BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Discovering patient groups in sequential electronic healthcare data using unsupervised representation learning

Jingteng Li[1,2†], Kimberley R. Zakka[1,2†], John Booth[1,2], Louise Rigny[2,3], Samiran Ray[1,2], Mario Cortina-Borja[1,2], Payam Barnaghi[1,2,3] and Neil Sebire[1,2*]

Abstract

Introduction Unsupervised feature learning methods inspired by natural language processing (NLP) models are capable of constructing patient-specific features from longitudinal Electronic Health Records (EHR).

Design We applied document embedding algorithms to real-world paediatric intensive care (PICU) EHR data to extract patient-specific features from 1853 patients' PICU journeys using 647 unique lab tests and medication events. We evaluated the clinical utility of the patient features via a K-means clustering analysis.

Results We trained a document embedding model under a unique evaluation pipeline and obtained latent patient feature vectors for all 1853 patients. We performed unsupervised clustering to the patient vectors as a downstream analysis and obtained 5 distinct clusters via hyperparameter optimisation. Significant variations ($p<0.0001$) within both patient characteristics and surgery intervention and diagnostic profiles were detected.

Conclusion The K-means clustering results demonstrated the clinical utilities of the patient-specific features learned from the embedding algorithms. The latent patient features obtained via the embedding process enabled direct applications of other machine learning algorithms. Future work will focus on utilising the temporal information within EHR and extending EHR embedding algorithms to develop personalised patient journey predictions.

Introduction

Electronic Health Records (EHRs) are a major source of structured, high-dimensional medical data. Clinical information systems store a wealth of longitudinal clinical information collected throughout a patient's healthcare trajectory [1]. Typically, these systems amalgamate data from various sources such as laboratory analysis, diagnosis records, interventions and medication histories [2]. In clinical practice, analysis of EHR data has the potential to drive the development of personalised healthcare, enabling clinicians to making informed, patient-specific decisions from similar patients' cases [3–5]. Additionally, improvements in care quality and patients' experiences have been attributed to the efficient utilisation of EHR data [6, 7]. However, deriving clinically significant insights from EHR data is often a challenging process [8].

Besides the inherent high dimensionality [9], EHR data analysis exhibits additional challenges such as missing value imputation, sparsity and irregular temporal dependency [10, 11]. The overwhelming diversity of data stored within EHR makes developing clinically applicable solutions that can utilise unprocessed EHR data directly

† Jingteng Li and Kimberley R. Zakka contributed equally to this work.

*Correspondence:
Neil Sebire
neil.sebire@gosh.nhs.uk
[1] Great Ormond Street Institute of Child Health, University College London, London, UK
[2] Data Research Innovation and Virtual Environment, Great Ormond Street Hospital for Children, London, UK
[3] Department of Brain Sciences, Imperial College London, London, UK

obtained from clinical information systems a challenging task. In recent years, machine learning (ML) models have emerged as powerful tools to analyse EHR data. Applications of modern learning algorithms have seen success in many data-intensive industries, and there has been a growing trend in their integration into medical data analysis [12]. While most learning algorithms work well on data with uniform features, real-life EHR data is rarely uniform and typically contains both structured and unstructured data.

Complexity and non-uniformity are not unique to EHR datasets. Many data-driven analysis of real-world data, such as natural language processing (NLP) [13], face similar issues. In many cases, extracting low-dimensional representations from high-dimensional data is the key to learning from complex datasets. NLP employs a common approach known as word embedding. This involves transforming individual word tokens into numerical feature vectors. In other words, word embedding is a feature learning method that learns to summarise all unique words from a given corpus in a vector space. The extracted word features often capture the semantic relationships between words and allow complex algorithmic operations in the embedding space.

In this paper, we adopted a document embedding algorithm, namely, Doc2Vec [14], to extract latent patient-specific features, which, in turn, can be used for downstream applications. Similar word and document embedding algorithms have been previously adopted to clinical domains to construct both code [15–17] and patient-level embeddings [18, 19]. By applying a clustering analysis to obtained patient vectors, we aim to demonstrate that the embedded patient features are effective at capturing patient-specific characteristics and can be used to derive insights from EHR data.

Methods

To determine the practical applicability of NLP models in the context of real-life PICU EHR data, we applied the Doc2Vec family of semantic embedding algorithms to derive patient-specific feature vectors from processed EHR medical sequences and assessed the resulting feature vectors via unsupervised clustering techniques. We imposed no specific medical assumptions during data pre-processing and assessed the clustering outcomes categorically based on patients' demographic and diagnostic profiles. In general, this research adhered to the standard pipeline of unsupervised learning applied to EHR data. The key stages are shown in Fig. 1.

Data collection and pre-processing

This retrospective study was conducted in collaboration with University College London (UCL) and the Great Ormond Street Hospital for Children (GOSH) in London, United Kingdom. GOSH receives 242,694 outpatient visits and 42,112 inpatient visits yearly (as of 2021/2022)

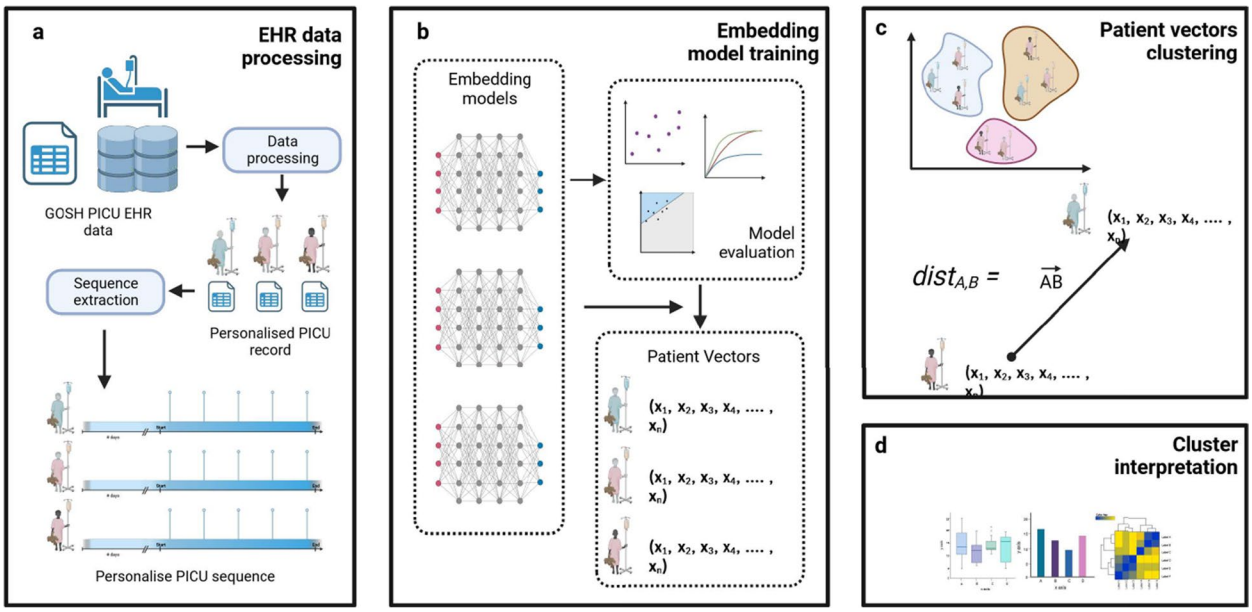


Fig. 1 The working pipeline of this study, **a** PICU EHR data was pulled from GOSH EHR repository and processed to a sequential format compatible with the embedding algorithm, **b** Embedding models were trained on the extracted PICU EHR sequences and were categorically evaluated, **c** Clustering of the obtained patient-specific embedding using distance-based metrics, **d** data-driven cluster interpretation with clinical domain knowledge

and is home to the largest critical paediatric care facility in the United Kingdom [20]. We exclusively used data from its general paediatric intensive care unit (PICU) for this work. All data used in this study was extracted from GOSH's EHR database and was accessed via the GOSH's secured digital research environment (DRE). The records were anonymised based on GOSH's data-sharing protocol, where patients' real names were replaced across all records with randomly generated identifiers. In addition, patients' actual dates of birth were obscured and are only accurate to the exact month.

The PICU EHR data contained 980,182 total entries drawn from the medical histories of 1853 patients, each of whom has been admitted to the PICU at least once during the period spanning May 01, 2019, to January 08, 2023. A more detailed summary of the EHR data is provided in Summary of the EHR data section. The medical records incorporate data from multiple sources and can be primarily categorised into five categories: administrative, demographic, diagnostics, laboratory tests and therapeutic drug interventions. Administrative records provide information about patients' interactions with GOSH, including hospitalisations and ward stays, which we used to identify PICU episodes. Demographic records provided information about patients' profiles. We calculated the patients' age at admission as their age at the beginning of PICU stay and used them as a metric to characterise the patient clusters from the clustering analysis. Due to data anonymisation, some very young patients' ages at admissions are precise only to the level of months. For longitudinal patient records such as diagnostic codes, tests and intervention records, we limited our analysis to entries occurring exclusively within PICU admission episodes.

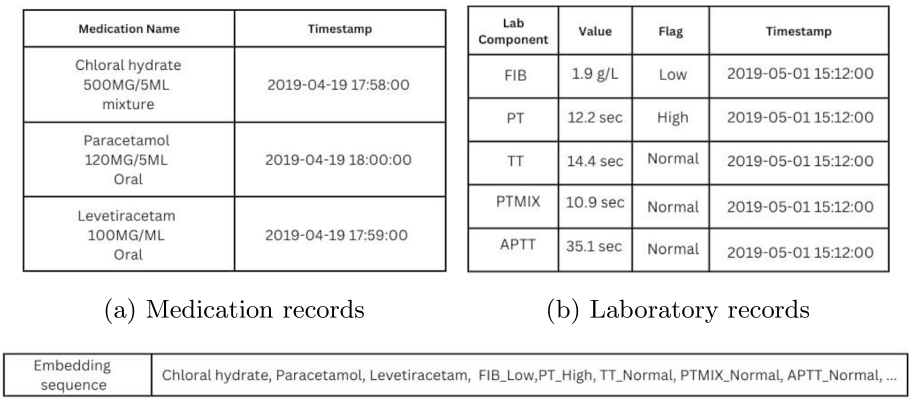
For training the document embedding Doc2Vec model, we constructed medical event sequences for each PICU episode using diagnosis, laboratory tests and medication records. PICU episodes without any medication or laboratory records were excluded or merged into other PICU episodes where appropriate, as shown in Summary of the EHR data section. We expected only categorical sequences for training Doc2Vec and aggregated all numerical lab test results to categorical features based on reference ranges. The PICU sequences are time-ordered, and duplicate events within an hour time-frame were removed.

Generating EHR medical sequences

In natural language processing applications, it is a common practice to treat each unique word from a given corpus as a 'token' and each sentence as a sequence of tokens. A similar approach to treat each patient's journey as a sequence of medical events has been previously adopted by several EHR embedding models [21–23]. However, the temporal sequence of information and events within a patient's journey holds significant implications on interpreting and presenting that journey and for the development of predictive and decision-support models.

In this study, we modified the data processing strategy from a Bidirectional Encoder Representation from Transformer (BERT) [24] extension model known as Med-BERT [23]. Instead of using observed ICD-10 codes, we focused exclusively on frequently-occurring laboratory and medication events and treated each unique laboratory and medication event as a medical event 'token'. Each patient's journey in the PICU was represented as a time-ordered sequence of medical tokens, which we then used to train the Doc2Vec embedding model. Each laboratory test component is combined with its associated categorical result (e.g., normal, high, and low) to form unique tokens for all possible combinations of test components and results. Concerning medications, only the names of the medications are preserved. An illustrative Doc2Vec training sequence is shown in Fig. 2. For duplicated tokens occuring within less than an hour, such as repeated medication or lab tests with the same outcomes, we retained only the first token to simplify the training sequences. Repeated tokens close to each other contribute minimal information when training word embedding networks and are often removed as 'stop words' in NLP.

Rare tokens, defined as occurring in fewer than five patients' journeys, were excluded from training. As explained in Appendix A, Doc2Vec assesses each token contextually with its surrounding tokens, making it less likely to learn useful embeddings for rare tokens due to a lack of contextual information. In practice, the original Word2Vec [25] implemented a frequency-based sampling routine to counter-balance frequent and rare tokens during training, but it is unclear how such a method can be translated to our study. We selected 5 as a minimum threshold to filter out a subset of very rare events that typically occurred only once or twice across the whole dataset. Since the number of tokens affects the number of learnable parameters, we removed tokens with very few occurrences to limit the size of our embedding models. We note that in embedding EHR data, the rate of occurrence is not equivalent to informativeness; intuitively, rare medical events often provide more information about a patient than common events. Limiting the training process to using only frequently-occurred tokens will reduce the precision of the resulting patient-specific embeddings. In total, out of the initial pool of 232 lab tokens and 761 medication tokens, 169 (73%) lab tokens



Medication Name	Timestamp
Chloral hydrate 500MG/5ML mixture	2019-04-19 17:58:00
Paracetamol 120MG/5ML Oral	2019-04-19 18:00:00
Levetiracetam 100MG/ML Oral	2019-04-19 17:59:00

(a) Medication records

Lab Component	Value	Flag	Timestamp
FIB	1.9 g/L	Low	2019-05-01 15:12:00
PT	12.2 sec	High	2019-05-01 15:12:00
TT	14.4 sec	Normal	2019-05-01 15:12:00
PTMIX	10.9 sec	Normal	2019-05-01 15:12:00
APTT	35.1 sec	Normal	2019-05-01 15:12:00

(b) Laboratory records

Embedding sequence	Chloral hydrate, Paracetamol, Levetiracetam, FIB_Low,PT_High, TT_Normal, PTMIX_Normal, APTT_Normal, ...

(c) Training sequence

Fig. 2 Example of a **c** Doc2Vec training sequence using a fragment of a patient's PICU journey from 10/04/2019 to 16/06/2019, **a** medication records and **b** laboratory results are processed into **c** Doc2Vec training sequences. Medication names are often reported together with dosages, so regular expressions were used to isolate the names of medications

and 478 (63%) medication tokens were retained for training the embedding algorithms.

Experiments and validations

We employed various regression and classification models alongside the main embedding/clustering models to assess the performance of our models and validate the choice of parameter values. The $\mathcal{L}2$ regularised linear regression was used for numerical data, the logistic regression for binary classification, and the one-vs-one support vector classifier for multiclass classifications. The area under the curve (AUC) was calculated for the precision-recall curve due to low positive rates for binary predictors [26]. Unless stated otherwise, all learning-based validation studies were repeated 50 times with 80 – 20 train test splits, and the average accuracy on the test set was reported. For K-means clustering analysis, the distribution of categorical variables among clusters was tested via Pearson's chi-squared test. Similarly, variations of numerical variables among clusters were tested via the Kruskal-Wallis test. To ensure we identify only significant characteristics among clusters from a diverse input, a p-value smaller than 0.0001 was considered to be significant.

Application of embedding algorithms

After pre-processing the EHR data, every PICU episode is represented by an ordered sequence of medical event tokens. While several EHR-specific feature learning/embedding algorithms exist[21–23], these algorithms are often designed with general visit-level data in mind and would translate poorly to longitudinal records generated within the PICU. Instead, we opted for Doc2Vec family of document embedding algorithms which offer a relatively versatile embedding network capable of processing sequential data, such as strings and sentences.

Doc2Vec [14] is a family of document embedding algorithms inspired by language neural networks and functions similarly to the word embedding algorithm Word2Vec [25]. Word2Vec was developed to extract token-specific features from a given corpus. Doc2Vec inherits the word embedding concept and extends it to construct document-specific features. Doc2Vec can learn sequence embedding in two different ways, namely, the distributed bags of words (D2V-dbow) model, and the distributed memory(D2V-dm) model. The D2V-dbow model forces the prediction network to predict tokens appearing in a sequence based on its vector embedding whilst the D2V-dm model predicts neighbouring tokens around a given token in a sequence. In our study, we treated each unique PICU sequence as a 'sentence' and used Doc2Vec to construct sequence-specific vectors. A detailed description of the Doc2Vec embedding algorithm is provided in Appendix A.

Overall, apart from the choice of learning methods, two additional Doc2Vec model parameters required tuning: the window size W and the embedding vector size N. The window size affects how token predictors are drawn during training, and the vector size determines the length of the embedded sequence vectors. Since the embedding size N affects the number of internal variables in the embedding model, we used the elbow method to determine the optimal choices of N to avoid overfitting. Experiments were repeated at different window sizes W to find an optimal combination between window size W and vector size N. The results

of the evaluation studies are discussed in EHR data embedding and representation section.

Application of k-means clustering

One notable characteristic of word embedding algorithms is that the distance between learned word embeddings reflect semantic similarities [25]. Previous research in developing EHR embedding algorithms has used K-means clustering to quantitatively evaluate the quality of the obtained features [21]. In this study, we applied K-means clustering to obtain clusters of patients that exhibit closer proximity in the embedding space.

K-means clustering is a non-parametric clustering algorithm in which data points are divided into K clusters based on Euclidean distances. The number of clusters K is the primary modelling parameter. In this study, we determined the optimal value of K through stability analysis supplemented with the silhouette score. The clustering stability analysis is a series of bootstrap methods designed to test the reproducibility of the obtained clusters [27]. The silhouette coefficient[28] is a distance-based measurement of cluster separations, normalised between −1 and +1. A silhouette score close to +1 implies well-separated clusters, while a silhouette score close to −1 generally indicates erroneous cluster assignments.

Model evaluation

The Doc2Vec embedding models and the K-means clustering results were evaluated independently to avoid confirmation bias. We selected the optimal D2V model based on the quality of obtained token/sequence embeddings and used the elbow method to determine the optimal model. The resulting patient-specific features from the selected D2V model were used during K-means clustering.

The goal of evaluating Doc2Vec embedding models was to identify an optimal combination of model variation (D2V-dbow and D2V-dm), window size W and embedding size N. The Doc2Vec model performance was evaluated based on the conformity of the obtained token features and the mortality prediction accuracy of the patient-specific features. We employed the conformity analysis from Med2Vec [21] and used unsupervised clustering on the obtained laboratory test token features to 'recover' the components of the laboratory tests. We used the adjusted mutual information score between the lab token clusters and their original component labels to measure the conformity of the learned token features. High conformity was preferred as it implies that laboratory tests of different components are distinguishable in the embedding space. For predicting mortality, we used patient-specific features as input and took the average AUC to measure the 'informativeness' of the patient features.

Patient-specific features extracted by the optimal Doc2Vec model were used as features for the K-means embedding algorithm. The optimal number of clusters was determined via stability analysis and silhouette score. To investigate the homogeneity of patients in clusters, we compared the distribution of demographics, diagnoses and surgeries among the clusters. In particular, we were interested in the distributions of specific diagnoses between patient clusters. We isolated the diagnostic records for all patients and performed Pearson's chi-square for all unique ICD-10 diagnoses. Diagnoses with skewed distributions among clusters ($p<0.0001$) were subjected to an additional two-tailed Spearman's correlation test to identify diagnoses strongly correlated to a specific cluster.

Results

Summary of the EHR data

We used EHR for 1853 patients who required pediatric intensive care from May 01, 2019, to January 08, 2023 (53.75% Male, 44.20% Female). The PICU dataset contained 980,182 total entries in 3193 PICU episodes across 1853 patients. Of these, 403 patients required more than one PICU admission. For patient admitted to PICU multiple times during disjointed hospital admissions, we only included PICU episodes in their longest continuous admission. PICU episodes less than 5 hours apart were merged to account for patients receiving procedures outside of PICU. For PICU episodes without medication or laboratory records (481,15.9%), we first checked if the episode could be merged with adjacent episodes and excluded episodes that cannot be merged. In total, 222(7.0%) PICU episodes were excluded, and 423(13.9%) PICU episodes merged into other episodes, leaving 915,939 entries in 2548 PICU episodes across 1853 patients. Patient demographic information is summarized in Table 1.

For training the embedding model, we included 169 laboratory tests and 478 medication tokens that appeared in more than 5 PICU episodes. Numerical laboratory test results were categorised according to reference ranges, and among 169 laboratory tests, we included all observed categorical outcomes for 87 lab components. Since GOSH uses internal codes for medication that included both medication name and dosage, we opted to retain solely the medication names.

To generate sequences for embedding, 2548 PICU sequences were constructed using the laboratory test results and the medication records. We applied Doc2Vec embedding to all 2548 PICU sequences to obtain sequence embedding. Embedded features for

Table 1 Selected characteristics of patients (n=1853) admitted to PICU from May 01, 2019 to January 8, 2023

Characteristic	n	Missing values n(%)
Median age at admission (IQR) (years)	2.5 (9.4)	0(0%)
Age group n (%)		0(0%)
Term neonatal (0-28d)	183(9.9%)	
Infancy (28d-12mo)	519(28%)	
Toddler (13mo-2y)	167(9%)	
Early childhood (2-5y)	311(16.8%)	
Middle childhood (6-11y)	337(18.2%)	
Early adolescence (12-18y)	336(18.1%)	
Late adolescence (19-21y)	0(0%)	
Ethnicity n (%)		169 (9.1%)
White	725 (39.1%)	
Asian	283 (15.3%)	
Prefer not to say	244 (13.2%)	
Other	191 (10.3%)	
Black	158 (8.3%)	
Mixed	83 (4.5%)	
Sex n (%)		0 (0%)
Female	819 (44.2%)	
Male	1,033 (53.8%)	
Indeterminate	1 (0.05%)	
Median LOS (IQR) (PICU,days)	3.4(4.9)	0(0%)
Total deaths n (%)	228 (12.3%)	

PICU sequences belonging to the same patient were averaged to form the patient-specific features, resulting in one feature vector for each patient. We used each patient's age at the start of their PICU admission period as their age on admission. We summed their time spent in PICU (during the admission period) to calculate their length of stay.

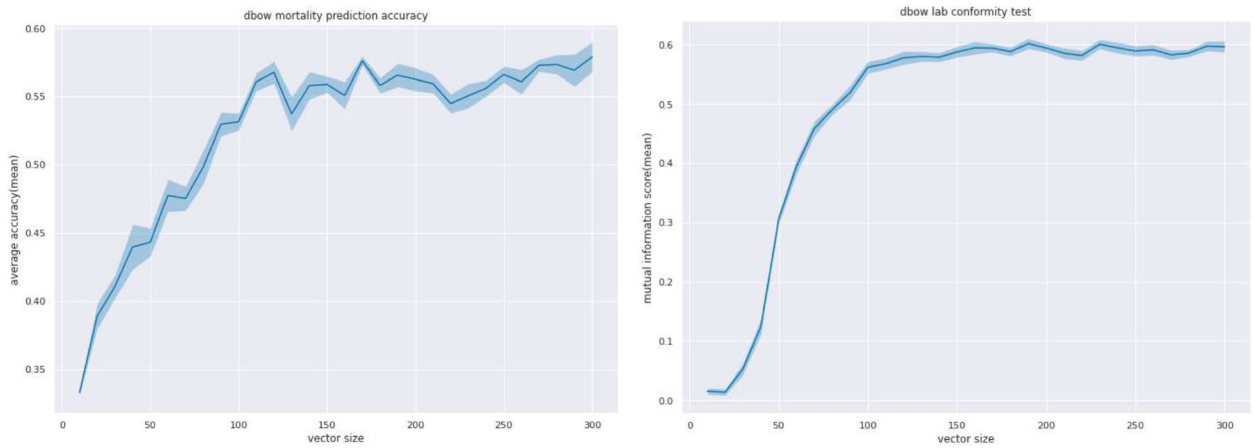
For patient diagnosis, we used ICD-10 codes provided at default levels. Diagnoses within the ICD chapters XVIII 'Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified' and XXI 'Factors influencing health status and contact with health services' were excluded from our analysis. Only diagnoses made during PICU stay were used for cluster analysis. In total, 2678 unique ICD-10 diagnoses were observed, of which 255 were unique primary diagnoses and 2423 were unique co-morbidities. The most common primary diagnosis was ICD-10 code J398: other specified diseases of the respiratory tract, while ICD chapter XVII 'Congenital Malformations, Deformations and Chromosomal Abnormalities' contained the largest number of patients. We identified 417 patients who required a surgical procedure during their PICU admission, 354 of which were unscheduled/emergency. The most common procedure involved microtherapeutic endoscopic operations on the larynx.

EHR data embedding and representation

Based on the lengths of PICU sequences and the number of tokens, we tested window sizes from 5 to 10 and vector sizes from 10 to 300 (at increments of 10) for both D2V-dbow and D2V-dm. Since the D2V-dbow model does not construct token embeddings independently, we trained skip-gram Word2Vec alongside D2V-dbow to obtain token-specific features for conformity evaluation. Overall, D2V-dbow demonstrated better and more stable performance across both evaluation tasks. For both models, the optimal window sizes were apparent. While the mortality prediction accuracy against vector sizes converged to similar values for all window sizes, the conformity plots overwhelmingly favoured smaller window sizes. The optimal window size was 5 for both the D2V-dbow and D2V-dm models. The mortality prediction accuracy and the lab token conformity plots for D2V-dbow with window size $W = 5$ are shown in Fig. 3. Evaluation metrics for all parameter choices are shown in the Appendix.

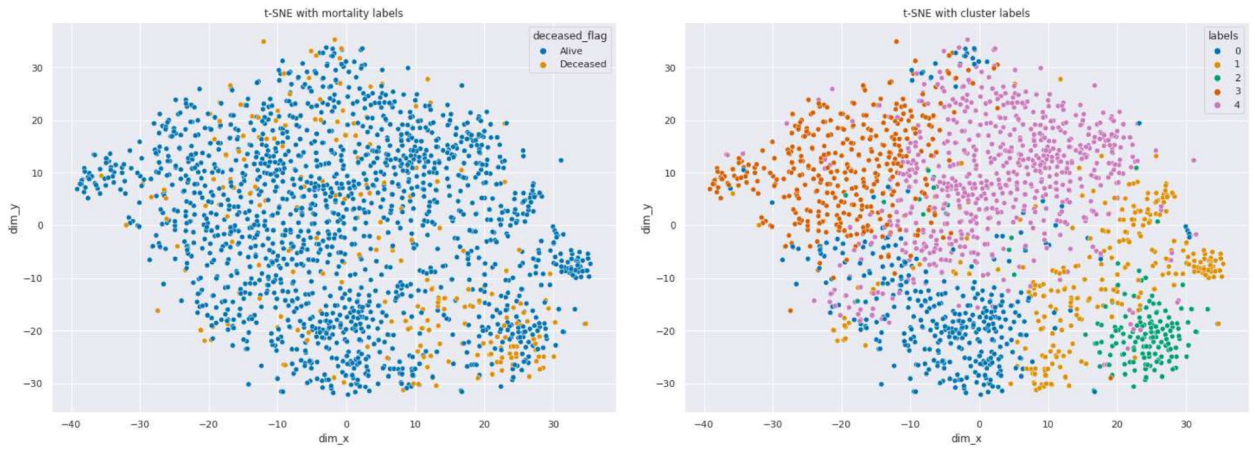
As shown in Fig. 3, both evaluation metrics produced an elbow-shaped curve. The mortality prediction accuracy and the laboratory token embedding conformity increased as the vector size increased. The outcome was expected as the embedding model contained more parameters, and the convergence of evaluation metrics was likely caused by the constraint on the intrinsic dimensions of the embedded vectors. Since the knee points for both plots occurred at the vector size $N = 100$, we considered it the optimal embedding size for the D2V-dbow model. We used t-distributed stochastic neighbour embedding (t-SNE) to project the patient-specific features on a 2D plane for visualisation. The t-SNE plot for patient-specific features obtained via D2V-dbow is shown in Fig. 4a, with patient mortality as labels.

We additionally applied intrinsic dimension analysis [29, 30] to patient-specific features obtained at all vector sizes for the D2V-dbow model. We investigated the number of principal components needed to explain 95% of the variances in patient-specific features. The results showed that the intrinsic dimensions of the patient-specific features started to converge at vector size $N = 210$, suggesting the optimal vector size should be lower than $N = 210$. We also investigated the possibility of using patient-specific features obtained by D2V-dbow to predict the patient length of stays in PICU ($R^2 = 0.34$). We retained patient-specific features produced by D2V-dbow with window size $W = 5$ and vector size $N = 100$ as features for the K-means clustering analysis.



(a) Mortality regression (b) Laboratory conformities

Fig. 3 **a** The AUCs for predicting patient mortality with 95% CI (shaded) and **b** the conformities of lab token embeddings for D2V-dbow at window size $W = 5$



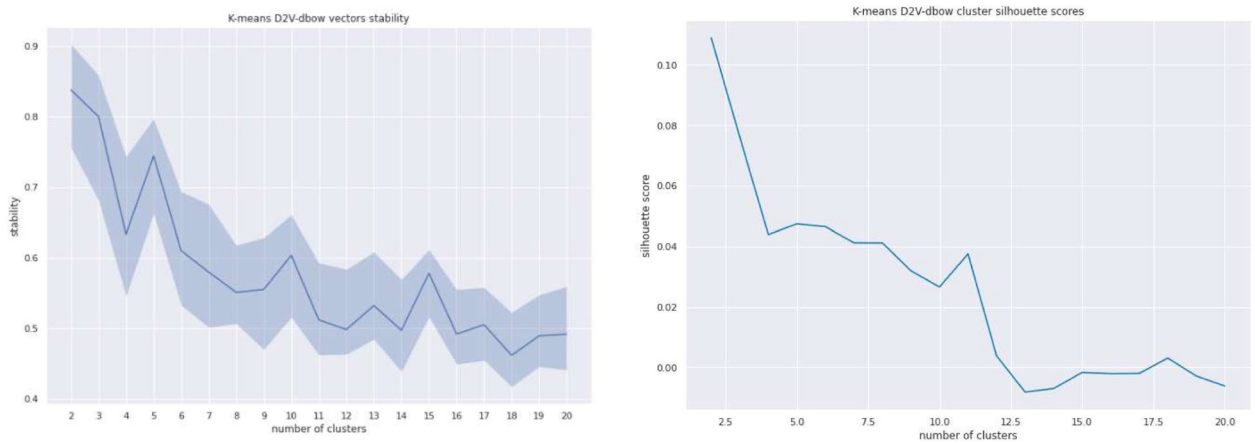
(a) Mortality t-SNE (b) Cluster t-SNE

Fig. 4 The t-distributed stochastic neighbour embedding (t-SNE) visualisation of the D2V-dbow patient-specific features with **a** mortality and **b** cluster labels. Cluster 0 contains a large number of children with atrial septal defects who require some type of neurosurgery. Cluster 1 combines patients with genitourinary diseases and their known complications. Cluster 2, with the highest mortality, consists of children with neoplasms. Cluster 3 contains younger patients with infectious conditions of the respiratory tract, which usually require prolonged hospital stays. Cluster 4, with shorter length-of-stay and low mortality, represents children post-surgery, particularly spinal surgery

K-means clustering

The choice of the optimal number of clusters was not immediately evident from the heuristics alone. The obtained patient-specific features were not well-separated and contained potential outliers that affected the clustering process and the validation experiments. Outlier data within the patient-specific features were expected as the patient cohort was diverse, with significant variations in age and length of stay. While outliers in the patient feature embedding space were detected via Local outlier factors, no patients were removed to remain consistent with the initial exclusion criteria. Since low Silhouette scores were observed for all choices of K, we used stability plot Fig. 5a as the main criteria for determining the optimal value for K. Additionally, the Silhouette scores across all choices of K are shown in Fig. 5b.

Stability above 0.7 was only observed for $K = 2, 3, 5$, and silhouette scores close to zero were observed for all



(a) Cluster stability with IQR (b) Silhouette score

Fig. 5 The **a** cluster stability and **b** cluster silhouette score at different choices of *K* for the K-means clustering, the cluster stability was used as the main heuristic for choosing *K* due to overall poor silhouette score

clustering results. The low silhouette scores were likely caused by a combination of factors including imperfect separation, high dimensionality and presence of outlier data. Mortality prediction via cluster memberships for $K = 2, 3, 5$ showed that $K = 5$ ($AUC = 0.45$) performed better on average than $K = 2$ ($AUC = 0.25$) and $K = 3$ ($AUC = 0.27$). We therefore selected $K = 5$ as the optimal number of clusters based on clustering stability and mortality prediction accuracy. The t-SNE plot for D2V-dbow patient vectors with cluster labels is shown in Fig. 4b.

Cluster characteristics

Selected cluster characteristics are shown in Table 2. Significant variations ($p<0.0001$) of patient mortality, age at admission and length of stay were observed (Fig. 6).

Patient mortality varied significantly across the clusters (Fig. 6a). Larger clusters 0, 3 and 4 had mortality rate lower or close to the global rate. Cluster 1 and 2 had elevated mortality rates, with more than a third of patients dying in cluster 2. The area under the precision-recall curve for predicting patient mortality via D2V-dbow patient vectors was 0.58, significantly higher than the expected baseline *AUC* (0.12) [26, 32].

The median length of stays differed among clusters. Cluster 4 had a significantly lower length of stay, with 75% of patients staying less than three days. Cluster 3 and Cluster 2 had a longer average length of stay. In contrast, Cluster 1 had an average length of stay closer to the global average. The age at admissions also differed significantly ($p<0.0001$) among clusters. Patients in clusters 0 and 3 were admitted much younger, while Clusters 1,2,4 had more varied age profiles among clusters. Patients in cluster 4 are older on average.

Cluster diagnoses

Variations of patient characteristics among clusters indicates distinguished diagnostic profiles among the clusters. We investigated the distribution of ICD-10 diagnoses among the clusters based on frequencies. In this case, each unique ICD-10 diagnosis was only counted once for each patient, even if it was recorded multiple times during their stay in the PICU. Table 3 shows the top 5 most frequent ICD-10 diagnoses in each cluster.

Some unique diagnostic characteristics were observed. More than 30% of patients in cluster 1 had unspecified acute renal failure and more than 50% of patients in cluster 2 had Agranulocytosis or Thrombocytopenia. In particular, cluster 2's top five most frequent diagnoses differed significantly from others. Due to the diversity of ICD-10 diagnoses observed in the PICU EHR data, we additionally studied the distribution of each ICD-10 diagnosis separately. In total, we identified 161 recurring ICD-10 diagnoses with unexpected distributions (Pearson Chi-square,$p<0.0001$) among clusters. We then used Spearman's correlation test with a two-tailed alternative hypothesis ($p<0.0001$) to identify the diagnoses' correlations with cluster assignments. Figure 7 shows the correlation coefficient between ICD-10 diagnoses (averaged among ICD-10 chapters) and cluster assignments.

The distributions of all 161 significant recurring diagnoses are included in the Appendix. Outstanding diagnostic characteristics were observed, such as cluster 2's strong correlations with diagnoses in neoplasms and blood and blood-forming organs or cluster 3's positive correlations with respiratory diagnoses. Cluster 4 had positive correlations with diagnoses related to mental development and the musculoskeletal system. In

Table 2 Various characteristics of cluster (0–4)

Characteristic	0	1	2	3	4
n	374	276	147	393	662
total %	20.2	14.9	7.9	21.2	35.8
Death n	28	54	50	43	53
Mortality %	7.5	19.6	31.0	11	8.1
Male %	59.1	48.6	57.1	60	54.2
LOS					
LOS mean	4.9	5.4	8.2	11.6	3.1
LOS std	4.3	8.5	11.5	20.0	5.1
LOS 25%	2.0	1.5	1.9	2.7	1.0
LOS 50%	4.0	3.5	6.0	8.1	1.9
LOS 75%	6.0	6.1	10.0	11.2	3.0
Age group n (%)					
Term neonatal (0-28d)	73(9.9)	56(20.3)	2(1.4)	20(5.1)	32(4.8)
Infancy (28d-12mo)	149(39.8)	34(12.3)	38(25.9)	199(50.6)	99(15.0)
Toddler (13mo-2y)	36(9.6)	13(4.7)	19(12.9)	57(14.5)	41(6.2)
Early childhood (2-5y)	64(17.1)	48(17.4)	25(17.0)	66(16.8)	108(16.3)
Middle childhood (6-11y)	35(9.4)	66(23.9)	40(27.2)	35(8.9)	161(24.3)
Early adolescence (12-18y)	17(4.5)	59(21.4)	23(15.6)	16(4.1)	221(33.4)
Late adolescence (19-21y)	0(0)	0(0)	0(0)	0(0)	0(0)
ICD chapters					
Top	Respiratory	Metabolic	Blood	Respiratory	Congenital
Top %	54.0	74.0	84.4	89.3	57.0
Secondary	Nervous	Circulatory	Circulatory	Congenital	Nervous
Secondary %	50.8	66.7	77.6	62.6	55.4
ICD chapters (Primary)					
Top	Congenital	Congenital	Neoplasms	Respiratory	Congenital
Top %	50.0	24.3	41.0	38.7	24.3
Secondary	Respiratory	Neoplasms	Blood	Congenital	Neoplasms
Secondary %	16.7	21.6	12.8	34.2	16.5

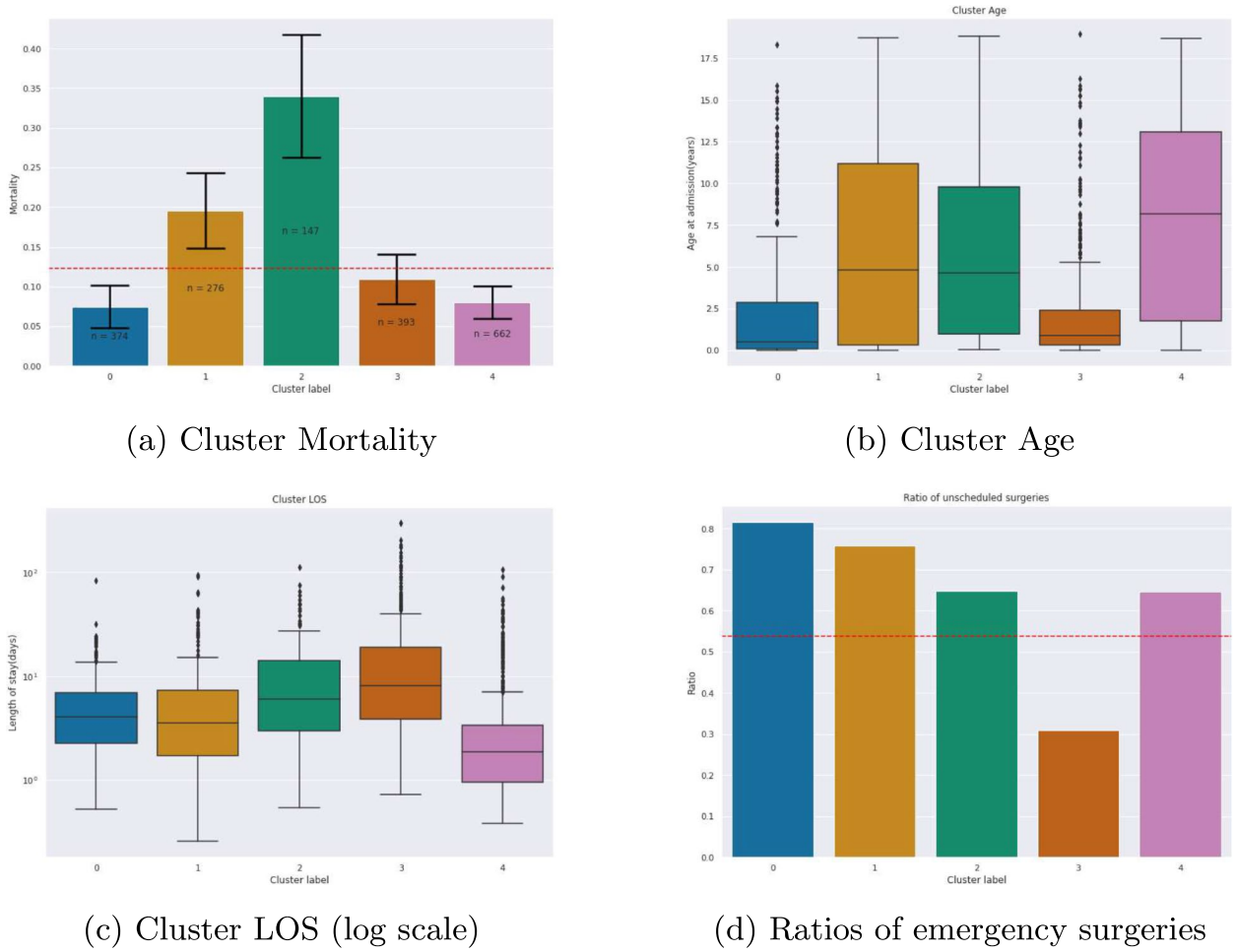
Total % is the total percentage of total population ($n = 1853$) in each cluster. Length of stay (LOS) is the total time spent in PICU in days, expressed as mean and quartiles. Age at admission is expressed as % of patients in a given age group. Age groups are calculated based on the NIH-recommended age strata [31]. Top and secondary diagnoses chapters are the most common ICD-10 diagnoses chapters in clusters, measured as % of patients in clusters with diagnoses in a given chapter

particular, cluster 4 contained almost all cases of Neuromuscular scoliosis (86 of 96). Most (78%) of the observed instances of neurosurgeries were in clusters 0 (41 of 114) and 4 (48 of 114), while most (70%) of the patients who received Ear, Nose & Throat (ENT) surgeries were in cluster 3 (102 of 144).

Discussion

Electronic health data are difficult to process and analyze because of data type heterogeneity, label inconsistencies, and missing values [11]. Through the embedding method, we demonstrated that it is possible to construct patient-specific features from real-world EHR data with minimal data processing and feature selection, and to use those features in downstream machine learning algorithms. In this study, K-means clustering was used to explore the phenotypes of PICU patients at GOSH. Five clinically sound clusters with unique characteristics were obtained.

The profiling reported that patients with genitourinary diseases such as acute renal failure and their known complications, including fluid overload, acidosis and pleural effusion, were predominately grouped into Cluster 1. Cluster 2 gathers patients with longer lengths of stay and highest mortality. These patients had a strong positive correlation with a diagnosis of neoplasms and were more likely to have agranulocytosis and thrombocytopenia, which are commonly seen as part of the disease process or secondary to chemotherapy [33]. Cluster 3 demonstrated a majority grouping of younger patients with infectious conditions of the respiratory tract, which usually require a more prolonged hospital course for antibiotic management and respiratory support. These patients were also found to have atelectasis and atrial septal



(a) Cluster Mortality

(b) Cluster Age

(c) Cluster LOS (log scale)

(d) Ratios of emergency surgeries

Fig. 6 Plots for **a** Cluster mortality with 95% CI against global mortality (line), **b** Cluster Length of stay (LOS) on the log scale, **c** Cluster age at admission (accurate to months) and **d** ratio of emergency surgeries across the clusters with the global ratio (line). The area under the precision-recall curve for predicting patient mortality via patient vectors = 0.58. R^2 for predicting LOS via D2V-dbow patient vectors = 0.34

defects (ASD). In a recent study, Nyboe *et al.* showed that patients with ASD had a significantly higher risk of hospitalization for pneumonia and used more antibiotics than those without ASD [34]. More than half of the patients in cluster 3 underwent surgeries performed by otolaryngology specialists; most commonly rigid bronchoscopy, microtherapeutic endocopic operations on the larynx, and tracheostomies. Membership in cluster 4 was strongly correlated with having a diagnosis of musculoskeletal disease and is associated with a significant number of neurosurgeries.

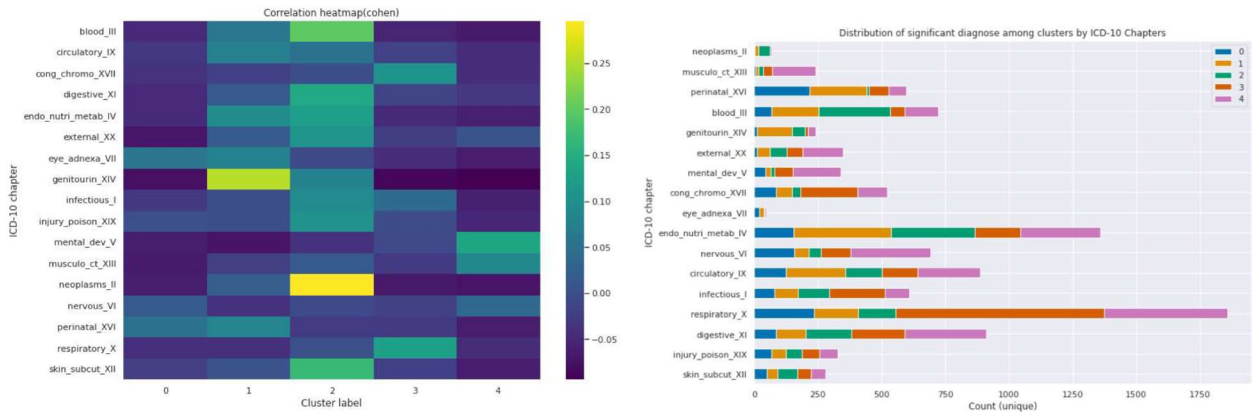
Despite known correlations in the literature, the data-driven approach of this study uncovers patterns in health data that warrant further investigation. The presented analysis showed that a large number of children with ASD required neurosurgery (see cluster 0). Herein lies the importance of developing robust methods to process and use the routinely collected data to gain in-depth insights from past patient experiences. Our clustering results showed that despite being constructed from fragmented and potentially noisy EHR data, the patient vectors could preserve patient-specific characteristics which may not be immediately apparent to human perception. We developed a pipeline for applying token embedding models to EHR data, complete with metrics for model tuning and evaluation. Beyond the evaluations, the learning process constructed uniform patient-specific features fit for other learning algorithms with minimal supervision.

The main strength of this analysis lies in its potential utility, extracting uniform latent features from EHR data with minimal supervision, offering improved modelling and versatility for various downstream tasks. These features may enhance model generalisability by capturing shared patterns across larger, more diverse EHR data. Digitally-accessible EHR provides ample opportunities

Table 3 Top 5 most common ICD-10 diagnoses among the clusters

	0	1	2	3	4
1st diag	Atrial Septal defect	Acute renal failure unspecified	Agranulocytosis	Pulmonary collapse	Gastroesophageal reflux disease without oesophagitis
1st diag %	21.9	30.1	51.7	33.1	19.0
2nd diag	Pulmonary collapse	Pleural effusion, not elsewhere classified	Thrombocytopenia	Atrial septal defect	Pulmonary collapse
2nd diag %	17.6	27.5	51.0	28.2	18.5
3rd diag	Acidosis	Acidosis	Hypokalaemia	Gastroesophageal reflux disease without oesophagitis	Constipation
3rd diag %	15.5	26.1	36.7	26.2	17.5
4th diag	Hypotension, unspecified	Fluid overload	Secondary hypertension, unspecified	Lobar pneumonia unspecified	Other surgical procedures
4th diag %	15.2	25.4	36.1	19.6	14.8
5th diag	Other complications following infusion, transfusion and therapeutic injection	Hypotension, unspecified	Fluid overload	Other specified diseases of upper respiratory tract	Pleural effusion, not elsewhere classified
5th diag %	12.6	23.9	33.3	19.6	14.8

The percentage denotes the percentage of patients in clusters with a given diagnosis



(a) Correlation heatmap (b) Distribution of ICD-10 chapters

Fig. 7 **a** Correlation heatmap between ICD-10 diagnoses (averaged among chapters) and cluster assignments. Lighter areas indicate positive correlations, while near-dark areas indicate negative correlations. **b** The stacked bar chart indicates the distributions of the ICD-10 diagnoses (summed by chapters) among clusters. Diagnoses in ICD-10 chapters XVIII ('Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified') and XXI ('Factors influencing health status and contact with health services') were not included during the analysis

for applying modern machine learning algorithms. At the moment, this potential remains under-utilised due to a lack of uniform features in EHR data.

While extracting latent features from EHR via word embedding models has been previously investigated, our method is distinct in its focus on extracting patient-specific features. The extraction of uniform latent features holds several key uses: they may further reduce data complexity by simplifying the representation of patient records, allowing for increased interpretability, as well as enhancing the efficiency of the model. The downstream applications of these features include risk and outcomes prediction, disease classification and forecasting. Improved modelling will facilitate these tasks and enable improved research for personalised healthcare and clinical support. We believe the potential benefit of

constructing patient-specific features would be more beneficial for personalised healthcare.

The main drawback of the current method is the lack of temporal information. Although we ordered medical events temporally for training the Doc2Vec embedding model, we did not explicitly utilise the temporal information. Time is a major factor in making clinically significant predictions, which is why many clinical prediction models favoured recurrent structures. In addition, we did not address the temporal data entry irregularities in the PICU EHR dataset. A potential future research direction would be to address the temporal properties of the EHR data via attention mechanisms [35] or time-sensitive positional encoding. Additionally, since we focused on applying our method to support an unsupervised analysis of GOSH's PICU cohort, we further acknowledge the possibility of validating our method on larger and publicly available EHR databases such as the MIMIC-III [36].

We have also made compromises to generalise the data pre-processing. Only patient medication and lab tests were included during the embedding process. Additional variables relevant to intensive care, such as vital signs and support needed, were not included during embedding or analysis. Factors such as irregular sampling of medical events were not accounted for in favour of scalability and generalisation. Temporal irregularities could potentially explain why D2V-dm, more sensitive to the ordering of tokens, performed worse than the D2V-dbow model.

Conclusion

In this paper, we investigated the possibility of using NLP-inspired feature learning algorithms to extract patient-specific features useful for downstream analytical tasks. We demonstrated the utility of the learned patient features by creating patient clusters. Five clinically relevant clusters with unique characteristics were identified using an unsupervised method. The profile analysis conducted for each cluster highlighted dominant patient data and features in each sub-group. The analysis identified variations within both patient characteristics, surgery intervention and diagnostics. Notably, a significant number of children with atrial septal defects that require neurosurgery were identified.

Overall, the findings from this study underscore how routinely collected EHR data can be used for regular observations and clinically valuable analysis. We demonstrated the ability to identify nuanced patterns that may not be apparent to human perception in day-to-day clinical practice. These findings may extend to improving the clinical practice and providing more in-depth insights into features and outcomes in common patient groups. The latter will have an effect on improving the clinical practice and provide more in-depth insights into the overall patient groups and their common features and outcomes. Future research will focus on developing time-sensitive feature learning algorithms to utilise temporal information within EHR data. Additionally, the possibility of making predictions using patient-specific features learned from time-censored EHR data should be explored.

Abbreviations

AUC	Area under the curve
ASD	Atrial septal defects
dbow	Distributed bags of words
dm	Distributed memory
EHR	Electronic health records
ENT	Ear, nose & throat
GOSH	Great ormond street children's hospital
ICD	International classification of diseases
LOS	Length of stay
ML	Machine learning
NLP	Natural language processing
NIH	National institutes of health
PICU	Paediatric intensive care unit
t-SNE	t-distributed stochastic neighbour embedding
UCL	University College London
UK	United Kingdom

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02812-9.

Supplementary Material 1.

Acknowledgements

Not applicable

Authors' contributions

JL, KZ: Conceptualisation, Data processing, Software, Computational analysis, Writing, Review and Editing. KZ: Clinical interpretation. JB: Data extraction. SR: Reviewing, Clinical interpretation. LR: Reviewing, proofreading. MB: Reviewing. NS,PB:Conceptualisation, Methodology, Writing, Review and Editing, Supervision, Funding acquisition.

Funding

This work was supported by the Great Ormond Street Hospital (GOSH) Charity (grant number 21PP30).

Data availability

Access to de-identified GOSH PICU patient data is restricted to researchers affiliated with GOSH and are not available for public sharing. Python packages used in this project are listed in the appendix. Additional supporting data and code required to reproduce the results using GOSH's PICU data will be made available upon resonable request.

Declarations

Ethics approval and consent to participate

The use of de-identified, routinely collected electronic healthcare data was approved by the London-South East Research Ethics Committee under REC approval [21/LO/0646]: Use of routine healthcare and operational hospital data for research. The research data obtained from GOSH DRE was de-identified in which patient-identifiable information has been redacted. In accordance with the UK Policy Framework for Health and Social Care Research, the requirement for informed consent in this study were waived following approval from GOSH.

Consent for publication
Not applicable.

Competing interests
The Authors declare no competing financial or non-financial interests.

Received: 10 July 2024 Accepted: 8 December 2024
Published online: 28 January 2025

References

1. Ehrenstein V, Kharrazi H, Lehmann H, et al. Obtaining Data From Electronic Health Records. In: Gliklich R, Leavy M, Dreyer N, editors. Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition. Rockville: Agency for Healthcare Research and Quality (US); 2019. pp. 10–13.
2. Bai T, Egleston BL, Zhang S, et al. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. KDD. 2018;169(2):43–51. https://doi.org/10.1145/3219819.3219904.
3. Glover WJ, Li Z, Pachamanova D. The AI-Enhanced Future of Health Care Administrative Task Management. Catal Non-Issue content. 2022;3(2). https://doi.org/10.1056/CAT.21.0355.
4. Parimbelli E, Marini S, Sacchi L, Bellazzi R. Patient similarity for precision medicine: A systematic review. J Biomed Inform. 2018;83:87–96. https://doi.org/10.1016/j.jbi.2018.06.001.
5. Sharafoddini A, Dubin JA, Lee J. Patient Similarity in Prediction Models Based on Health Data: A Scoping Review. JMIR Med Inform. 2017;5(1):e7. https://doi.org/10.2196/medinform.6730.
6. Cebul RD, Love TE, Jain AK, Hebert CJ. Electronic Health Records and Quality of Diabetes Care. N Engl J Med. 2011;365(9):825–33. https://doi.org/10.1056/NEJMsa1102519.
7. McWilliams JM, Landon BE, Chernew ME, Zaslavsky AM. Changes in Patients' Experiences in Medicare Accountable Care Organizations. N Engl J Med. 2014;371(18):1715–24. https://doi.org/10.1056/NEJMsa1406552.
8. Holmes JH, Beinlich J, Boland MR, Bowles KH, Chen Y, Cook TS, et al. Why Is the Electronic Health Record So Challenging for Research and Clinical Care? Methods Inf Med. 2021;60(1–02):32–48.
9. Berisha V, Krantsevich C, Hahn P, Hahn S, Dasarathy G, Turaga P, et al. Digital medicine and the curse of dimensionality. npj Digit Med. 2021;4. https://doi.org/10.1038/s41746-021-00521-5.
10. Sauer C, Chen L, Hyland S, et al. Leveraging electronic health records for data science: common pitfalls and how to avoid them. Lancet Digit Health The. 2022;4(12). https://doi.org/10.1016/S2589-7500(22)00154-6.
11. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25(10):1419–28. https://doi.org/10.1093/jamia/ocy068.
12. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8.
13. Locke S, Bashall A, Al-Adely S, Moore J, Wilson A, Kitchen GB. Natural language processing in medicine: A review. Trends Anaesthesia Crit Care. 2021;38:4–9. https://doi.org/10.1016/j.tacc.2021.02.007.
14. Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32. Beijing: JMLR.org; 2014. p II-1188-II–1196.
15. Feng Y, Min X, Chen N, Chen H, Xie X, Wang H, et al. Patient outcome prediction via convolutional neural networks based on multi-granularity medical concept embedding. In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017. pp. 770–7. https://doi.org/10.1109/BIBM.2017.8217753.
16. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Making. 2017;17(2):67. https://doi.org/10.1186/s12911-017-0468-7.
17. Choi E, Bahadori MT, Searles E, Coffey C, Thompson M, Bost J, Tejedor-Sojo J, Sun J. Multi-layer Representation Learning for Medical Concepts. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: Association for Computing Machinery; 2016. p 1495–504.
18. Zhu Z, Yin C, Qian B, Cheng Y, Wei J, Wang F. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). 2016. pp. 749–58. https://doi.org/10.1109/ICDM.2016.0086.
19. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In: Doshi-Velez F, Fackler J, Kale D, Wallace B, Wiens J, editors. Proceedings of the 1st Machine Learning for Healthcare Conference. vol. 56 of Proceedings of Machine Learning Research. Northeastern University. Boston: PMLR; 2016. pp. 301–18. https://proceedings.mlr.press/v56/Choi16.html.
20. Great Ormond Street Hospital for Children - Who We Are. https://www.gosh.nhs.uk/about-us/who-we-are/. Accessed 30 Mar 2023.
21. Caroprese L, Veltri P, Vocaturo E, Zumpano E. Deep Learning Techniques for Electronic Health Record Analysis. 2018. pp. 1–4. https://doi.org/10.1109/IISA.2018.8633647.
22. Li Y, Rao S, Solares J, et al. BEHRT: Transformer for Electronic Health Records. Sci Rep. 2020;10(1). https://doi.org/10.1038/s41598-020-62922-y.
23. Rasmy L, Xiang Y, Xie Z, et al. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit Med. 2021;4(86). https://doi.org/10.1038/s41746-021-00455-y.
24. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019 Vol 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics; 2019. p 4171–86.
25. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. In: Bengio Y, LeCun Y (eds) 1st International Conference on Learning Representations, ICLR 2013. Scottsdale: Workshop Track Proceedings; 2013.
26. Davis J, Goadrich M. The Relationship between Precision-Recall and ROC Curves. In: Proceedings of the 23rd International Conference on Machine Learning. ICML '06. New York: Association for Computing Machinery; 2006. pp. 233–40. https://doi.org/10.1145/1143844.1143874.
27. Von Luxburg U. Clustering Stability: An Overview. Found Trends Mach Learn. 2010;2(3):235–74. https://doi.org/10.1561/2200000008.
28. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7.
29. Fukunaga K, Olsen DR. An Algorithm for Finding Intrinsic Dimensionality of Data. IEEE Trans Comput. 1971;C-20(2):176–183. https://doi.org/10.1109/T-C.1971.223208.
30. Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biol Direct. 2007;2(1):2. https://doi.org/10.1186/1745-6150-2-2.
31. Williams K, Thomson D, Seto I, Contopoulos-Ioannidis DG, Ioannidis JPA, Curtis S, et al. Standard 6: Age Groups for Pediatric Trials. Pediatrics. 2012;129(Supplement_3):S153–S160. https://doi.org/10.1542/peds.2012-0055I.
32. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
33. Jiffry M, Khan A, Pires F, Okam N, Vargas J, Moin K, et al. Agranulocytosis Secondary to Cancer Chemotherapy Associated With Higher In-Hospital Mortality in Patients With Central Line Insertion During a Hospital Stay. Cureus. 2023;15. https://doi.org/10.7759/cureus.34717.
34. Nyboe C, Olsen MS, Nielsen-Kudsk JE, Johnsen SP, Hjortdal VE. Risk of Pneumonia in Adults With Closed Versus Unclosed Atrial Septal Defect (from a Nationwide Cohort Study). Am J Cardiol. 2014;114(1):105–10. https://doi.org/10.1016/j.amjcard.2014.03.063.
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc.; 2017. p 6000–6010
36. Johnson AEW, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):160035. https://doi.org/10.1038/sdata.2016.35.

Publisher's Note