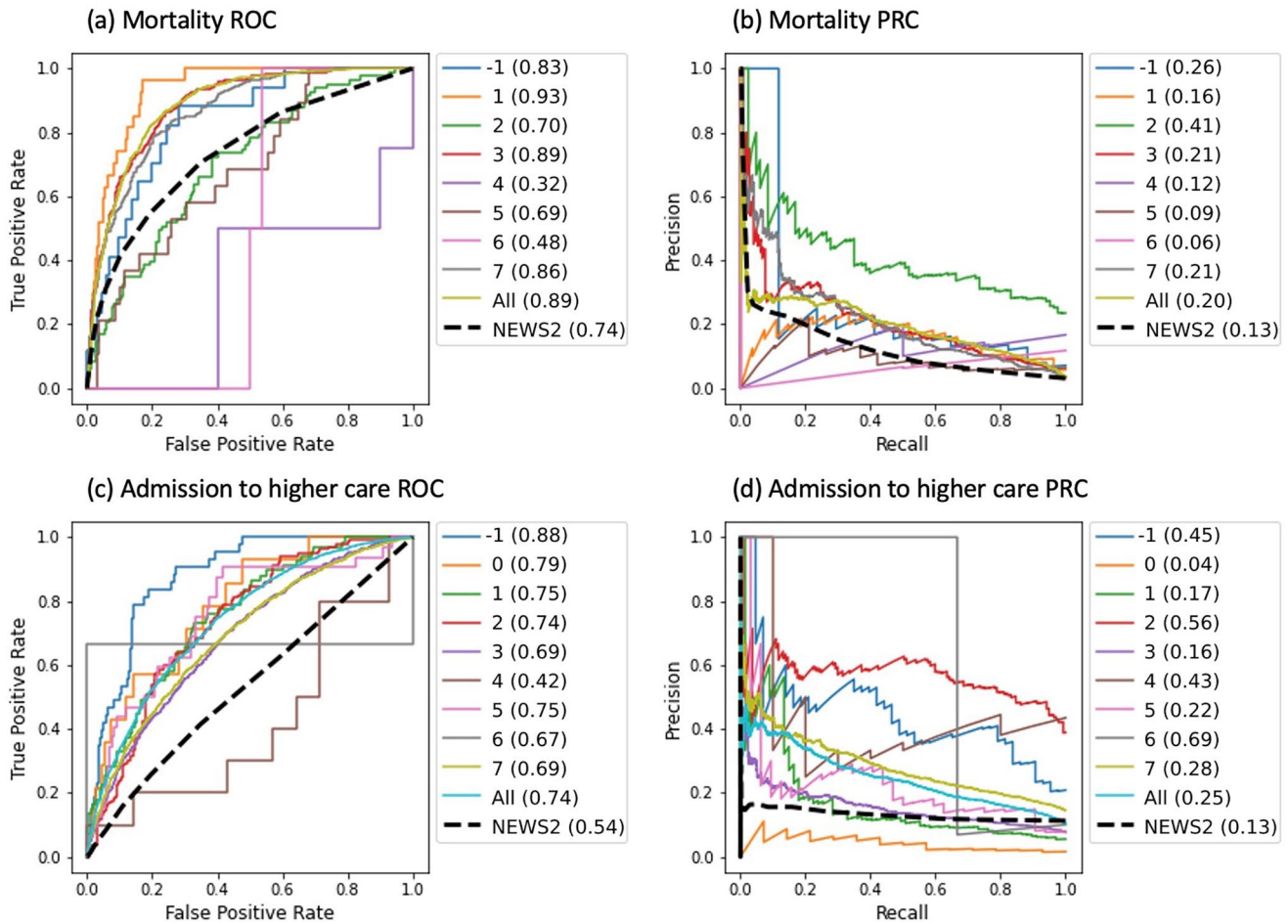


Figure 12. Predictive performance for classification models compared against the existing NEWS2 risk scoring system for the two predicted outcomes: in-hospital mortality and admission to higher care (general ICU, cardiac ICU, and critical care unit). "All" refers to the entire unclustered patient cohort. Mortality was not predicted for cluster 0 since only one positive case occurred: (a) mortality ROC, (b) mortality PRC, (c) admission to higher care ROC, and (d) admission to higher care PRC. Figures in brackets are the area under the curve. ROC: receiver operating characteristic curves; PRC: precision recall curve.

identified. Vranas *et al.*³ found "Sepsis" as the most common diagnosis in ICU patients in five out of six clusters, whereas Castela Forte *et al.*¹ determined a different leading cause for each cluster. In this study, "Sepsis" is found as the most common diagnosis in two out of nine clusters plus the unclusterable patient group. However, all three clusters account for only about 2.23% of the hospital population. The two largest clusters found "Respiratory Infections" and "Heart disease" as the most common ICD-10 and account for 41.73% and 38.41% of the hospital population, respectively. Castela Forte *et al.*¹ also identified two clusters with high prevalence of respiratory failure.

The identified NEWS2 thresholds for best predicting mortality and admission to higher care are lower than the national guidelines' triggers for emergency assessment.¹⁸ This is unsurprising since our study focuses on the initial set of observations at the time of admittance and the patient may only deteriorate days or weeks later. Hence, other studies often limit performance evaluation to outcomes predicted within 24h of a recorded score.³⁷ However, clusters and subclusters identified in this study were better able to predict early signs of these patient outcomes compared to

the existing NEWS2 system even though the average stay per cluster was often two or more days (Table 2). For example, cluster 2 saw consistent improvements in predictive performance, specifically F1 score, for both targets when compared to NEWS2. A defining characteristic of cluster 2 independently noted by both clinicians was cardiovascular disease. This provides further evidence that certain subtypes could benefit from improved predictive performance via a clustering process if deployed clinically. Of note, cluster -1 (unclusterable patients) models performed well compared to NEWS2 and often better than other clusters. This suggests that even those identified patients not in a clearly defined cluster could benefit from the presented approach. Generally, admission to higher care predictive performance for each cluster was better than mortality when compared to the NEWS2 baseline. For mortality, individual clusters and subclusters generally showed improved sensitivity compared to NEWS2, whereas for admission to higher care, they had both improved sensitivity and specificity. Admission to higher care was less imbalanced than mortality with 10.2% of cases in the positive class compared to 2.9% for mortality (Table 1), which may have contributed to the improved

predictive performance. However, class imbalance, and small sample size, remained an issue and likely influenced performance for some clusters, in particular for metrics such as area under precision recall curve (AUPRC), recall, and F1 score which are affected by imbalances. These results illustrate the potential utility of applying the presented pipeline for prediction of patient outcomes. The presented prediction results involved no hyperparameter optimization and utilized only the first time point of measurements for each hospital stay. Predictive performance may be further improved by hyperparameter tuning, addressing the class imbalance, using more data point for each patient, and/or implementing a more sophisticated predictive model.

Surrogate explainers were generated to improve cluster explainability. A previous analysis showed that the level of consciousness and SATS are the key criterion for separating clusters.³¹ The importance of the level of consciousness has also been previously identified as the key feature in predicting discharge from ICU.⁵ However, in this study, the level of consciousness appeared less important and SATS only in some cases. It was also shown that the number and type of features varies between clusters. For example, clustering the entire population mostly only utilized five features, whereas subclustering cluster 0 almost exclusively relies on SATS. This result was verified by a 74% overlap between features independently identified by clinicians and the surrogate explainers (Table 4). The integration of surrogate explainers and clinicians helped validate and verify the presented results. Future studies and the deployment in hospital settings should consider this approach to increase fairness, accountability, and transparency. This also aids in building trust between the clinicians and machine learning systems. However, the identified patient subtypes and associated predictive performance should be treated with care, and considered merely illustrative, as the whole analysis is based on a data set from one hospital. Adding data from other hospitals as well as considering additional features may reveal other or alter the identified patient subtypes. Further work should be completed to validate the presented results before clinical application of the pipeline. Furthermore, two clinicians were part of the team in order to co-design the process. Future studies will increase the number of clinicians for additional feedback, increasing the acceptance of and trust in the identified patient subtypes. In summary, once validated, the presented pipeline could become a useful tool to assign patients to subtypes and could thereafter inform clinical decisions for improved patient care.

AUTHORS' CONTRIBUTIONS

EW and JNC participated in all aspects of the study. RSB and AH contributed to the development of the machine learning pipeline. AH extracted the data and assisted in writing the manuscript. CPB and MA clinically evaluated the data. CPB, CJM, RBR, and MA consulted throughout the study.




DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by Health Data Research UK via the Better Care Partnership Southwest (HDR CF0129) within the P-NEWS project (personalized early warning scores for preventing unplanned critical admissions); EW, JNC, AH, and RSR are funded by the UKRI Turing AI Fellowship (grant no. EP/V024817/1).

ORCID IDS

Enrico Werner  <https://orcid.org/0009-0006-3682-3078>
 Jeffrey N Clark  <https://orcid.org/0000-0003-0118-3999>
 Michael Ambler  <https://orcid.org/0000-0001-7884-6522>

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

- Castela Forte J, Yeshmagambetova G, van der Grinten ML, Hiemstra B, Kaufmann T, Eck RJ, Keus F, Epema AH, Wiering MA, van der Horst ICC. Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering. *Sci Rep* 2021;**11**:12109
- Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient subtyping via time-aware LSTM networks. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, Halifax, NS, Canada, 13–17 August 2017, pp.65–74. New York: ACM
- Vranas KC, Jopling JK, Sweeney TE, Ramsey MC, Milstein AS, Slatore CG, Escobar GJ, Liu VX. Identifying distinct subgroups of intensive care unit patients: a machine learning approach. *Crit Care Med* 2017;**45**:1607–15
- Anand RS, Stey P, Jain S, Biron DR, Bhatt H, Monteiro K, Feller E, Ranney ML, Sarkar IN, Chen ES. Predicting mortality in diabetic ICU patients using machine learning and severity indices. *AMIA Jt Summits Transl Sci Proc* 2018;**2017**:310–9
- McWilliams CJ, Lawson DJ, Santos-Rodriguez R, Gilchrist ID, Champneys A, Gould TH, Thomas MJC, Bourdeaux CP. Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. *BMJ Open* 2019;**9**:e025925
- Carr E, Bendayan R, Bean D, Stammers M, Wang W, Zhang H, Searle T, Kraljevic Z, Shek A, Phan HTT, Muruet W, Gupta RK, Shinton AJ, Wyatt M, Shi T, Zhang X, Pickles A, Stahl D, Zakari R, Noursadeghi M, O'Gallagher K, Rogers M, Folarin A, Karwath A, Wickstrøm KE, Köhn-Luque A, Slater L, Cardoso VR, Bourdeaux C, Holten AR, Ball S, McWilliams C, Roguski L, Borca F, Batchelor J, Amundsen EK, Wu X, Gkoutos GV, Sun J, Pinto A, Guthrie B, Breen C, Douiri A, Wu H, Curcin V, Teo JT, Shah AM, Dobson RJB. Evaluation and improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study. *BMC Med* 2021;**19**:23
- Oei SP, van Sloun RJG, van der Ven M, Korsten HHM, Mischi M. Towards early sepsis detection from measurements at the general ward through deep learning. *Intell Based Med* 2021;**5**:100042
- Giannini HM, Ginestra JC, Chivers C, Draugelis M, Hanish A, Schweickert WD, Fuchs BD, Meadows L, Lynch M, Donnelly PJ, Pavan K, Fishman NO, Hanson CW 3rd, Umscheid CA. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Crit Care Med* 2019;**47**:1485–92
- Cheng F-Y, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, Kohli-Seth R, Levin MA, Timsina P, Kia A. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. *JCM* 2020;**9**:1668

10. Rothman MJ, Rothman SI, Beals J 4th. Development and validation of a continuous measure of patient condition using the Electronic Medical Record. *J Biomed Inform* 2013;**46**:837–48
11. Subbe CP, Kruger M, Rutherford P, Gemmel L. Validation of a modified Early Warning Score in medical admissions. *QJM* 2001;**94**:521–6
12. Balkan B, Essay P, Subbian V. Evaluating ICU clinical severity scoring systems and machine learning applications: APACHE IV/IVa Case Study. In: *40th Annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Honolulu, HI, 18–21 July 2018, pp.4073–76. New York: IEEE
13. Khwannimit B. A comparison of three organ dysfunction scores: MODS, SOFA and LOD for predicting ICU mortality in critically ill patients. *J Med Assoc Thai* 2007;**90**:1074–81
14. Gerry S, Bonnici T, Birks J, Kirtley S, Virdee PS, Watkinson PJ, Collins GS. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* 2020;**369**:m1501
15. Johnson AEW, Mark RG. Real-time mortality prediction in the Intensive Care Unit. *AMIA Annu Symp Proc* 2017;**2017**:994–1003
16. Jiang H, Su L, Wang H, Li D, Zhao C, Hong N, Long Y, Zhu W. Non-invasive real-time mortality prediction in intensive care units based on gradient boosting method: model development and validation study. *JMIR Med Inform* 2021;**9**:e23888
17. King Z, Farrington J, Utley M, Kung E, Elkhodair S, Harris S, Sekula R, Gillham J, Li K, Crowe S. Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digit Med* 2022;**5**:104
18. Royal College of Physicians. National Early Warning Score (NEWS) 2: standardising the assessment of acute-illness severity in the NHS—updated report of a working party 2017, 2021. <https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2>
19. Abbott TEF, Cron N, Vaid N, Ip D, Torrance HDT, Emmanuel J. Pre-hospital National Early Warning Score (NEWS) is associated with in-hospital mortality and critical care unit admission: a cohort study. *Ann Med Surg* 2018;**27**:17–21
20. Alam N, Vegting IL, Houben E, van Berkel B, Vaughan L, Kramer MH, Nanayakkara PW. Exploring the performance of the National Early Warning Score (NEWS) in a European emergency department. *Resuscitation* 2015;**90**:111–5
21. Baker KF, Hanrath AT, Schim van der Loeff I, Kay LJ, Back J, Duncan CJ. National Early Warning Score 2 (NEWS2) to identify inpatient COVID-19 deterioration: a retrospective analysis. *Clin Med* 2021;**21**:84–9
22. Kostakis I, Smith GB, Prytherch D, Meredith P, Price C, Chauhan A, Portsmouth Academic Consortium For Investigating COVID-19 (PACIFIC-19). The performance of the National Early Warning Score and National Early Warning Score 2 in hospitalised patients infected by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). *Resuscitation* 2021;**159**:150–7
23. Costa DK, Kahn JM. Organizing critical care for the 21st century. *JAMA* 2016;**315**:751
24. Dlugacz YD, Stier L, Lustbader D, Jacobs MC, Hussain E, Greenwood A. Expanding a performance improvement initiative in critical care from hospital to system. *Jt Comm J Qual Improv* 2002;**28**:419–34
25. Mirnezami R, Nicholson J, Darzi A. Preparing for precision medicine. *N Engl J Med* 2012;**366**:489–91
26. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979;**1**:224–7
27. Geri G, Vignon P, Aubry A, Fedou AL, Charron C, Silva S, Repessé X, Vieillard-Baron A. Cardiovascular clusters in septic shock combining clinical and echocardiographic parameters: a post hoc analysis. *Intensive Care Med* 2019;**45**:657–67
28. Wu W, Bang S, Bleecker ER, Castro M, Denlinger L, Erzurum SC, Fahy JV, Fitzpatrick AM, Gaston BM, Hastie AT, Israel E, Jarjour NN, Levy BD, Mauger DT, Meyers DA, Moore WC, Peters M, Phillips BR, Phipatanakul W, Sorkness RL, Wenzel SE. Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma. *Am J Respir Crit Care Med* 2019;**199**:1358–67
29. Ribeiro MT, Singh S, Guestrin C. “Why should i trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, CA, 13 August 2016, pp.1135–44. New York: ACM
30. Zhou Z, Sun M, Chen J. A model-agnostic approach for explaining the predictions on clustered data. In: *2019 IEEE international conference on data mining (ICDM)*, Beijing, China, 2019, pp.1528–33. New York: IEEE
31. Werner E, Clark JN, Bhamber RS, Ambler M, Bourdeaux CP, Hepburn A, McWilliams CJ, Santos-Rodriguez R. Identification, explanation and clinical evaluation of hospital patient subtypes. In: Shaban-Nejad A, Michalowski M, Bianco S (eds) *Artificial intelligence for personalized medicine—W3PHAI 2023: studies in computational intelligence*, vol 1106. Cham: Springer, 2023, pp.137–49
32. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *JOSS* 2018;**3**:861
33. Campello RJGB, Moulavi D, Sander J. Density-based clustering based on hierarchical density estimates. In: Pei J, Tseng VS, Cao L, Motoda H, Xu G (eds) *Advances in knowledge discovery and data mining*. Berlin: Springer, 2013, pp.160–72
34. Moulavi D, Jaskowiak PA, Campello RJGB, Zimek A, Sander J. Density-based clustering validation. In: *Proceedings of the 2014 SIAM international conference on data mining*, Philadelphia, PA, 24–26 April 2014, pp.839–47. Philadelphia, PA: SIAM
35. Sokol K, Hepburn A, Santos-Rodriguez R, Flach P. bLIMEy: surrogate prediction explanations beyond LIME. In: *2019 Workshop on human-centric machine learning (HCML 2019) at the 33rd conference on neural information processing systems (NeurIPS 2019)*, Vancouver, BC, Canada, 29 October 2019
36. Sokol K, Santos-Rodriguez R, Flach P. FAT forensics: a Python toolbox for algorithmic fairness, accountability and transparency. *Softw Impact* 2022;**14**:100406
37. Pankhurst T, Sapey E, Gyves H, Evison F, Gallier S, Gkoutos G, Ball S. Evaluation of NEWS2 response thresholds in a retrospective observational study from a UK acute hospital. *BMJ Open* 2022;**12**:e054027

(Received June 2, 2023, Accepted October 25, 2023)