

Harnessing temporal patterns in administrative patient data to predict risk of emergency hospital admission



Benjamin Post, Roman Klapaukh, Stephen J Brett, A Aldo Faisal



Summary

Background Unplanned hospital admissions are associated with worse patient outcomes and cause strain on health systems worldwide. Primary care electronic health records (EHRs) have successfully been used to create prediction models for emergency hospitalisation, but these approaches require a broad range of diagnostic, physiological, and laboratory values. In this study, we aimed to capture temporal patterns of patient activity from EHR data and evaluate their effectiveness in predicting emergency hospital admissions compared with conventional methods.

Methods In this retrospective observational study, we used the Secure Anonymised Information Linkage databank to extract temporal patterns of primary care activity from undifferentiated electronic health record timestamp data for 1·37 million patients in Wales aged 18–80 years with at least one recorded Read code between the years 2016 and 2018. Using Gaussian mixture modelling we grouped patients into distinct temporal clusters, performed a three-stage validation of our approach and calculated the risk of emergency hospital admission for each temporal cluster group. Finally, these temporal clusters were combined with five administrative variables and incorporated into four emergency hospital admission prediction models (logistic regression, naive Bayes, XGBoost, and multilayer perceptron [MLP]) and compared with a more traditional, but data-intensive, modelling technique. The primary outcome was emergency hospital admission as the next health-care event.

Findings Six distinct temporal cluster patterns of primary care EHR activity were identified, associated with varying risks of future emergency hospital admission risk. These patterns were visually interpretable, repeatable at a population-level, and clinically plausible. The best emergency hospital admission prediction model (MLP) achieved an area under the receiver operating characteristic (AUROC) of 0·82 and precision of 0·94 in regional cohorts. In external validation in regional cohorts, similar model performance was observed (AUROC 0·82 and precision 0·92). This model also matched the performance of a more complex model (extended feature model) requiring 33 clinical parameters (AUROC 0·82 vs 0·83; precision 0·94 vs 0·90) for the same task on the same dataset.

Interpretation We developed a novel machine learning pipeline that extracts interpretable temporal patterns from simple representations of EHR data and can be incorporated into emergency hospital admission predictors. This framework might enable more rapid development of parsimonious clinical prediction models.

Funding UKRI CDT in AI for Healthcare, UKRI Turing AI Fellowship, NIHR Imperial Biomedical Research Centre, and Research Capability Funding.

Copyright © 2025 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY 4.0 license.

Introduction

Internationally, an urgent need to reduce hospital admissions has been identified,^{1,2} prompting interest in using electronic health records (EHRs) to help reduce this burden by identifying patients for intervention or care planning before hospital admission.³ Although EHRs are a well-established resource for clinical prediction,⁴ their use is expanding into new areas such as health-care planning, scheduling, and pandemic preparedness.^{5,6} Accurate and representative data are needed for these tasks, but improving data quality in EHRs can be expensive and resource-intensive.^{7,8} This conventional approach to EHR prediction modelling is fraught with challenges including noisy and complex data, algorithmic bias, mixed data types, data missingness, inconsistent quality, and interoperability

issues.^{4,9} As such, the use of prediction modelling has been limited to health systems with sophisticated data infrastructures.

For two decades, EHRs have been used to build statistical and machine learning predictors for emergency hospital admission.^{10,11,12} For example, the landmark QAdmissions study achieved an area under the receiver operating characteristic (AUROC) of 0·78 for prediction of emergency admission within 2 years.¹¹ However, the practical application of such models might be limited by their real-world performance (eg, high false positive rates). This phenomenon was demonstrated in the PRISMATIC trial published in 2018 which showed increased health-care costs without patient benefits when such prediction software was used in primary care practices.¹³ However, recent advancements suggest that

Lancet Digit Health 2025;
7: e124–35

See [Comment](#) page e106

Department of Bioengineering
(B Post MBChB,
Prof A A Faisal PhD) and
Department of Computing
(B Post, A A Faisal) and
Department of Chemical
Engineering (R Klapaukh PhD)
and Department of Surgery and
Cancer (Prof S J Brett MD),
Imperial College London,
London, UK; UKRI Centre in AI
for Healthcare, Imperial College
London, London, UK (B Post,
Prof S J Brett, A A Faisal), Chair in
Digital Health, Universität
Bayreuth, Bayreuth, Germany
(Prof A A Faisal)

Correspondence to:
Prof A Aldo Faisal, UKRI Centre in
AI for Healthcare, Imperial
College London,
London SW7 2AZ, UK
a.faisal@imperial.ac.uk

Research in context

Evidence before this study

We searched PubMed and Google Scholar from inception to July 31, 2024 for publications in the English language that aimed to predict emergency hospital admission from primary care electronic health record (EHR) data. We used the keywords “emergency hospital admission”, “prediction”, “risk”, “machine learning”, “primary care”, and “electronic health records”. Research in this domain, so far, has focused on curating accurate and detailed primary care medical data to maximise the number of model input variables. The landmark QAdmissions study from 2013 showed that emergency hospital admission could be predicted from primary care data. Larger and more sophisticated EHR datasets, combined with newer machine learning modelling techniques, have allowed more accurate predictors to be created. The current state-of-the-art approach used in one study improved upon the QAdmissions model’s predictive performance. Although these models achieve strong predictive performance they require a lot of clinical parameters to be measured and suffer from limitations, such as high false positive rates, which preclude them being used as primary care alert tools. In summary, existing methods use only a single timepoint of a large number of clinical variables to predict patient trajectory.

Added value of this study

In this study, we incorporate a dynamic representation of patients’ interaction with health care that allows us to map

their health activity over time with their risk of emergency hospital admission. We show that temporal patterns of primary health-care interactions can be decoded via our so-called general practice event clustering method and then provide a three-step validation of our novel approach to predict hospitalisation. We integrate this dynamic, temporal measure with five administrative variables, and show how we match or outperform a more traditional approach to emergency hospital admission prediction.

Implications of all the available evidence

We show here instead that simple data, likely to be present in most EHRs, can reveal interpretable and informative temporal patterns of primary care interaction that can replace more complex data. Our approach requires only a sequence of timestamps, easily extracted from most EHR databases, that we decode using machine learning. We use these temporal pattern signals with five other purely administrative variables (sex, age, ethnicity, Welsh Index of Multiple Deprivation, and total hospital admissions in the last 12 months) to train accurate and clinically useful predictors of emergency hospital admission. This work demonstrates a substantially lower data burden than a traditional approach (moving from >100 clinical parameters to six administrative parameters) for the prediction of emergency hospitalisation from primary care data. As we retain sufficiently high predictive performance, with a low false positive rate, this approach has potential utility for a range of health systems.

machine learning might outperform more traditional prediction models.¹⁴ A 2018 study from Rahimian and colleagues used standard machine learning techniques to enhance the QAdmissions model, achieving a benchmark predictive performance (AUROC 0.848)¹⁵ with an average precision-recall score of 0.35. This model relied on over 100 clinical parameters, limiting its use to health systems with sophisticated data infrastructures. Yet the simple presence of recorded medical data might be just as revealing as specific measured variables.¹⁶

These data, used by conventional EHR prediction models, are typically single-timepoint features, such as age or recent clinical measurements, which might not fully capture a patient’s dynamic health status or health trajectory.¹⁷ Changes in these trajectories might precede significant health-care events, such as unplanned hospital admissions, and can be associated with worse outcomes.^{18,19} Nonetheless, most published models use only simplified summary statistics of past health-care interactions to predict emergency hospital admissions.

In this paper, we harness the datetime labels of EHR administrative data, which are automated, low-cost, reliable, ubiquitous, and require minimal data preprocessing. We aimed to determine the usefulness of the datetime labels using a purpose developed machine

learning pipeline (figure 1A) to analyse patient trajectories as manifested in EHRs and read their temporal activity (figure 1B) by creating a novel, interpretable feature with a principled, multifaceted approach to validating this new feature, and show it can enrich the performance of emergency hospital admission prediction compared with a conventional approach.

Methods

Study design and data collection

In this retrospective observational study, we explored how patient trajectories can be extracted from primary care EHR data and be rendered informative to predicting future emergency hospital admission. We identified the ideal number of primary care interactions needed for this task and show that these sequences are not random. We then incorporated this optimal configuration into prediction models and compared this with more conventional approaches to the prediction of emergency hospital admission.

We used the Secure Anonymised Information Linkage (SAIL) databank for model development.^{20,21} SAIL is a national dataset from Wales (population: 3.1 million²²), consisting of billions of person-based health records and covering 86% of the population. This includes 43 linked and anonymised health data assets covering primary and

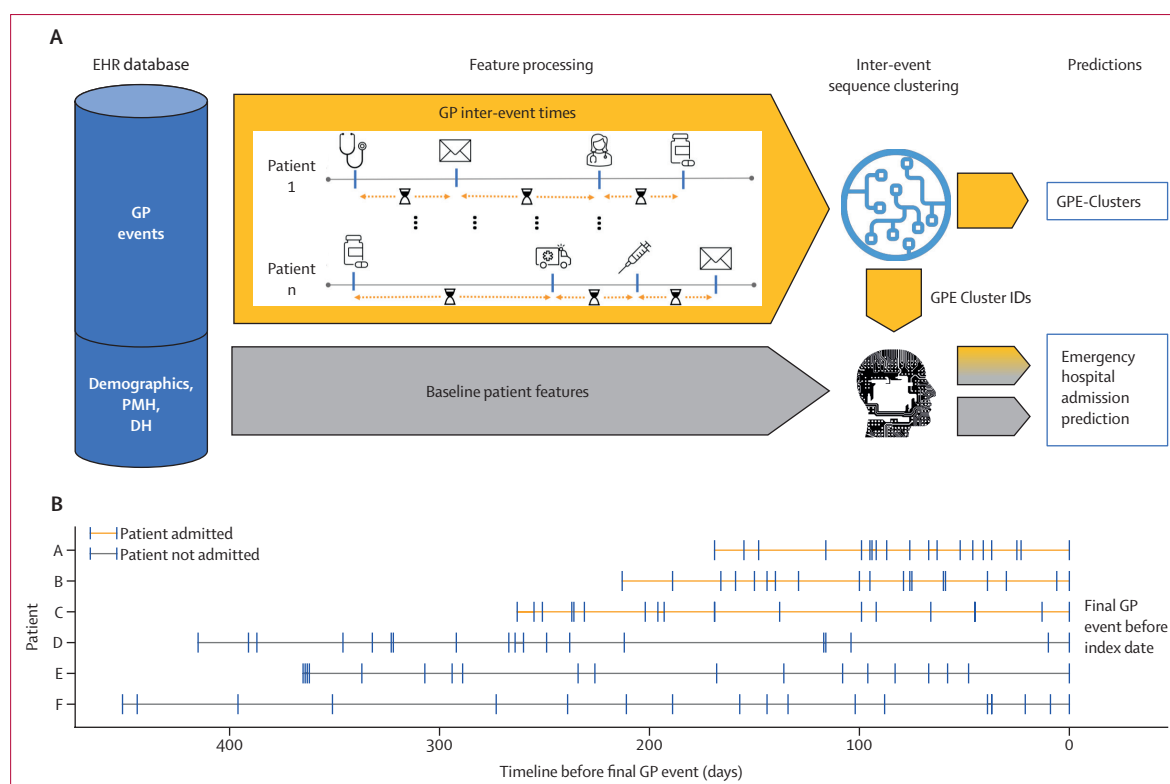


Figure 1: Research pipeline (A) and real-world patient data (B)

(A) GPE data are separated and the associated timestamps extracted to create inter-event sequence clusters using machine learning clustering. The outputs from this process are then merged with routinely collected patient data and used to train an emergency hospital admission prediction model. (B) Real GP inter-event times, before an index date, for a sample of anonymised patients. Blue markers indicate individual GPE. Random noise is added to inter-event times to hide true patient identities. This image is for demonstrative purposes only and shows the inter-event times for three patients (A–C) chosen at random from the admitted outcome group and three patients (D–F) chosen at random from the not admitted outcome group. EHR=electronic health record. GP=general practice. GPE=general practice event. PMH=past medical history. DH=drug history.

secondary health care. Of these available assets, four data resources were used: the Welsh Demographic Service Dataset, the Welsh Longitudinal General Practice Dataset, the Patient Episode Database for Wales, and the Annual District Death Extract (appendix p 2).

Approval for this project was granted by the SAIL Health Information Research Unit's Information Governance Review Panel (project approval number 1323). All data used in this study were anonymised and arise from the delivery of routine public services, so additional consent or ethical approval were not required as per Health Research Authority guidance.²³

Patients

All patients aged 18–80 years with at least one recorded Read code between the years 2016 and 2018 were included. All patients with five or more emergency hospital admissions in 2018 were excluded. All patients without a recorded pseudo-identifier, year of birth, sex identifier, or Welsh Index of Multiple Deprivation (WIMD)²⁴ were excluded as these formed key demographic data for our analysis. These exclusions were chosen as a minimum requirement for individual health

record completeness based on those chosen by Rahimian and colleagues.¹⁵

Patients who died during the study period were also excluded. This decision was made as we could not reliably identify patients with do not resuscitate orders, treatment escalation plans, and patients with expected deaths. These different subgroups of patients were deemed likely to have interacted with health care differently and risked artificially amplifying or adversely influencing our findings. However, these patients were reintroduced, to form a sensitivity analysis of our final prediction model (see appendix pp 3–4 for further details).

Outcomes

The outcome of interest was emergency hospital admission in the year 2018. Mental health and psychiatry admissions were excluded as these patients generally receive treatment that is partitioned from the main hospital infrastructure and have different treatment needs. Patients with five or more emergency admissions in 2018 (“frequent attenders” as defined by the Royal College of Emergency Medicine²⁵) were excluded, to

See Online for appendix

focus on identifying individuals with high but potentially undetected risk and avoid patients likely already well known to health services. However, all patients with one to four admissions were included and each repeat admission was treated as a distinct outcome point. This decision was made as each admission event, even if in relation to the same patient, is preceded by a unique pattern of primary care events. Our control group consisted of all patients that did not undergo an emergency hospital admission in 2018. Our method predicts emergency hospital admission as the next health-care event, so to assess our models on the control group we generated a random date within the outcome period (drawing on Jan 1, to Dec 31, of 2018 using a discrete uniform distribution). Thus, every patient in the study cohort was assigned an index date in 2018, even if an emergency admission did not occur. This approach was chosen over other methods, such as assigning the same index date to all non-admitted patients, to avoid confounding from seasonal variations in medical consumption and hospital admissions.

General practice events (GPEs)

To extract patterns of patient interaction with primary care services during the study period, we captured the unique dates on which each patient had at least one datapoint recorded in their individual primary care electronic health record: a general practice event (GPE). We defined a GPE as the date associated with any patient-specific clinical or administrative datapoint (symptoms, laboratory test results, medical appointments, telephone calls, etc). This approach was data-agnostic and treated all data similarly, regardless of the format or type of encoded data. For each study patient, all dates associated with any primary care EHR datapoint were extracted, but the data fields themselves were discarded. Duplicate date-values were removed, thus a list of dates was created demonstrating how patients temporally interacted with primary health-care services over the study period. The extraction of these data was purposefully straightforward as it involved the collection of raw timestamps only.

Baseline features

We extracted person-level data for demographics (patient age, gender, ethnicity, and WIMD decile), comorbidities, the total number of unique drug prescriptions in the year 2016, the total number of hospital admissions 1 year before the index date, and the total number of GPEs 2 years before the index date. These features were then divided into sparse and extended feature sets (appendix p 21). The sparse features included five simple variables (age, sex, ethnicity, WIMD, and number of hospital admissions in the last 1 year) normally available at an administrative level. The extended feature set was comprised of 26 comorbidities and clinical datapoints. These variables were selected to represent data used in more traditional

prediction models, informed by Rahimian and colleagues¹⁵ previous work.

All Welsh Longitudinal General Practice datapoints referring to hospital admissions and emergency department attendances were excluded, to avoid data leakage and duplication from secondary care data stored in the Patient Episode Database for Wales dataset.

Novel feature engineering: GPE intervals

We extracted GPEs for each unique patient and calculated the time difference (in days) between sequential GPEs (ie, inter-event times). Thus, each patient was mapped to a sequential array of integers representing the time difference between GPEs: we refer to this as their GPE-Sequence. In other words, for n GPEs, a total of $n-1$ inter-event times can be calculated, thus making the GPE-Sequence length = $n-1$. For example, a patient with GPEs on Jan 1, Jan 8, and Jan 10 would have a corresponding GPE-Sequence length of 2, containing the values 7 days and 2 days.

We used each study patient's index date to extract the preceding 24 months' GPE-Sequence. These GPE-Sequences were then clustered using Gaussian Mixture Models (GMMs) to create GPE-Clusters (figure 1A). A GMM is a probabilistic unsupervised machine learning technique that assumes all datapoints, in a given dataset, are generated from a combination of Gaussian distributions²⁶ which yields a principled approach to handle the natural variability in the data. This approach performs well with highly heterogeneous data and was chosen because the distribution of GPE inter-event times was heavily skewed (long-tail distribution; appendix p 8). However, GMMs require the dimensionality of all input features to be the same (each observation must possess the same number of variables). For this task, this meant that all GPE-Sequences had to be the same length. For the years 2016 to 2018, the study cohort had a variable number of GPEs (and thus GPE-Sequence lengths; appendix p 9). To facilitate comparison between GPE-Sequences with varying lengths, longer GPE-Sequences were reduced to the requisite length by removing datapoints representing the oldest GPEs and the most recent values were retained. This approach captured primary care activity over a fixed number of events rather than a fixed period of time. A visual demonstration of this procedure is provided in the appendix (p 10).

To determine the most effective GPE-Sequence length to predict emergency hospital admissions, we generated GPE-Clusters for lengths from 4 to 30 and incorporated these into corresponding logistic regression models (further details of this procedure are provided in the appendix [p 5]). This length was chosen as we felt that three or fewer inter-event times was not enough to be considered a pattern. The models were directly comparable since the GPE-Clusters were the only variable that differed between them. The optimal model was determined by balancing incremental prediction

performance against the number of patients that could be included in the model; longer GPE-Sequences resulted in a reduced number of patients meeting inclusion criteria, decreasing the patient sample size for model training (illustrated in appendix p 11).

GPE-intervals: validation

The approach outlined previously allowed us to capture patterns of primary care interaction from EHR data. To ensure our data were robust and not a chance phenomenon, we conducted a three-step validation experiment (details of which are described in the appendix pp 6–7).

Advanced prediction modelling

To demonstrate how our novel feature could be integrated into a real-world clinical predictor, we developed a two-stage pipeline that first generated GPE-Clusters using a GMM and then incorporated these into emergency hospital admission prediction models.

The optimal length GPE-Sequence was used to train a GMM and assign all patients in the cohort to a GPE-Cluster (GPE-Cluster 0, GPE-Cluster 1, etc). All patients with fewer than the required GPEs were assigned to a separate cluster label: GPE-Cluster-1; meaning patients who were not eligible for clustering could still be included in the final emergency hospital admission prediction model. Each patient's cluster membership was combined with their sparse baseline features and used to train four machine learning models—logistic regression, naive Bayes, extreme gradient boosting (XGBoost), and a multilayer perceptron (MLP)—to predict emergency hospital admission at the next health-care interaction. These models were each chosen as they belong to different categories of machine learning methods and offer different approaches to model estimation. XGBoost was chosen as the ensemble tree-based method over similar methods (eg, Random Forest, Gradient Boosting Classifier) as this approach has demonstrated superiority over deep neural networks for tabular data.²⁷ Additionally, EHR data are composed of mixed data types (binary, categorical, and numerical values), which can be handled by each of these methods. Hyperparameters for all models (eg, learning rate, tree-depth, and hidden-layer size) were tuned using a grid search (a complete summary of the search space can be found in appendix p 23). We performed internal validation using five-times cross-validation. We report the area under the receiver operating characteristic curve (AUROC), recall, precision, and Brier score. Model calibration curves were generated with 95% CIs. A sensitivity analysis was performed by rerunning the above models and including all patients who died within the study period (see appendix pp 3–4 for further details).

Advanced prediction modelling: external validation

External validation of the two-stage approach described previously was performed by emulating the methodology of Rahimian and colleagues,¹⁵ as recommended by Riley

and colleagues.²⁸ Specifically, the study population was divided into three non-random regional groups according to the 22 local authorities in Wales (see appendix p 22).

The best-performing GMM and machine learning classifier were identified, along with their hyperparameters. The first regional group was used as the training cohort and a GMM was trained to generate GPE-Clusters. The GPE-Clusters, for every patient in this region, were then combined with their sparse baseline features to create an emergency hospital admission prediction model. The optimal hyperparameters learned previously were used for model training. Further tuning was deemed unnecessary as these regional groups represented sub-sections of the whole study population.

This new two-stage pipeline was then tested on the second regional group (testing cohort 1) and the third regional group (testing cohort 2). Again, we report the AUROC, recall, precision, and Brier score. CIs were calculated by performing bootstrapping five times by sampling 90% of the population with replacement.

Comparative modelling including additional features

To demonstrate the predictive use of GPE-Clusters we trained separate XGBoost models, as above, using the sparse baseline features with and without the novel GPE-Clusters feature to predict emergency hospital admission as the next health-care interaction. An XGBoost classifier was chosen for this task, as model feature importances could easily be extracted. Hyperparameters were tuned with a grid search and five-times cross-validation was performed for each model. Performance metrics were then compared between the two best models.

Finally, we identified our best sparse feature model (GPE-Cluster + Sparse Baseline Feature model) and used the same class of machine learning predictor to train a more traditional predictive model (Extended Baseline Feature model) using the extended baseline feature set (appendix p 21) without GPE-Clusters. Internal five-times cross-validation was performed and the models performance were compared using AUROC, recall, precision, and Brier score. We then compared the performance of this conventional model with our best-performing sparse model.

Our findings are reported according to the TRIPOD²⁹ and STARD statements.³⁰ We performed all statistical analyses using Python version 3.9.12.

Role of the funding source

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

From the SAIL databank, a total of 2794146 patients were identified with at least one GPE in the study period. 71458 patients died during the study period and 1288858 patients did not meet age inclusion criteria.

	Emergency hospital admission			p value
	Overall (n=1 373 048)	Admitted (n=156 467)	Not admitted (n=1 216 581)	
Age, years	46.5 (17.2)	52.1 (18.7)	45.8 (16.9)	<0.001
Sex				
Female	726 068 (52.9%)	86 324 (55.2%)	639 744 (52.6%)	..
Male	646 980 (47.1%)	70 143 (44.8%)	576 837 (47.4%)	<0.001
WIMD decile	5.6 (2.9)	5.0 (2.8)	5.7 (2.9)	<0.001
Ethnicity				
African	2745 (0.2%)	205 (0.1%)	2540 (0.2%)	..
Arab	1164 (0.1%)	57 (0.0%)	1107 (0.1%)	..
Asian other	3718 (0.3%)	204 (0.1%)	3514 (0.3%)	..
Bangladeshi	1488 (0.1%)	168 (0.1%)	1320 (0.1%)	..
Black	1107 (0.1%)	90 (0.1%)	1017 (0.1%)	..
British or mixed British	181 586 (13.2%)	21 101 (13.5%)	160 485 (13.2%)	..
Caribbean	347 (0.0%)	30 (0.0%)	317 (0.0%)	..
Chinese	3055 (0.2%)	54 (0.0%)	3001 (0.2%)	..
Indian	2923 (0.2%)	178 (0.1%)	2745 (0.2%)	..
Mixed	4159 (0.3%)	312 (0.2%)	3847 (0.3%)	..
Not stated	993 655 (72.4%)	111 783 (71.4%)	881 872 (72.5%)	..
Other	43401 (3.2%)	3495 (2.2%)	39 906 (3.3%)	..
Pakistani	1842 (0.1%)	184 (0.1%)	1658 (0.1%)	..
White	131 858 (9.6%)	18 606 (11.9%)	113 252 (9.3%)	..
Comorbidities				
Hypertension	181 180 (13.2%)	35 918 (23.0%)	145 262 (11.9%)	<0.001
Diabetes	77 127 (5.6%)	20 598 (13.2%)	56 529 (4.6%)	<0.001
Atrial fibrillation	19 155 (1.4%)	8241 (5.3%)	10 914 (0.9%)	<0.001
Ischaemic heart disease	37 946 (2.8%)	15 629 (10.0%)	22 317 (1.8%)	<0.001
Venous thromboembolic disease	13 216 (1.0%)	4712 (3.0%)	8504 (0.7%)	<0.001
Chronic kidney disease	3304 (0.2%)	1363 (0.9%)	1941 (0.2%)	<0.001
Valvular heart disease	11 286 (0.8%)	3714 (2.4%)	7572 (0.6%)	<0.001
Congestive cardiac failure	8052 (0.6%)	4172 (2.7%)	3880 (0.3%)	<0.001
Cancer	155 506 (11.3%)	27 029 (17.3%)	128 477 (10.6%)	<0.001
Asthma or chronic obstructive pulmonary disease	259 298 (18.9%)	44 559 (28.5%)	214 739 (17.7%)	<0.001
Epilepsy	22 259 (1.6%)	5315 (3.4%)	16 944 (1.4%)	<0.001
Falls	1130 (0.1%)	607 (0.4%)	523 (0.0%)	<0.001
Depression	97 431 (7.1%)	19 607 (12.5%)	77 824 (6.4%)	<0.001
Hypercholesterolaemia	92 360 (6.7%)	19 058 (12.2%)	73 302 (6.0%)	<0.001
Manic depression or schizophrenia	49 809 (3.6%)	9974 (6.4%)	39 835 (3.3%)	<0.001
Chronic liver disease or pancreatitis	11 987 (0.9%)	3543 (2.3%)	8444 (0.7%)	<0.001
Rheumatoid arthritis or systemic lupus erythematosus	11251 (0.8%)	3151 (2.0%)	8100 (0.7%)	<0.001
Osteoarthritis	133 215 (9.7%)	28 351 (18.1%)	104 864 (8.6%)	<0.001
Connective tissue disease	3530 (0.3%)	998 (0.6%)	2532 (0.2%)	<0.001
Hemiplegia	2952 (0.2%)	1088 (0.7%)	1864 (0.2%)	<0.001
Learning difficulties	5889 (0.4%)	1114 (0.7%)	4775 (0.4%)	<0.001
Obesity	48 697 (3.5%)	11 071 (7.1%)	37 626 (3.1%)	<0.001
Osteoporosis	20 740 (1.5%)	5994 (3.8%)	14 746 (1.2%)	<0.001
Peripheral vascular disease	8825 (0.6%)	3451 (2.2%)	5374 (0.4%)	<0.001
Peptic ulcer disease	462 (0.0%)	141 (0.1%)	321 (0.0%)	<0.001
Substance misuse	6616 (0.5%)	1836 (1.2%)	4780 (0.4%)	<0.001
Total unique prescriptions (2016)	2.0 (0.0–6.0)	7.0 (2.0–13.0)	2.0 (0.0–5.0)	<0.001
Hospital visits in last 1 year	0.0 (0.0–0.0)	0.0 (0.0–1.0)	0.0 (0.0–0.0)	<0.001
Total GPEs in last 2 years	19.0 (6.0–41.0)	49.0 (24.0–78.0)	17.0 (6.0–37.0)	<0.001

Data are n (%), mean (SD), or median (IQR). WIMD=Welsh Index of Multiple Deprivation. GPE=general practice event.

Table: Baseline characteristics

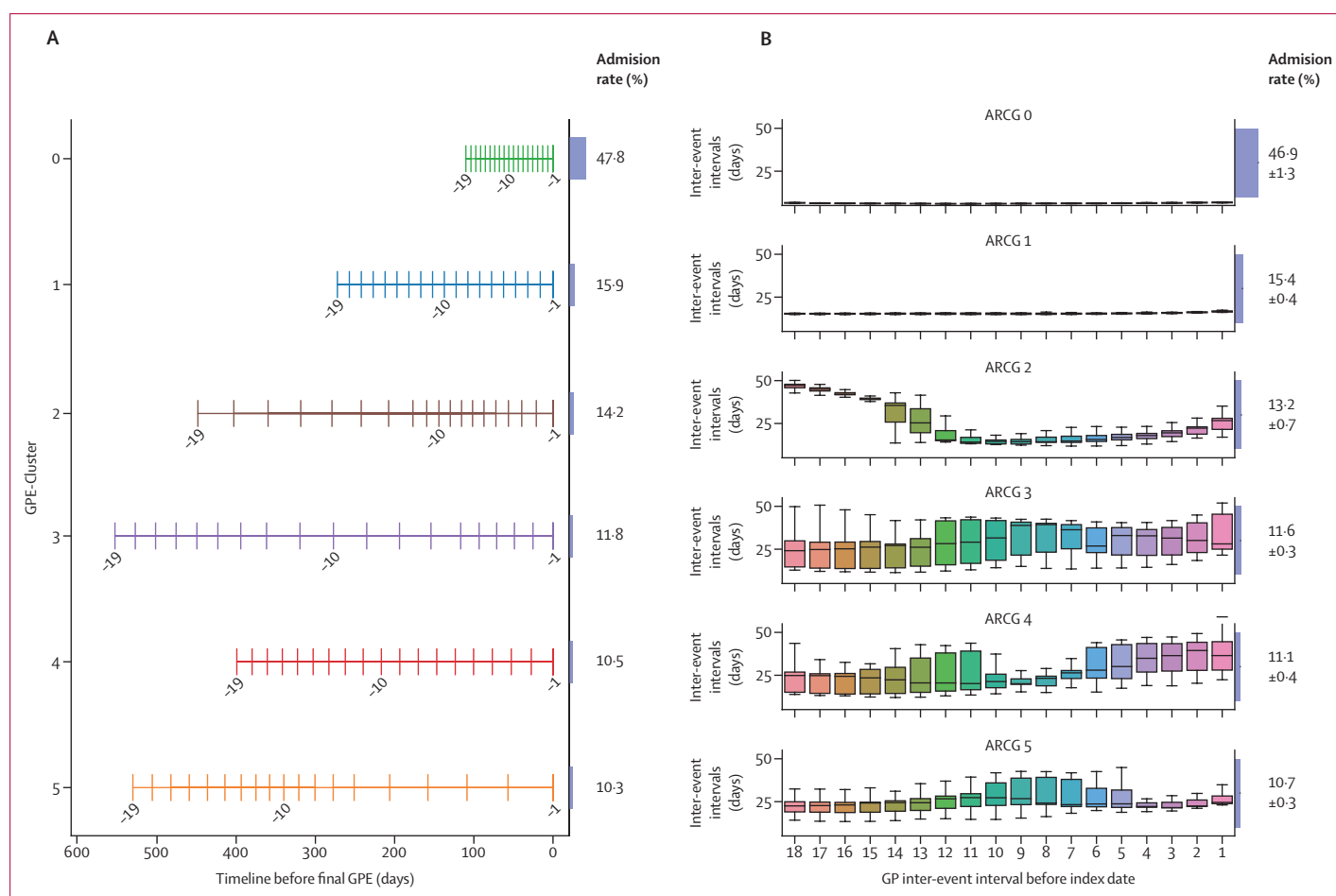


Figure 2: GPE-Sequence clusters patterns (A) and stability of GPE-Sequence clusters (B)

(A) The 6 GPE-Cluster and their centroids, based on 19 GPEs. Each line demonstrates cluster-specific GP event patterns and mean inter-event times leading up to an index date. The first (earliest) GPE is represented by the left-most tick on the horizontal line with the penultimate GPE, before the index date, represented by the right-most tick. The associated admission rate is shown to the right of each cluster (blue). (B) Summary of 1000 clustering trials based on different random initialisations—across multiple trials there is a consistent pattern seen in the three highest ARCGs. Each ARCG is the amalgamation of one given cluster over 1000 experimental trials. This is a summary plot that shows the spread of values over 1000 trials for this specific cluster. Each box represents the distribution of inter-event times (in days) after 1000 trials. The different colours in the box plots are to aid visualisation, they are not representative of any values. Clusters 0 and 1 demonstrate a frequent and regular attendance pattern, whereas Cluster 2 demonstrates a crescendo (increasing frequency of GPEs). The distributions around each cluster in ARCG 0–2 are tight, suggesting that these activity patterns are highly stereotyped and reliably present across the population. However, ARCGs 3–5 have wider distributions around most cluster centroids, suggesting that these patterns might be less stable across the population and account for a wider variety of primary care activity behaviour. GP=general practice. GPE=general practice event. ARCG=Admission Risk Cluster Group.

After excluding patients with incomplete demographic information and all patients with five or more emergency hospital admissions, we identified 1373 048 study patients (from 1336 010 unique patients after accounting for repeat admissions) with a total of 156 467 emergency hospital admissions in the year 2018 (population admission rate of 11.4%). The time-to-event distributions for both outcome groups are shown in the appendix (pp 13–14). Individuals admitted to hospital were older (52.1 years [SD 18.7] vs 45.8 years [16.9]) and more likely to be female (86 324 [55.2%] of 156 467) than individuals not admitted to hospital (639 744 [52.6%] of 1216 581). 72.4% of the study cohort did not have an ethnicity recorded, but where reported, ethnicity was approximately equally distributed amongst both outcome groups (table). The total number of unique prescriptions was higher in

individuals admitted to hospital than not admitted (7 [IQR 2–13] vs 2 [0–5]). The total number of GPEs, in the 2 years before the index date, was also higher for patients admitted to hospital (49 [IQR 24–78] vs 17 [6–37]).

Logistic regression predictor models were created for GPE-Sequence lengths from 4 to 30 (4–30 inter-event times before the index date). A step-change in model performance occurred at GPE-Sequence lengths greater than 15 for both emergency hospital admission prediction and for the high-risk cluster admission rate (appendix pp 11, 15). An optimal GPE-Sequence length was determined to be 18 as model performance plateaued for both experiments at this point. As such, we selected an optimal GPE-Sequence length of 18 as model improvements beyond this point were outweighed by the loss of patients eligible for clustering. We identified

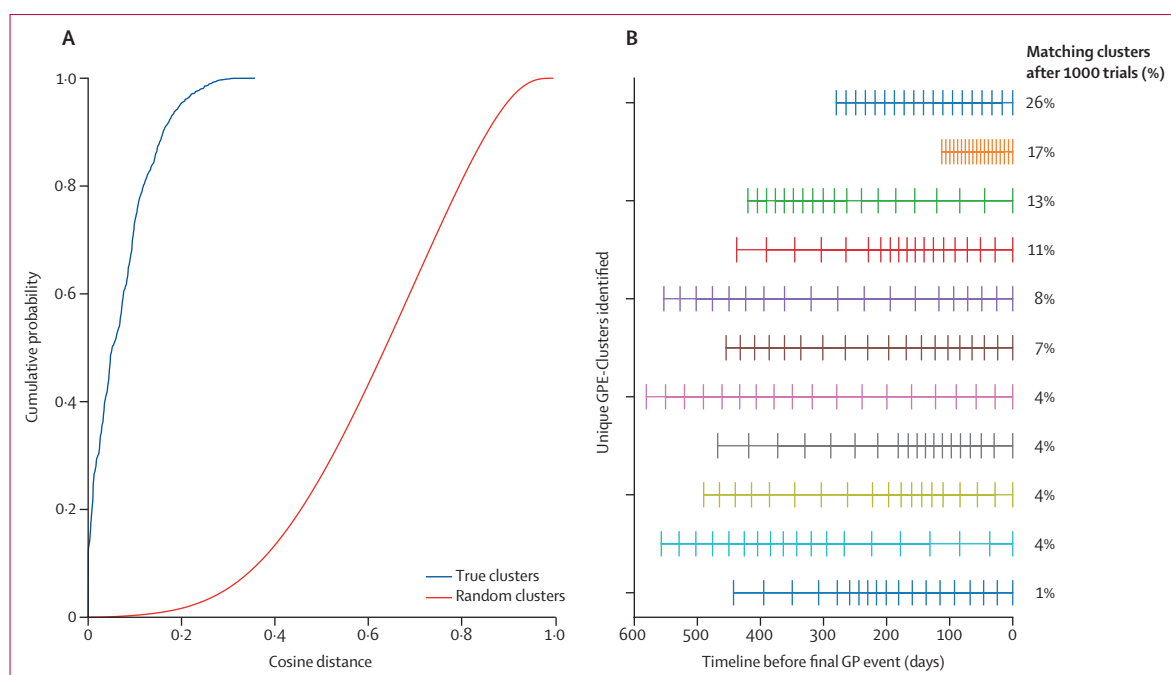


Figure 3: Cumulative distribution function of cosine distances for true and random clusters (A) and dictionary of clusters (B)

(A) For 1000 simulated trials (6000 clusters), the distributions of cosine distances between all clusters are markedly different for the true clusters (GPE-Clusters) and randomly generated clusters. (B) Demonstration of the 11 most commonly occurring unique clusters identified after 1000 clustering trials (6000 clusters) based on the cosine distance threshold of similarity. After each clustering trial, each new set of six clusters was compared with previous unique clusters by calculating the cosine distance. If the cosine distance between any new cluster was greater than the threshold of similarity (appendix p 12) between all other clusters, then this cluster was considered a new unique cluster and added to the dictionary of unique clusters. If the calculated cosine distance was below the threshold of similarity, this was considered a match with a previous cluster, and the new cluster was discarded. Each unique cluster, with at least one match, is shown here and ordered by the frequency of matching (ie, the most commonly occurring cluster patterns). GPE=general practice event.

six clusters as the optimal number of components (appendix p 16), approximating six distinct phenotypes of primary care EHR activity before an index date. The choice of six clusters was reinforced by the findings of our validation experiments.

For all lengths of GPE-Sequence explored, a specific pattern of interest emerged: regular and frequent primary care consumption was associated with a higher hospital admission rate than all other clusters. The results of clustering based on a GPE-Sequence of length 18 is shown in figure 2A, with the highest risk cluster (GPE-Cluster 0) associated with a GPE approximately every week and an emergency hospital admission rate of 47.8%.

We used the optimal GPE-Sequence length (18) and optimal number of clusters (six), as identified above, to perform 1000 clustering trials with different random initialisations. When the clusters were matched across trials by admission rate rank, this resulted in six admission risk cluster groups, each composed of 1000 18-dimensional clusters. Within each admission risk cluster groups, the median and IQR of the 18 sequential centroids are demonstrated in figure 2B. The admission risk cluster groups associated with the highest admission rates (admission risk cluster groups 0, 1, and 2) showed clear stability across all trials

evidenced by tight distributions across each dimension. Specifically, admission risk cluster groups 0 and 1 showed frequent and repeating patterns where each centroid (indicative of mean inter-event times) had a very similar value (6.2 [IQR 0.29] and 15.5 [0.60], respectively) with admission rates of 46.9 (95% CI 1.3) and 15.4 (0.4). Admission risk cluster group 2 showed a clear step-like decrease in centroid values (inter-event time), before the index date, with an admission rate of 13.2 (95% CI 0.7).

The distribution of cosine distances between all clusters from the 1000 trials was compared with the simulated pseudo-clusters and found to be visually different (figure 3A). The distribution of the cosine distances between the true clusters (appendix p 12) was then used to identify a value of 0.02 as the threshold of similarity. From a total of 6000 clusters, created after 1000 clustering trials (six clusters per trial), a total of 24 unique clusters (cosine distances between each cluster and all other dictionary clusters >0.02) were identified with 11 having at least one match across all trials. Unique cluster patterns, with at least one match across 1000 trials, are shown in figure 3B. The top three most common clusters created accounted for more than 50% of all patterns generated, representing stability of these cluster patterns across all 1000 clustering trials. The most

common unique clustering patterns identified in this dictionary closely resemble the final six clusters shown in figure 2A.

Using the five sparse baseline features in conjunction with the length 18 GPE-Clusters, the logistic regression, XGBoost, and MLP models achieved AUROC performances of 0·813 (95% CI 0·813–0·813), 0·821 (0·820–0·821), and 0·821 (0·820–0·821), respectively, for the prediction of emergency hospital admission as the next health-care interaction. The naive Bayes classifier had significantly worse performance: AUROC 0·753 (95% CI 0·753–0·754). The MLP demonstrated significantly better precision when compared with the other models: logistic regression 0·402 (95% CI 0·388–0·415), naive Bayes 0·691 (0·691–0·692), XGBoost 0·450 (0·421–0·478), and MLP 0·935 (0·935–0·935). The MLP demonstrated the best calibration (appendix p 17) and the lowest Brier scores and negative log loss (figure 4). A full account of model performance metrics and tuned hyperparameters are located in the appendix (p 23). Model performance for the sensitivity analysis, including all patients who died within the study period, are reported in the appendix (pp 3–4). These results show comparable performance metrics to our baseline analysis, demonstrating robustness of our approach to a different composite outcome.

The most effective model for predicting emergency hospital admissions using five sparse baseline features and 18 GPE-Clusters was the MLP classifier. A similar two-stage model, trained exclusively on patients from seven Welsh regions, achieved similar performance to the model trained on the entire study population. Model performance on testing cohorts 1 and 2 (appendix p 18) achieved AUROCs of 0·810 (95% CI 0·809–0·811) and 0·822 (0·821–0·822), respectively. Precision (0·899 [0·898–0·899] and 0·916 [0·915–0·916]) and recall (0·404 [0·404–0·405] and 0·429 [0·428–0·430]) metrics were also congruous with the predictive performance. A full account of model performance metrics for these test cohorts is provided in the appendix (p 24).

The addition of GPE-Clusters to the sparse baseline features significantly improved the performance of our XGBoost model to predict emergency hospital admission. The median AUROC for the standard model was 0·776 (95% CI 0·726–0·826) and improved to 0·821 (0·789–0·853) with the addition of the novel cluster feature. The feature importances are shown in the appendix (p 19).

Our best prediction model trained on the sparse baseline feature set combined with the GPE-Cluster was identified to be the MLP. As such, we trained a second MLP model using the extended baseline feature set but excluding the GPE-Cluster feature. Results were comparable between the models: AUROC 0·821 (95% CI 0·820–0·821) for the GPE-Cluster and sparse baseline feature model versus 0·834 (0·830–0·839) for the extended baseline feature model, recall 0·418

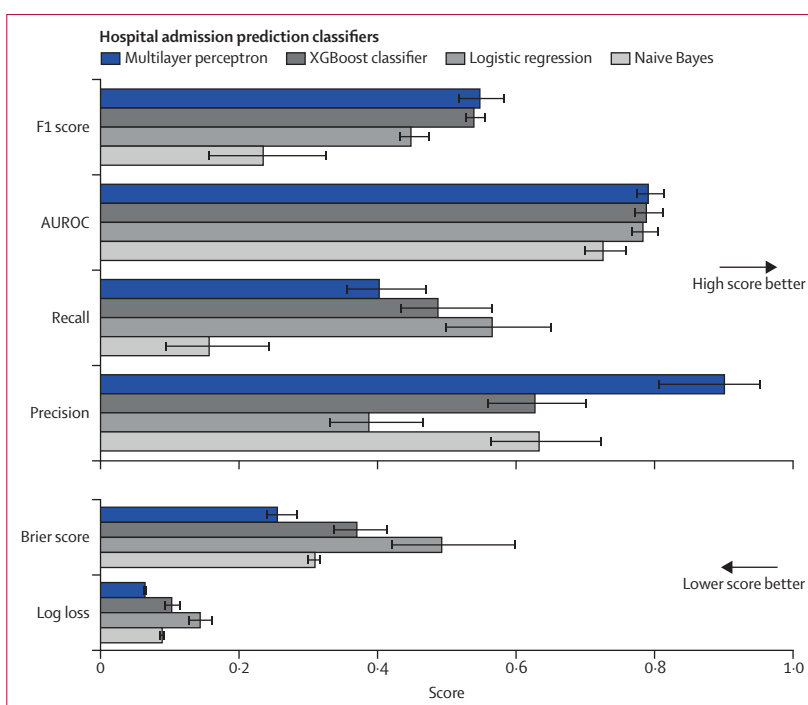


Figure 4: Model performance for the full pipeline

Model performance for four machine learning classifiers trained with five baseline administrative features in combination with the GPE-Sequence clusters, to predict emergency hospital admission. The Brier score and Log loss score are probabilistic prediction evaluation metrics, with the multilayer perceptron showing the best (lowest) scores. AUROC=area under the receiver operating characteristic. GPE=general practice event.

(0·418–0·418) versus 0·432 (0·432–0·432), precision 0·935 (0·935–0·935) versus 0·902 (0·890–0·915; figure 5). To demonstrate that the MLP was the fairest comparator model here we also trained XGB models (see appendix p 20). This result suggests that our novel GPE-Cluster feature has the same predictive value as an extended list of medical diagnoses, total drug prescriptions, and total primary care visits in 24 months.

Discussion

We have shown a novel approach to EHR research by discovering and interpreting a temporal code in simple timestamp data and transforming this code into a novel temporal feature, the GPE-Cluster, which can be incorporated into patient outcome predictors. Crucially, this approach matched a conventional predictive approach that relied on extensive clinical data. Specifically, our GPE-Clusters, based on simplified representations of routinely collected data, reveal interpretable and meaningful primary care interaction patterns. A three-stage validation of these GPE-Clusters confirmed their mathematical stability and demonstrated consistent interaction patterns associated with high hospital admission risk. Finally, these meaningful GPE-Clusters can enrich emergency hospital admission prediction models.

Previous studies have shown that health-care behaviours change before emergency hospitalisation.^{15,18}

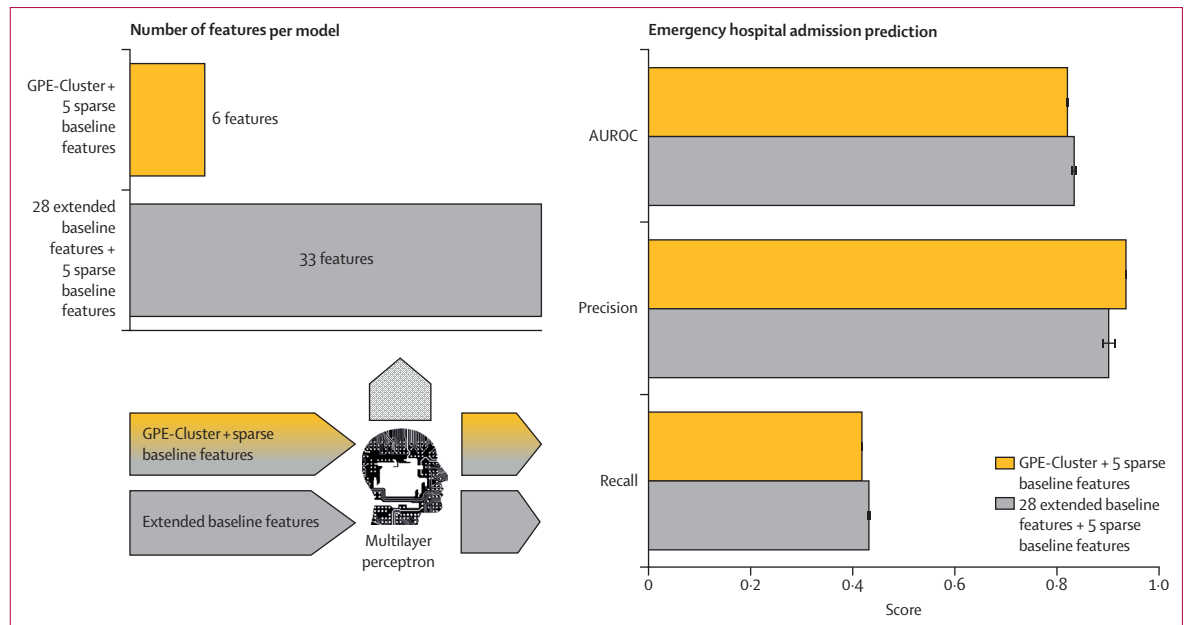


Figure 5: Feature exchangeability

Emergency hospital admission prediction models achieve similar performance when 28 commonly used features are replaced with our single novel GPE-Cluster feature. GPE=general practice event.

Our findings expand on this, by demonstrating that complex temporal patterns of health-care use might be more informative than simplistic summary metrics. Our data-driven approach extracts a comprehensible, temporal representation of patient trajectories, meaning that clinical predictions are accompanied by inbuilt, bespoke interpretability. Similar published works^{11,15} that adopt a more traditional approach to prediction modelling tend to be more data-intensive, requiring sophisticated and effortful approaches to high-quality feature extraction. In addition, the explainability of these models is limited to methods such as feature importance, which might provide a poor understanding of the inner workings of a model.³¹ However, our prediction models are not directly comparable with these previously published works, due to subtle differences in our defined outcomes. Our novel GPE-Cluster feature is retrospectively constructed from patient outcome dates, as we were focused on identifying imminent emergency hospital admissions. Instead, we trained an internal comparator predictor to represent a more established modelling approach, partially based on these previously published models. Compared with this conventional approach, our novel feature prediction model used 18% as many features of the data (6 vs 33) to achieve similar prediction performance.

Our GPE-Clusters are derived from readily available EHR data and require straightforward preprocessing, but are able to replace more expensive features which require medical expertise to parse (eg, past medical history and prescriptions). Nonetheless, this novel feature was able to extract interesting and interpretable patterns of primary

care interaction. For example, GPE-Cluster 0 could encompass a patient with poorly controlled diabetes needing regular treatment for chronic complications like an infected ulcer; patients in this cluster are associated with an emergency hospital admission rate of approximately 47.8%. The next temporal pattern, GPE-Cluster 1, was similar to GPE-Cluster 0, but had roughly twice the periodicity with bi-weekly events. In this cluster, for example, we might find a patient with multiple chronic diseases that are well managed due to frequent preventive actions by health services (corresponding to a lower admissions rate of 15.9% for patients falling into this cluster). The third, GPE-Cluster 2, might represent individuals with a sudden deterioration in health, such as an acute infection, as it shows a pattern that increases in its rhythm, from monthly to bi-weekly health-care interactions (14.2% are admitted). By contrast, GPE-Cluster 5 showed a reversal of this pattern going from bi-weekly to monthly intervals, for example, a patient showing improvement after apparent resolution of complications due to chronic illness (10.3% of whom are admitted). Over time, patients might transit between different GPE-Clusters (eg, from deterioration in GPE-Cluster 2 to GPE-Cluster 1 to GPE-Cluster 0), showing potential fluctuation of health activity over time.

To demonstrate the potential for real-world applicability, the novelty of the GPE-Clusters needed to be paired with robust validity and clinically useful prediction models. Our methodology captured 60% of the approximately 2.2 million people aged 18–80 years old in Wales (with at least one primary care Read code recorded). This sample

was demographically representative (age and sex) of the official population statistics,²² making our findings applicable to the country of Wales. The stability of our GPE-Clusters was cross-validated across this large group and confirmed through thousands of randomised initialisations, giving us high confidence in the precise nature of these clusters. Finally, our MLP model matched AUROC scores of logistic regression and XGBoost models, but with superior precision (positive predictive value), minimising false positives at the expense of true positives. As a population-level screening tool, this is a justifiable trade-off as these patients would not otherwise have been identified and most health services would not have the resources to intervene pre-emptively in this cohort.

Our study should be interpreted considering its limitations. First, although our study population was large, it was confined to a single country with a unified health system, free at the point of care, with unrestricted access to primary care which controls access to secondary and specialist care. More comprehensive testing (eg, validation on an external dataset from a different country) is warranted before real-world clinical deployment. Second, over 70% of our study population did not have a recorded ethnicity, which might have masked ethnicity-specific GPE patterns. Different ethnic groups might interact differently with health-care services.³² Third, to construct our GPE-Clusters, we did not explore GPE-Sequence lengths less than 4, as we felt that three inter-event times did not constitute a pattern of primary care interaction and were reassured by model performance improvements above this minimum threshold. Finally, we have trained models using four established machine learning approaches. Although it is possible that time-to-event models (to account for right censoring) or state-of-the-art classifiers might have outperformed these approaches, the focus of our work here was to introduce the novel GPE-Cluster approach to EHR data.

One of the key advantages of our study is that we chose established predictor variables, readily available in most modern EHRs, allowing our approach to easily be replicated using most primary care data repositories. Second, our novel GPE-Cluster features were generated using immediately intuitive clustering approaches, making them tangible and physically measurable quantities. This built-in interpretability facilitates visualisation of specific patterns of primary care activity and clinical understanding of stereotypical patterns of health interaction before emergency hospital admission. Simple escalating or de-escalating temporal patterns of behaviour have been shown to be predictive of human nature, performance, or intent in non-clinical domains, such as voting in the US Congress,³³ sick leave patterns in employee performance,³⁴ and social media communication.³⁵ Although our GPE-Clusters contain such escalating and de-escalating temporal patterns of health-care consumption, we also found more nuanced temporal patterns. Finally, the data required to create GPE-Clusters

are derived from complex clinical data, but our feature requires only a simplified representation of these data. Our approach is agnostic to specific data types, requiring only timestamps, meaning all data are treated equally, thus reducing preprocessing requirements and the need for domain expertise for variable interpretation. This framework for analysis is easily generalisable to other health-care outcomes, presenting opportunities for application and validation to other predictive tasks.

In conclusion, we have shown, using an exemplar application of emergency hospital admission prediction, a general principle that EHR administrative data contain valuable temporal predictive information that can be extracted using purpose-built artificial intelligence models. We provide a principled approach to evaluating these GPE-Clusters and demonstrate how they can facilitate clinical predictions using straightforward EHR data. In addition to standard clinical data, EHRs possess temporal patterns of activity that resemble a kind of health Morse code, which can be extracted at a national level, opening new avenues for EHR research.

Contributors

BP prepared the original manuscript draft. All authors had access to all data in the study. BP, RK and AAF accessed and verified the data. Formal analysis and software development was done by BP supported by RK and AAF. The study was conceptualised by BP, SJB, and AAF. The method was developed by BP, RK, SJB, and AAF and software was developed and maintained by BP. All authors wrote the manuscript and were responsible for the final decision to submit the manuscript for publication.

Declaration of interests

We declare no competing interests.

Data sharing

The data used in this study are available in the SAIL databank at Swansea University, Swansea, UK. All proposals to use SAIL data are subject to review by an independent Information Governance Review Panel (IGRP). Before any data can be accessed, approval must be given by the IGRP. The IGRP gives careful consideration to each project to ensure proper and appropriate use of SAIL data. When access has been granted, it is gained through a privacy-protecting safe haven and remote access system referred to as the SAIL Gateway. SAIL has established an application process to be followed by anyone who would like to access data via SAIL (<https://saildatabank.com/data/apply-to-work-with-the-data/>). Our research code can be found here: <https://doi.org/10.6084/m9.figshare.21696281.v1>.

Acknowledgments

BP was supported by the UKRI CDT in AI for Healthcare (<http://ai4health.io>; grant no. EP/S023283/1). AAF holds a UKRI Turing AI Fellowship (grant no. EP/V025449/1). This research was also funded by the NIHR Imperial Biomedical Research Centre (BRC), and by Research Capability Funding. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

References

- 1 Steventon A, Deeny S, Friebe R, Gardner T, Thorlby R. Briefing: Emergency hospital admissions in England—which may be avoidable and how? The Health Foundation, 2018.
- 2 van den Broek S, Heiweggen N, Verhofstad M, et al. Preventable emergency admissions of older adults: an observational mixed-method study of rates, associative factors and underlying causes in two Dutch hospitals. *BMJ Open* 2020; **10**: e040431.
- 3 Ben-Assuli O, Shabtai I, Leshno M. Using electronic health record systems to optimize admission decisions: the Creatinine case study. *Health Informatics J* 2015; **21**: 73–88.

- 4 Honeyford K, Expert P, Mendelsohn EE, et al. Challenges and recommendations for high quality research using electronic health records. *Front Digit Health* 2022; 4: 940330.
- 5 Harris SL, May JH, Vargas LG. Predictive analytics model for healthcare planning and scheduling. *Eur J Oper Res* 2016; 253: 121–31.
- 6 Mizani MA, Dashtban A, Pasea L, et al. Using national electronic health records for pandemic preparedness: validation of a parsimonious model for predicting excess deaths among those with COVID-19—a data-driven retrospective cohort study. *J R Soc Med* 2023; 116: 10–20.
- 7 Feder SL. Data Quality in electronic health records research: quality domains and assessment methods. *West J Nurs Res* 2018; 40: 753–66.
- 8 Reimer AP, Milinovich A, Madigan EA. Data quality assessment framework to assess electronic medical record data for use in research. *Int J Med Inform* 2016; 90: 40–47.
- 9 Sauer CM, Chen L-C, Hyland SL, Girbes A, Elbers P, Celi LA. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health* 2022; 4: e893–98.
- 10 Bottle A, Aylin P, Majeed A. Identifying patients at high risk of emergency hospital admissions: a logistic regression analysis. *J R Soc Med* 2006; 99: 406–14.
- 11 Hippisley-Cox J, Coupland C. Predicting risk of emergency admission to hospital using primary care data: derivation and validation of QAdmissions score. *BMJ Open* 2013; 3: e003482.
- 12 Wallace E, Smith SM, Fahey T, Roland M. Reducing emergency admissions through community based interventions. *BMJ* 2016; 352: h6817.
- 13 Snooks H, Bailey-Jones K, Burge-Jones D, et al. Predictive risk stratification model: a randomised stepped-wedge trial in primary care (PRISMATIC). *Health Serv Deliv Res* 2018; 6.
- 14 Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA Netw Open* 2020; 3: e1918962.
- 15 Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: development and validation using linked electronic health records. *PLoS Med* 2018; 15: e1002695.
- 16 Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018; 361: k1479.
- 17 Luta X, Diernberger K, Bowden J, et al. Healthcare trajectories and costs in the last year of life: a retrospective primary care and hospital analysis. *BMJ Support Palliat Care* 2024; 14: e807–15.
- 18 Lone NI, Gillies MA, Haddow C, et al. Five-year mortality and hospital costs associated with surviving intensive care. *Am J Respir Crit Care Med* 2016; 194: 198–208.
- 19 Gross JL, Borkowski J, Brett SJ. Patient or family perceived deterioration in functional status and outcome after intensive care admission: a retrospective cohort analysis of routinely collected data. *BMJ Open* 2020; 10: e039416.
- 20 Ford DV, Jones KH, Verplancke J-P, et al. The SAIL Databank: building a national architecture for e-health research and evaluation. *BMC Health Serv Res* 2009; 9: 157.
- 21 Lyons RA, Jones KH, John G, et al. The SAIL databank: linking multiple health and social care datasets. *BMC Med Inform Decis Mak* 2009; 9: 3.
- 22 Office for National Statistics. Population and household estimates: Census 2021, unrounded data. Office for National Statistics. June, 2022.
- 23 Jones KH, Ford DV, Thompson S, Lyons RA. A profile of the SAIL Databank on the UK Secure Research Platform. *Int J Popul Data Sci* 2019; 4: 1134.
- 24 StatsWales. Welsh Index of Multiple Deprivation. 2019. <https://statswales.gov.wales/Catalogue/Community-Safety- and-Social-Inclusion/Welsh-Index-of-Multiple-Deprivation> (accessed Aug 1, 2022).
- 25 Hayhurst C, Smith S, Chambers D. Frequent attenders in the emergency department. The Royal College of Emergency Medicine, 2017.
- 26 Reynolds D. Gaussian mixture models. In: Li SZ, Jain A, eds. *Encyclopedia of biometrics*. Springer US, 2009: 659–63.
- 27 Schwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. *arXiv* 2021; published online Nov 23. <https://doi.org/10.48550/arXiv.2106.03253> (preprint).
- 28 Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 2016; 353: i3140.
- 29 Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann Intern Med* 2015; 162: 55–63.
- 30 Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015; 351: h5527.
- 31 Dunn J, Mingardi L, Zhuo YD. Comparing interpretability and explainability for feature selection. *arXiv* 2021; published online May 11. <https://doi.org/10.48550/arXiv.2105.05328> (preprint).
- 32 Wan YI, Robbins AJ, Apea VJ, et al. Ethnicity and acute hospital admissions: multi-center analysis of routine hospital data. *EClinicalMedicine* 2021; 39: 101077.
- 33 Critchfield TS, Haley R, Sabo B, Colbert J, Macropoulos G. A half century of scalloping in the work habits of the United States Congress. *J Appl Behav Anal* 2003; 36: 465–86.
- 34 Kauermann G, Ortlieb R. Temporal pattern in number of staff on sick leave: the effect of downsizing. *J R Stat Soc Ser C Appl Stat* 2004; 53: 355–67.
- 35 Tavares G, Faisal A. Scaling-laws of human broadcast communication enable distinction between human, corporate and robot Twitter users. *PLoS One* 2013; 8: e65774.