Predicting Acute Clinical Deterioration with Interpretable Machine Learning to support Emergency Care Decision Making

Stelios Boulitsakis Logothetis[1], **Darren Green**[2,3], **Mark Holland**[4], **and *Noura Al Moubayed**[1,5]

[1]Department of Computer Science, University of Durham, Durham, United Kingdom
[2]Northern Care Alliance NHS Foundation Trust, Department of Renal Medicine, Manchester, United Kingdom
[3]Division of Cardiovascular Sciences, University of Manchester, Manchester, United Kingdom
[4]School of Clinical and Biomedical Sciences, University of Bolton, United Kingdom
[5]Evergreen Life Ltd, Manchester, United Kingdom

ABSTRACT

The emergency department is a fast-paced environment responsible for large volumes of patients with varied acuity. Operational pressures on emergency departments are increasing, which creates the imperative to efficiently identify patients at risk of acute deterioration and minimise patients' risk of morbidity and mortality. We apply state-of-the-art machine learning methods to predict patient deterioration early, based on their first recorded vital signs, observations, laboratory results, and other predictors documented in Electronic Patient Records. We build on prior work by incorporating interpretable machine learning and fairness-aware modelling, and achieve improved classification performance over the current standard, the National Early Warning Score 2, measured by average precision and daily alert rate. We use a cross-sectional Electronic Patient Record dataset comprising $121,058$ unplanned admissions at Salford Royal Hospital, UK, to systematically compare model variations for predicting mortality and critical care utilisation within 24 hours of admission. Our models yield up to 0.375 increase in average precision, up to 18.5% reduction in daily alert rate, and a median 0.577 reduction in differential bias amplification across the protected demographics of age and sex. We use Shapely Additive exPlanations to justify the models' outputs, verify that the captured data associations align with domain knowledge, and pair predictions with the causal context of each patient's most influential characteristics. We encourage future research to follow a systematised approach to data-driven risk modelling and help obtain clinically applicable support tools.

Introduction

When patients deteriorate, care providers must be able to recognise their worsening condition immediately and intervene accordingly[?]. Delayed identification of deterioration is associated with preventable hospital deaths[?], while delaying the transfer of critically ill patients to intensive care puts them at higher risk of morbidity and mortality[?]. The importance of timely identification and appropriate response to clinical instability has motivated the development of 'track-and-trigger' systems. These systems tie clinical observations that are antecedent to patient deterioration with recommended interventions to be executed by care staff or dedicated response teams as part of a rapid response system[?]. In the United Kingdom, this system is recommended by both National Institute for Health and Care Excellence (NICE) and the Royal College of Physicians (RCP) to monitor all adult patients in acute hospital settings[?,?].

In most cases, acute clinical instability and deterioration are preceded by abnormal vital signs[?], therefore standard practice in acute secondary care settings is to monitor patients using basic homeostatic measures, which include heart rate, blood pressure, inspired oxygen, oxygen saturation, temperature, and level of consciousness[?]. To assist this process, weighted aggregate scores of these measures, known as Early Warning Scores (EWS), have been developed to characterise the patient's acuity[?]. These scores can act as the afferent component of a rapid response system, tying them to an escalation protocol or a set of recommended clinical interventions[?].

Historically, data pertaining to an EWS were manually recorded and tallied on paper charts. As such, they often fell short of including the full breadth and variety of available predictive information[?]. The gradual phasing-out of bedside paper charts has brought the transition to digital EWS solutions that draw patient data in real-time from Electronic Patient Records (EPR). Beyond digitising conventional EWS, EPR systems collate comprehensive patient data, which can be used to improve performance and clinical utility[?]. In particular, the large volume of available data makes it feasible to develop a purely or partly data-driven solution using machine learning. AI-based systems have already demonstrated suitability for assisting in medical

imaging tasks, which makes AI-powered prognostic modelling a key research area of interest[?]. Our study concentrates on analysing EPR data to model clinical risk, as we use machine learning methods to potentially identify acute clinical deterioration in patients presenting to the emergency department (ED).

Prior work has used machine learning to model inpatient admission, deterioration, critical care admission, cardiac arrest, and mortality, among other outcomes[?]. In a systematic review of studies published from 2009-2017, Goldstein et al. identified 107 applications of EPR data to training statistical and ML models[?]. Recently, Klug et al.[?] used gradient-boosted decision trees (GBDT) on a single-centre cohort of approximately 800,000 ED episodes to predict short-term mortality risk and achieved improved performance over severity scores such as the Shock Index[?]. Romero et al.[?] developed a gradient-boosting machine (GBM) model for use as an EWS and demonstrated superior performance compared to the National Early Warning Score 2 (NEWS2)[?]. Finally, Fernandes et al.[?] investigated the predictive value of ED patients' presenting complaints compared to vital signs and other measurements. They used natural language processing (vectorisation with TF-IDF normalisation) to encode unstructured complaint information and trained models on a cohort of approximately 235,000 patients to predict mortality or cardiac arrest. Their experimental findings showed improved predictive performance and calibration when including the chief complaint as a predictor.

This study applies state-of-the-art methods from contemporary machine learning practice to estimate risk of deterioration for acute medical patients in the ED. We bring together findings from prior studies to improve the differentiation of at-risk patients and address challenges that are prerequisites to clinical deployment for a proposed solution. The ED is a fast-paced environment that treats a large volume of patients with varied acuity and is responsible for their initial assessment and clinical management[?]. Operational pressures in EDs are steadily increasing[?], creating an imperative to differentiate the patients with the highest risk efficiently. In our study setting, 'obvious cases' of imminent critical deterioration usually bypass the acute medical team and are escalated immediately. By elimination, the remaining patients are 'less obvious' cases and thus have a greater need for decision support. Conventional, general-purpose EWS are not optimised for specific patient populations or contexts, while 'off-the-shelf' EWS, such as the NEWS2, have variable performance[?]. Recent work argues in favour of centre-specific, locally tailored scores and risk models[?,?]; data-driven solutions deployable at scale can fulfil this role.

We systematically compare the performance of various learning algorithms based on logistic regression (LR), gradient-boosted decision trees (GBDT), and support vector machines (SVM) for predicting imminent clinical deterioration based on cross-sectional patient data extracted from EPR. Our outcome of interest is a composite of in-hospital mortality and admission to critical care to represent severe and time-sensitive medical conditions requiring intervention. We ensure the models' outputted probabilities are well-calibrated and reliable to fit into existing frameworks for assessing clinical utility[?]. Rather than prescribe a specific threshold for classifying high-risk cases, we measure our models' discriminative skill across sensitivities via precision-recall curves and through their daily alert rate, which expresses how they would operate when deployed. We compare our performance against NEWS2, the preferred EWS in the United Kingdom[?].

An extant practical challenge we address is models not generalising to new application environments due to structural differences compared to the development environment[?,?]. Solutions with rigid data requirements unrealistically require providers to conform to a specific pattern of testing or treatment to produce all the requisite data correctly[?]. To avoid making assumptions about data availability or its collection context (such as timing, reliability, or frequency), we conduct experiments using different sets of predictive features that providers might generate under their unique clinical workflow. Starting with vital signs, we gradually construct models with finer information, including manual observations, laboratory results, clinical notes, and service utilisation, to reveal the most influential features.

A further barrier is a requirement for models supporting the clinical workflow to be transparent, safe, fair, and traceable in their decision-making process[?,?]. Machine learning models have conventionally operated as 'black boxes'[?], obscuring their internal reasoning and biases[?,?]. Advances in interpretable machine learning and fairness-aware modelling allow us to address this. We incorporate methods from the fair machine learning literature[?,?] into our evaluation framework to ensure our constructed models do not exhibit unfair bias against individuals or protected demographic groups. Then, we utilise Shapely Additive exPlanations[?], a recently popularised model-agnostic framework for interpreting predictive models, to produce justifications for our models' risk predictions on the individual patient level. These justifications reveal the best-performing models' internal reasoning and allow us to examine and validate the relationships between the significant predictors and the outcome. In addition to predicting a patient's risk, our interpretable models can justify their prediction to the user by isolating the relevant characteristics of the patient that led them to that result[?,?].

Results

Our selected data comprised 121,058 presentations to the acute medical unit (AMU) at Salford Royal Hospital, Manchester, UK, corresponding to 62,162 distinct patients over the study period of January 2015 to March 2022. We identified 8,341 critical deterioration events, of which 2,893 occurred within 24 hours after admission. The Methods Table **??** summarises the dataset and presents the stratification of samples across the three data subsets we used in our analysis: we partitioned the

samples chronologically 2:1 into a model development set and a validation set, and we additionally considered an 'unseen' subset of the validation set that excludes the 8,139 patients (13.09%, making up 42.24% of the validation set's records) that had prior admission records in the training set. The rates of critical care admission, mortality, and composite critical deterioration were uniform across the chronological split.

Figure 1. Average precision **(a)** and Area under receiver operating curve **(b)** achieved by the best predictive models per learning algorithm across tested sets of data features. Each group corresponds to independent models trained with the indicated feature set concatenated to all the previous feature sets to its right. The error bars represent 95% bootstrapped confidence intervals. Obs: Supplemental observations & phenotype, Labs: Laboratory results, Notes: Clinical notes, Services: Triage & service utilisation. We detail the contents of the feature sets in Methods Table **??**.

We compared numerous modelling pipeline variations as described in the Methods section. From this comparison, we identified LightGBM, a variant of GBDT, as the best-performing learning algorithm overall and logistic regression with L2 penalty (LR-L2) as the best linear model. We summarise their performance in Table **??**. Figure **??** compares the average precision (AP) and area under the receiver operating curve (AUROC) of the best predictive models across classifier types on the complete validation set against the measured performance of the reference model (NEWS2) on this patient cohort. The groups in each plot correspond to incrementally augmenting the training data - the leftmost groups of each section present models using only vital signs as predictors, and subsequent groups give the results when we concatenated the indicated feature set (as described in Methods Table **??**) to the previous training inputs. We test these sets of features in order of 'centre-specificity', so that the most clinically standardised predictors, such as vital signs, are considered first. We provide the actual measurements with bootstrapped confidence intervals and the performance on the 'unseen' validation set in Supplementary Tables 7 and 8.

Figure 2. Alert Rate vs Sensitivity **(a,c)** and Precision-Recall curves **(b,d)**. Top Row **(a,b)**: All learning algorithms trained on the complete feature set (equiv. "& Services"). Bottom Row **(c,d)**: GBDT (LightGBM) across feature sets (concatenated incrementally). In **(a,c)**, the Alert Rate curve plots the arithmetic mean of daily positive predictions (alerts) across the validation period for a given sensitivity value (y-axis) against that sensitivity value (x-axis). The point where two lines intersect corresponds to the maximum achievable sensitivity for which the model with the lower line maintains a lower daily alert rate than the model with the upper line. In **(b,d)**, the Precision-Recall (PR) curve presents the positive predictive value (PPV, or precision) on the y-axis against sensitivity on the x-axis. On the PR curve, an unskilled model giving random outputs would yield a horizontal line at $y = P/(P+N)$, where P and N are the numbers of positive and negative samples in the data, respectively, while a theoretical 'perfect' model would yield a single point $(1,1)$ in the upper-right corner of the plot. The curves are plotted from each model's outputted predictions for the complete validation set.

Metric	Estimator	Dataset	Vitals	& Obs	& Labs	& Notes	& Services
AP	LR-L2	Complete	0.157	0.258	0.259	0.264	0.466
		Unseen	0.172	0.309	0.305	0.312	0.478
	LightGBM	Complete	0.179	0.314	0.314	0.323	**0.513**
		Unseen	0.206	0.370	0.365	0.374	**0.523**
AUROC	LR-L2	Complete	0.812	0.836	0.838	0.851	0.894
		Unseen	0.823	0.843	0.845	0.854	0.896
	LightGBM	Complete	0.835	0.868	0.888	0.891	**0.922**
		Unseen	0.842	0.879	0.896	0.897	**0.923**

Table 1. Summary of model performance. Average precision (AP) and Area under receiver operating curve (AUROC) of GBDT (LightGBM) and logistic regression with L2 penalty (LR-L2) for predicting 24-hour critical deteriorations on the two validation sets: 'Complete', the full validation set, and 'Unseen', which includes only patients who had no admissions in the training dataset. Each column corresponds to independent models trained with the indicated feature set concatenated to all the previous feature sets to its right.

Data-driven modelling matched or outperformed the reference model across all feature sets, with the complete feature set (rightmost group in each section of Figure **??**) giving the best performance. Both AP and AUROC trended upward as the

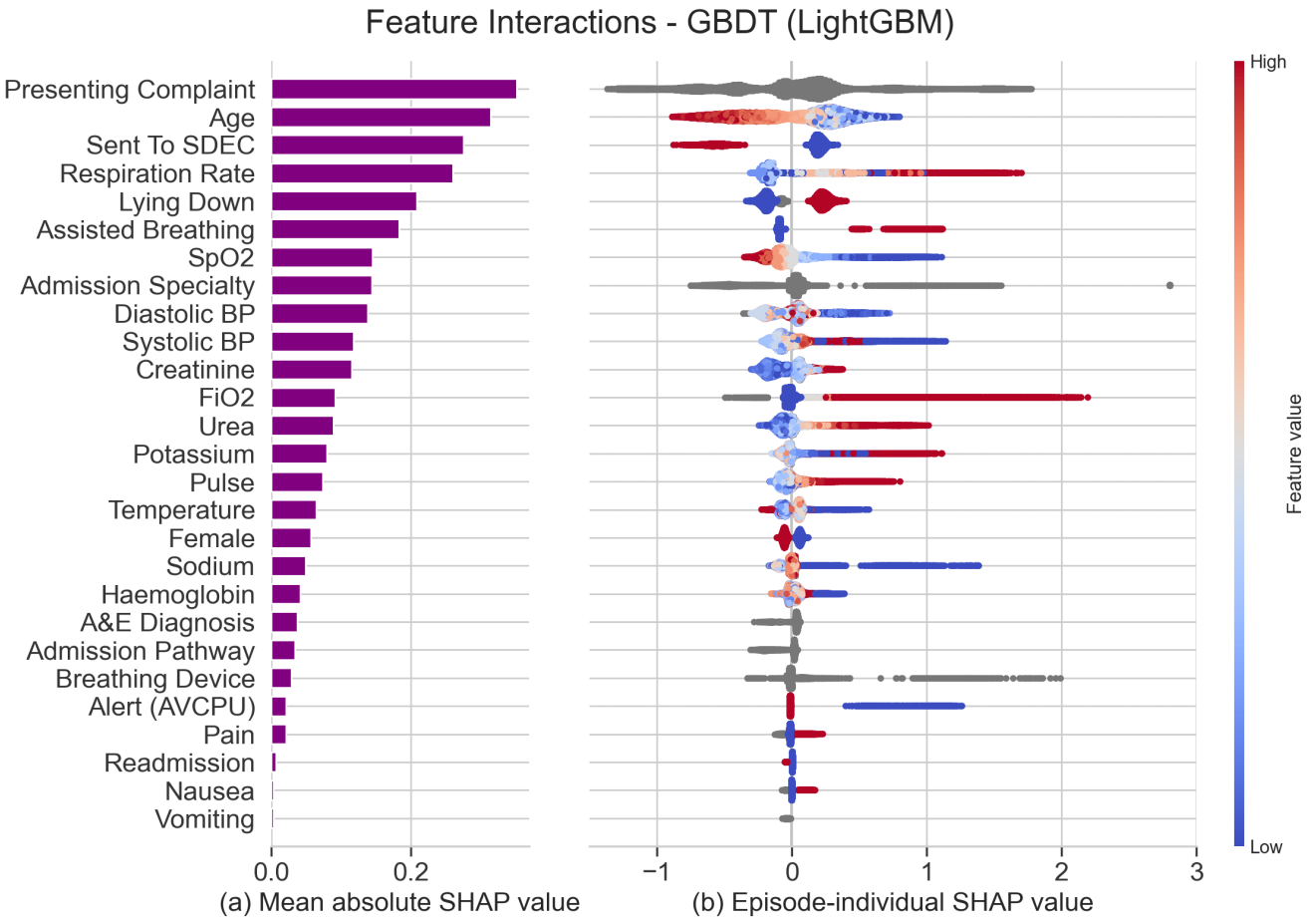

Figure 3. Induced feature importances for GBDT (LightGBM) in decreasing order of mean absolute impact. In **(a)**, the bar lengths represent the mean absolute impact of each feature on the model's predictions for the validation set. In **(b)**, each point represents a value from one admission record. The points' colour corresponds to numerical value, and their position on the x-axis represents the magnitude of their contribution towards increasing the predicted risk (if $x > 0$) or reducing it (if $x < 0$).

number of predictors grew, though phenotype and supplemental observations ("& Obs"), laboratory results ("& Labs"), and clinical notes ("& Notes") had a greater impact on the average precision while the AUROC remained more stable. Including triage and service utilisation ("& Services") yielded the largest singular boost in AP (increase from 0.324 $\rightarrow$ 0.513 for GBDT). Figure **??** illustrates the alert rate vs sensitivity and precision-recall curves for GBDT across different feature sets and for all classifier types trained on the complete feature set. GBDT produced fewer alerts per day on average compared to the reference model up to very high sensitivities (0.966), and all classifiers maintained an improved alert rate up to moderately high sensitivities (> 0.80). The largest reduction of alert rate was at sensitivity 0.864, where GBDT yielded 9.869 daily alerts, 18.50% less than NEWS2's 12.110. The positive predictive value (PPV) of GBDT-Vitals behaves similarly to the reference model as we vary sensitivity. Performance was stable between the "Complete" and "Unseen" validation sets, with a median increase of 0.054 for AP and 0.012 for AUROC when the 'known' patients were removed, as shown in Supplementary Tables 7 and 8. All models had satisfactory calibration, though with a tendency to underestimate the probability of critical deterioration, as illustrated in Supplementary Figure 8.

The feature interactions induced by SHAP for GBDT yielded further insight into the predictors' relationship with our clinical outcome. Figure **??** (a) ranks all the included predictors by their mean absolute impact towards positive predictions (deterioration) and negative ones (no deterioration) across the validation set, (b) illustrates the patient-individual impact of each feature, and Supplementary Figure 10 breaks down the relative impact of the values taken by categorical data features. The presenting complaint ranked the highest and contributed similarly towards positives ("diabetes", "GI bleeding") and negatives ("back pain", "facial problems"). The model captured a non-linear relationship between risk and indicators of kidney function, such as creatinine and urea levels, which is consistent with clinical findings differentiating the mortality risk of acute kidney

injury versus chronic disease[?]. Triage decisions were heavily influential, with same day emergency care (SDEC) invariably reducing the estimated risk, while certain clinical specialities, such as respiratory medicine, geriatric medicine, and general medicine (a catch-all for non-specialty cases), strongly contributed towards positives.

Similarly, we record the coefficients of the logistic regression models in Supplementary Tables 9 and 10 and find them to be consistent across the penalised models. SDEC, higher sodium levels, and specific presenting complaints (e.g., "facial problems", "ear problems") reduce the estimated risk. Conversely, elevated respiratory rate, potassium levels, requiring a bed, and certain clinical specialities and breathing devices yield increased risk estimates. It is interesting to notice that age is assigned a negative coefficient. Figure **??** reveals that GBDT also identified age as a strong predictor, with advanced age driving the model towards negative predictions rather than positive ones. We explore this non-intuitive and potentially spurious association in Figure **??**(a,b) which compares the two models' patient-individual SHAP values for the age feature. We theorise this relationship is partly due to high-frailty patients (aged ≥ 80 years), having the lowest proportion of 24-hour critical deterioration events out of all age groups (as shown in Supplementary Figure 7) despite being very frequent attendees at the ED.

Figure 4. Feature-specific importances extracted by SHAP from GBDT **(a,c)** and LR-L2 **(b,d)**. The top section **(a,b)** presents the importances of patient age, while the lower section **(c,d)** presents body temperature. Each point represents a value from one validation set record. The points' position on x-axis represents the numerical feature value, while the y-axis indicates their contribution to the prediction for that patient, with values above $y = 0$ (indicated in red) contributing towards making the prediction positive and values below $y = 0$ (indicated in blue) contributing towards making the prediction negative.

Figure 5. Average precision (AP) of **(a)** LR-L2 and **(b)** GBDT. Each pair of bars corresponds to incrementally including the indicated feature sets (from Methods Table **??**) as training data. For a given feature set, we measure the AP of two independently trained models, one using the direct measurements of vital signs (blue), and one with the vital signs encoded using the NEWS2 severity scales (red). The error bars represent 95% bootstrapped confidence intervals.

As an additional test, we trained logistic regression and GBDT models with vital signs encoded into integers $0-3$ per the NEWS2 severity scales[?]. We compared the results with the classifiers' performance when using the original vital sign values to investigate how each model type captures the non-linear relationship between vitals and clinical outcomes in Figure **??**. We observe that the 'handcrafted' scales boosted the performance of logistic regression across feature sets, while GBDT's performance either dropped or remained stable. Figure **??**(c,d) presents an example of a diverging relationship learned by GBDT and LR from the same feature, temperature. Note that the presented results thus far assume 24 hours after admission as the cut-off point for identifying deterioration events. Supplementary Figure 9 illustrates how the AUROC of GBDT and LR-L2 varied when we increased the (cumulative) time threshold gradually from 24 hours to 30 days. Across all feature sets, the AUROC trended downwards as the cut-off widened and the on-admission measurements for each newly included sample became more distant from the outcome.

Finally, Figure **??** presents the generalised entropy index vs sensitivity for GBDT across the tested feature sets and all models trained on the complete feature set. Supplementary Figure 11 isolates the between-group fairness component of the generalised entropy index when we consider the population groups defined by the protected demographic characteristics of age group and sex (as specified in Supplementary Figure 7). All models except for GBDT-Vitals achieve an improved fairness score compared to the reference model across sensitivity thresholds. NEWS2 produces a better between-group fairness and, correspondingly, a more significant unfairness within the demographic groups, under the complete feature set above sensitivities of ~ 0.82. To account for potential pre-existing inequalities in the cohort, we record the differential bias amplification of the models in Supplementary Table 11. These measurements corroborate the generalised entropy findings, with a positive bias amplification under the vital signs feature set when considering age groups. However, this diminishes when considering intersectional protected groups of both age and sex. Bias amplification values across all other feature sets are strongly negative - indicating removal of bias - or near zero. We theorise that this unusual amplification of inequality with respect to age groups is due to the vital signs feature set containing insufficient information to predict our tracked outcome correctly for patients of all ages.

Figure 6. Generalised Entropy vs Sensitivity curves of **(a)**: GBDT across the tested feature sets, and **(b)**: All classifier types trained on the complete feature set. We plot each model's generalised entropy index for a given sensitivity value (y-axis) against that sensitivity value (x-axis). A lower value on the y-axis indicates a more fair distribution of 'benefit', i.e. of receiving a positive prediction. A theoretical 'perfect' model would yield a single point $(0,1)$ in the lower-right corner of the plot.

Discussion

In a large cohort of ED admissions, we developed and validated predictive models that can differentiate patients likely to deteriorate shortly after admission. GBDT methods received the most focus as they are state-of-the-art for sparse classification tasks (even compared to deep neural networks[?]), they can capture non-linear interactions such as those present in clinical data, and they natively incorporate missing values, which are inevitable under typical clinical workflows. Using our trained models' coefficients and the extracted global justifications, we can identify which characteristics of our cohort were most predictive of the tracked clinical outcome both on the patient level and across the studied population. Features that encode the clinical context of the patient's condition, presentation, and comorbidities stood out as the most useful. These included presenting complaints, triage decisions such as the utilisation of SDEC, and the assigned clinical speciality, among others. Patient age stood out for being inversely correlated with our tracked outcome, against clinical intuition[?], which we theorise results from the low prevalence of the outcome within the highest age band. While it did not result in the model amplifying unfair bias, it presents a clear example of model interpretability revealing spurious associations that might require correcting prior to deployment.

This study differs further from prior literature in using as representative a data sample as possible and in prioritising practical concerns. Our cohort, with patients of varied acuity and conditions, reflects a typical real world ED acute medical workload. Frontline staff collected the patient data under everyday conditions, where operational pressures affect the timeliness and reliability of data entry. We excluded little data since, although comprehensive manual data curation is helpful for model development, it conflicts with scalable deployment and real-time use of data-driven systems[?] and can lead us to discard valuable information for uncommon cases[?]. We did not carry out a priori feature selection but instead used all available data and employed modelling methods that perform intrinsic feature selection and can differentiate useful features based on evidence. Healthcare digitalisation is an ongoing process[?], so we made no assumptions about the level of EPR integration. Instead, through our experiments with different feature sets, we accommodate different levels of data availability. The lack of a standardised benchmark dataset makes direct comparisons between studies on this topic challenging, so we minimised centre-specific assumptions and standardised our modelling pipelines' structure to establish reproducibility.

We similarly designed our assessment methodology around the extant practical challenges and presented results with the context of their resource cost. We used a temporal split of the study data to assess performance but retained the records where the patient had presented to the same ED during the training period as frequent repeat attendees reflect the reality of clinical practice. To strengthen our results, however, we also examined removing these records and still demonstrated good performance. Calibration is often underappreciated[?], and alert frequency deserves attention as alert fatigue is a key critique aimed at existing solutions from frontline staff[?]. We focused on measuring discriminative skill and avoided setting a threshold for positive or negative classifications, as setting it carries clinical, operational, and ethical complications. Directing care where it is needed promptly is vital and far outweighs the cost of false positives. However, excessive false alarms are detrimental to a model's utility due to alert fatigue[?,?,?]. Balancing clinical risk against available capacity is a well-researched problem beyond the scope of our study[?,?]; instead, we argue that early-stage researchers should aim to maximise the discriminative skill of their model, as might be measured by AUROC or the highest achievable sensitivity while preserving acceptable specificity.

Our observational dataset is limited to one acute secondary care centre, but many measured parameters and outcomes vary between providers. Even near-universal predictors such as vital signs may be measured differently. For example, manual measurement of respiratory rate is less precise than an electronic recording[?], provision of supplemental oxygen is subjective and depends on operational constraints, availability, guidelines, and expertise[?], and the same oxygen saturation may represent different levels of clinical risk depending on whether it was measured before or after commencing oxygen[?]. Furthermore, we recorded symptoms, vital signs, and laboratory results from the point of admission. This information gives a cross-sectional view of the patient's condition as seen by the admitting clinician but excludes longitudinal information, which prior work has collected via continuous vital sign monitoring and used to train highly effective models[?,?]. Finally, we investigated unfair bias and group inequalities in the models to the best of our ability but limited our assessment to the available protected characteristics. While patients face divergent clinical risks depending on characteristics such as sex, age, or ethnic background[?], finer data such as economic stability, education, community context, and other social determinants of health are also strong predictors of clinical risk[?]. We recommend that researchers investigate fairness thoroughly, especially if the models they construct are intended to autonomously screen or prioritise patients' access to care, to ensure healthcare inequalities are not perpetuated[?].

There are key considerations researchers should take into account before adopting similar modelling methodologies. It is essential to consider the validity of jointly modelling outcomes and the reliability of any composite outcome as a surrogate for clinical deterioration. We considered critical care admission and mortality as a single outcome because we expect both to be preceded by deranged physiology, and the clinical response to both, in terms of urgency and skill, is similar[?]. The joint outcome served as a surrogate for any severe and time-sensitive medical condition encountered at the ED; this is a common modelling choice in the literature[?,?] and one we find reasonable, as our focus is on clinical escalation, which is the primary purpose of an EWS[?]. However, critical care and mortality represent competing outcomes as the former intends to prevent the latter[?]. Future studies may prefer to avoid such assumptions and investigate multiclass modelling or compositing multiple

binary classifiers, each trained to identify a single measurable outcome. Some features we utilised, such as triage outcomes, directly represent clinical decision-making. Their inclusion is in contrast with the 'one-size-fits-all' approach taken by the NEWS2[?] or their explicit exclusion by some studies to avoid capturing and amplifying human-originated bias[?]. If the purpose of a system is to 'sense-check' clinical decisions, its input data should ideally be as isolated as possible from those decisions. However, our findings show that these features efficiently stratify patient risk, making them valuable for producing reliable clinical risk estimates as long as the risks are made clear and considered.

In conclusion, we demonstrated the development of predictive models on a large, real-world sample of general ED patients. Considering the high and rising pressures EDs face and the potential for missed diagnoses, models built from continuing our work could be clinically valuable for decision support. We contend that this study demonstrates the power of machine learning for modelling or adapting to patient populations for this task. By incorporating modularised modelling pipelines from contemporary machine learning practice and leveraging the advances in interpretable modelling, we encourage future research to follow a systematised model-building approach and help obtain clinically useful prognostic tools.

Methods

Data Collection and Preparation

Study Setting. Salford Royal Hospital is a digitally mature, 'paper light' NHS secondary care hospital with over $100,000$ emergency department attendances and $\sim 40,000$ unplanned admissions annually. The Hospital's EPR captures clinical episode data in real-time from arrival at the emergency department until discharge. Selected data are exported pseudonymously to an internal data warehouse to drive local quality improvement and service development projects. Our study considered all such records from 1st January 2015 to 31st March 2022. This starting date reflects the first calendar year after the introduction of electronic NEWS recording in the Hospital. We selected all patients aged ≥ 18 years admitted to the Acute Medical Unit (at Salford, known as the Emergency Admissions Unit, EAU). Our data include patients who received ambulatory emergency care (AEC) and same-day emergency care (SDEC[?]) but exclude planned admissions and day case reviews. We present summary statistics of the dataset in Table **??**, and further details of the collected categorical features in Supplementary Table 6.

Data Collection. As a routine part of EAU admission, the responsible staff member (nurse or support worker) records the patient's vital signs within a target of 30 minutes of arrival. The vital signs that make up the NEWS2[?] are measured in a standardised manner using Dinamap monitors, and manually transcribed into EPR. These data are body temperature ($^{\circ}C$), heart rate (beats/min), systolic (and diastolic) blood pressure (mmHg), and peripheral oxygen saturation (%). Other parameters are measured using manual observation and direct questions. These are the patient's level of consciousness (AVCPU), presence of pain, nausea, or vomiting, whether the patient was receiving oxygen at the time of SpO2 measurement and, if applicable, the oxygen flow rate and mode of delivery.

Independent of this, blood test results are automatically recorded in the laboratory information management system (LIMS) and copied to EPR in real time. Whether a patient receives routine blood tests depends on operational pressures and considerations at the ED, not on the patient's presentation. Other information available upon arrival at the EAU includes identifier data such as the unique patient number; basic phenotypic information, such as their age and sex; admission pathway (e.g., ED, emergency GP referral); arrival time; and unstructured notes indicating their presenting complaint and the ED staff's primary diagnosis. For patients with prior hospital visits, significant comorbidities and previous admission events are available from the point of admission.

Retrospective Data Augmentation. Following initial collection, our data are supplemented with downstream administrative and outcome information, including ICD-10, OPCS-4, and HRG codes, alongside service utilisation records, 30-day readmission, and community mortality events. The date and time of discharge, the total length of stay (LOS), the wards the patient was admitted to (in chronological order), and the LOS per ward are collated upon the patient's discharge. Each record includes up to seven concurrent diagnoses, represented by 4-7-character alphanumeric codes per the ICD-10-CM standard. These diagnoses are compiled after discharge by a clinical coding team based on the existing codes and clinical notes recorded in EPR. Procedures and service utilisation are similarly recorded in detail in EPR and coded in retrospect using the OPCS-4 standard.

Feature Engineering. Some of the collected data is not directly clinically relevant or may be unsuitable for modelling under a realistic use case. However, we can use it to engineer useful features or delineate subpopulations in the cohort for more detailed model evaluation. Other features are relevant but first require cleaning or modification. We derive the following features:

- 30-day Readmission. We mark as readmissions those patient records that are preceded by a record bearing the same unique patient ID if the two records' admission dates are ≤ 30 days apart.
- Unstructured Clinical (ED) Notes. The presenting complaint and ED diagnosis are unstructured text and thus could hold any string value. We cluster presenting complaints into a categorical variable representation since the 50 most

Group	Variable	Total	Training	Test (Complete)	Test (Unseen)
Episode	**LOS (days)**	2.27 (0.65-7.13)	2.06 (0.63-6.68)	2.88 (0.71-8.49)	2.20 (0.55-7.50)
	Patients	62162	44752	25549	17410
	Records	121058	81108	39950	23075
Outcomes	**30-day Mortality**	3967 (3.28%)	2589 (3.19%)	1378 (3.45%)	692 (3%)
	Critical Care	4012 (3.31%)	2822 (3.48%)	1190 (2.98%)	716 (3.10%)
	Critical Events	8341 (6.89%)	5489 (6.77%)	2852 (7.14%)	1537 (6.66%)
	In-hospital Mortality	5124 (4.23%)	3233 (3.99%)	1891 (4.73%)	965 (4.18%)
Vitals	**AVCPU-A**	119493 (98.71%)	80013 (98.65%)	39480 (98.82%)	22827 (98.93%)
	Assisted Breathing	12167 (10.05%)	7835 (9.66%)	4332 (10.84%)	2245 (9.73%)
	NEWS	1 (0-2)	1 (0-2)	1 (0-2)	1 (0-2)
	Pulse (beats/min)	80 (70-90)	80 (70-90)	80 (70-90)	80 (70-90)
	RR (breaths/min)	17 (16-18)	17 (16-18)	18 (16-18)	17 (16-18)
	SpO2 (%)	97 (96-98)	97 (96-98)	97 (96-98)	97 (96-98)
	Systolic BP (mmHg)	124 (113-139)	122 (112-138)	125 (114-140)	125 (114-140)
	Temperature (oC)	36.70 (36.40-37)	36.70 (36.40-37)	36.70 (36.40-37)	36.70 (36.40-37)
Supplemental Observations & Phenotype	**Age (years)**	69 (50-82)	69 (50-82)	68 (50-81)	64 (44-79)
	Diastolic BP (mmHg)	70 (60-80)	70 (60-78)	70 (62-80)	70 (62-80)
	Female	63414 (52.38%)	42768 (52.73%)	20646 (51.68%)	11571 (50.15%)
	FiO2 (%)	0 (0-0)	0 (0-0)	0 (0-0)	0 (0-0)
	Lying Down	56565 (46.73%)	36197 (44.63%)	20368 (50.98%)	11368 (49.27%)
	Nausea	1934 (1.60%)	1407 (1.73%)	527 (1.32%)	278 (1.20%)
	Pain	18504.0 (15.29%)	13308.0 (16.41%)	5196.0 (13.01%)	3227.0 (13.98%)
	Vomiting	607 (0.50%)	417 (0.51%)	190 (0.48%)	109 (0.47%)
Labs	**Creatinine (mmol/L)**	79 (63-105)	78 (63-103)	79 (64-106)	78 (63-100)
	Haemoglobin (g/L)	130 (114-143)	129 (114-142)	130 (114-143)	132 (117-145)
	Potassium (mEg/L)	4.20 (3.90-4.50)	4.20 (3.90-4.50)	4.20 (3.90-4.50)	4.20 (3.90-4.50)
	Sodium (mmol/L)	138 (135-140)	138 (135-140)	138 (135-140)	138 (135-140)
	Urea (mmol/L)	6.40 (4.60-9.60)	6.30 (4.60-9.40)	6.40 (4.60-9.70)	6 (4.50-8.90)
Service Utilisation	**30-day Readmission**	15387 (12.71%)	10732 (13.23%)	4655 (11.65%)	1919 (8.32%)
	SDEC	29096 (24.03%)	21020 (25.92%)	8076 (20.22%)	5719 (24.78%)

Table 2. Summary statistics of the study sample. Numerical patient characteristics of acute medical unit admissions, chronologically partitioned into training and testing sets. "Test (Unseen)" corresponds to the chronologically split validation set but excluding patients who had any prior admissions in the training set. Binary variables are reported as "number of positives (%)", while numerical variables are reported as quartiles.

frequent values account for nearly all records (97.06%), and we assign the remainder a sentinel value. In contrast, the ED diagnosis varies greatly between records, so we compile a list of clinically relevant word stems and abbreviations based on expert opinion and construct a boolean Bag-of-Words vector for each record indicating which ones are present. We provide the prevalent presenting complaint values and diagnosis stems in Supplementary Table 6.

- Vital Signs. We investigate training models directly on vital sign readings or encoding them into integers $0-3$ per the NEWS2 severity scales[?]. The former approach forces models to form evidence-based weightings for values that correlate with adverse patient outcomes, while the latter allows us to incorporate the domain knowledge embedded in the NEWS2 into the models. Recorded vitals must be checked for spurious values as they are the only parameters transcribed into EPR manually under a typical workflow. We check each record against fixed ranges (e.g., 0-100% for SpO2) and soft thresholds based on the range of physiologically possible values determined by expert clinical opinion. We provide further details on filtering these values in Supplementary Table 4.

Data Labelling. Our tracked outcome is a composite of in-hospital mortality or admission to critical care from the ward within a specified time threshold after presenting to the ED. The criteria to identify patient episodes that belong in the positive class are:

1. The episode is not excluded by our filtering criteria, i.e. they have at least recorded vital signs, are ≥ 18 years old, are not pregnant, and did not receive critical care interventions on-arrival.

2. The discharge/end-of-episode record indicates the patient died in the hospital AND the record's timestamp is within 24 hours of the admission timestamp, OR their service utilisation indicates admission to critical care or provision of critical interventions on the ward AND this occurred within 24 hours of the admission timestamp.

We identify critical care based on recorded admission into the hospital's critical care unit (CCU) or the high-dependency medical unit (H1). We use the length-of-stay per ward to determine how long after the patient's arrival they were admitted to critical care. A smaller subset of patients received critical care interventions without being moved to these wards, and we can detect most such cases through specific entries in their recorded procedures - OPCS-4 codes E85.1 (invasive ventilation), X50.3 (advanced cardiac pulmonary resuscitation), X50.4 (external ventricular defibrillation), or X56.* (intubation of the trachea).

Model Development.

Modelling Pipeline. We adopt a modularised model-building approach from contemporary machine learning practice. We consider *pipelines* as sequences of distinct tasks in the model-building process, where each task's output becomes the subsequent task's input. Some tasks modify the data samples in preparation for modelling. At least one task in each pipeline is a learning/model-building algorithm. Then, subsequent post-processing tasks may alter the predictive model's output or aggregate multiple models. We implement the following tasks, executed in order:

1. Data Pre-processing. Executes the data preparation tasks outlined previously to produce a vector representing each patient episode. We parameterise the processing component to include only the features we specify, so we may investigate selectively including features and the impact they have on performance. The sets of features we consider are listed in Table **??**.

2. Data Splitting. Partitions the data into two subsets; we use one for model construction and reserve the second for validation. We prefer a temporal train-test split over standard random splitting[?], and partition the dataset such that the first $2/3$ of records chronologically serve as the training set and the latter $1/3$ as the validation set. For some experiments we implement an additional filter that excludes any validation set records where the patient, as identified by their unique ID, had also appeared in the training set in a previous admission.

3. Data Imputation. Supplements standard values into data samples with empty fields. We apply this only to those modelling algorithms that are incompatible with missing data in their inputs (logistic regression). We impute numerical features with the median over the training dataset and binary and categorical variables with appropriate constant values. The imputed values correspond to a patient in stable condition.

4. Model Construction. A learning algorithm receives the data samples and produces a predictive model.

5. Calibration. As a post-processing step, we map the numerical outputs of the trained predictive model into well-calibrated probabilities, substituting the model's original output $C(\boldsymbol{x_i})$ on input $\boldsymbol{x_i}$ for an estimate of $Pr(y_i = 1|C(\boldsymbol{x_i}))$, the conditional probability of belonging to class y_i. We opt for isotonic calibration[?] and fit a meta-estimator that learns the isotonic (monotonically increasing) mapping m that minimises a loss function $\mathcal{L} = \sum_i w_i(y_i - m(C(\boldsymbol{x_i}))^2$.

Feature Set	Features (Units)
Vital signs (NEWS2)	Body temperature ($^{\circ}C$), heart rate (beats/min), systolic blood pressure (mmHg), peripheral oxygen saturation (%)
Supplemental Obs. & Phenotype	Sex (M/F), Age (years), Diastolic blood pressure (mmHg), breathing device (if applicable), prescribed oxygen (FiO2), presence of pain (Y/N), presence of nausea (Y/N), presence of vomiting (Y/N), lying down (Y/N)
Clinical Notes	Presenting complaint (text), ED diagnosis (text)
Laboratory Results	Haemoglobin (g/L), urea (serum, mmol/L), sodium (serum, mmol/L), potassium (serum, mEq/L), creatinine (mcmol/L)
Service Utilisation	Triaged to SDEC (Y/N), readmission within 30 days (Y/N), admission speciality (category), admission pathway (category)

Table 3. Dataset features and units categorised into feature sets. In the given units, "Y/N" indicates binary variables, "category" un-ordered categorical variables, "text" unstructured text data, and "M/F" indicates male or female.

Model Training and Tuning. We construct pipelines with each combination of available components. For each one, we execute a single-objective Bayesian optimisation process (Tree-Structured Parzen approach[?]) to sweep over the space of possible hyperparameter values and probabilistically settle on values that maximise our chosen performance metric, average precision. We construct the final models using the best-scoring hyperparameters after 1000 tuning iterations. We report the resultant hyperparameters in Supplementary Table 5. We avoid training the calibration meta-estimator on the same data that trained the classifier and, instead, we combine calibration with k-fold cross-validation. We randomly separate the training dataset into k equal-sized partitions (setting parameter $k = 5$), train a model on four of the subsets and fit the calibrator using the remaining subset. We iteratively repeat this k times to such that each partition serves as the calibration set once and produce k independent models to serve as sub-estimators of a model ensemble. The final 'representative' probability prediction of the ensemble C of sub-estimators $C_1, \cdots, C_k$ for input vector $\boldsymbol{x}$ is taken to be the arithmetic mean of the sub-estimators' predictions: $C(\boldsymbol{x}) = \frac{1}{k}\sum_{i=1}^{k} C_i(\boldsymbol{x})$.

Model Evaluation. We assess the discriminative skill of the models by constructing the precision-recall curve and measuring the average precision, which is the mean of the PPV (or precision) over the interval of sensitivity (TPR/recall) values from 0-1. We approximate this with the weighted mean of the measured PPV across the observed sensitivity thresholds, where the weight of each element is the difference in sensitivity from the previous element[?].

$$AP = \int_0^1 p(r)dr \approx \sum_{k=1}^{n} P(k)\Delta r(k)$$

where $p(r)$ is the PPV as a function of sensitivity r, $P(k)$ is the precision at cut-off k in the ranked sequence of data samples in the validation dataset, and $\Delta r(k)$ is the difference in recall $r_k - r_{k-1}$. We calculate the confidence intervals for our estimate of the AP by bootstrapping with 1000 bootstrap samples over the validation set[?]. We construct the PR curve by plotting the PPV on the y-axis against sensitivity on the x-axis[?]. On the PR curve, an unskilled model giving random outputs would yield a horizontal line at $y = P/(P+N)$, where P and N are the numbers of positive and negative samples in the data, respectively, while a theoretical 'perfect' model would yield a single point $(1,1)$ in the upper-right corner of the plot.

We construct the receiver-operating characteristics (ROC) curve and compute the area under the receiver operating curve (AUROC). We plot the false-positive rate (1 minus the specificity) on the x-axis against the sensitivity on the y-axis. The minimum possible area under the curve is 0.5, corresponding to a completely random relationship between the model's output and the ground truth. Generally, $0.7-0.8$ indicates reasonable discrimination, and values over 0.8 indicate good discrimination[?]. We compute confidence intervals for the AUROC as before.

The ROC and PR curves both provide a model-wide evaluation and, while the ROC curve is more common, we prefer the PR curve because it better indicates the skill of the model at predicting the minority (positive) class correctly and is less influenced by predicting the majority (negative) class correctly[?]. The PR curve further allows us to visually inspect how quickly PPV deteriorates as we increase model sensitivity[?], which is helpful in a task where it may be appropriate to value sensitivity over specificity.

Finally, we investigate how a model's daily alert rate varies with sensitivity[?]. We construct an alert rate curve by plotting the alert rate (the number of positive predictions divided by the number of days) on the y-axis over sensitivity on the x-axis.

The point where two lines intersect corresponds to the maximum achievable sensitivity for which the model with the lower line maintains a lower daily alert rate than the model with the upper line.

Model Bias We investigate two forms of undesirable bias: individual, representing how dissimilarly we treat individuals who deserve similar outcomes[?], and group-based, measuring the inequality of predictions between demographic groups defined by protected characteristics[?]. The generalised entropy index[?] applies to both notions concurrently. Given a patient record $\boldsymbol{x}_i$ with ground-truth outcome y_i, we define the *benefit* experienced by the patient due to model prediction $C(\boldsymbol{x}_i)$ as:

$$b_i = y_i - C(\boldsymbol{x}_i) + 1$$

Under this representation, a false-positive patient experiences a large benefit ($b = 2$), while a false-negative that the model missed has the heaviest penalty ($b = 0$). Then, given the vector of benefit values over the validation set, $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$, and their arithmetic mean $\mu(\boldsymbol{b})$, we measure the generalised entropy index fairness score $\mathscr{E}^2_{\boldsymbol{b}}$, where:

$$\mathscr{E}^{\alpha}(\boldsymbol{b}) = \frac{1}{n\alpha(\alpha-1)} \sum_{i=1}^{n} \left(\left(\frac{b_i}{\mu(\boldsymbol{b})} \right)^{\alpha} - 1 \right)$$

Furthermore, given protected groups $g \in G$, with each comprising n_g patient records with benefit vectors $\boldsymbol{b}^g = (b^g_1, b^g_2, \cdots, b^g_{n_g})$, we decompose the generalised entropy into its between-group component $\mathscr{E}^2_{\beta}$ and its within-group component $\mathscr{E}^2_{\omega}$, representing group and individual fairness, respectively. We measure the between-group component $\mathscr{E}^2_{\beta}$, where:

$$\mathscr{E}^{\alpha}_{\beta}(\boldsymbol{b}) = \mathscr{E}^{\alpha}(\boldsymbol{b}) - \mathscr{E}^{\alpha}_{\omega}(\boldsymbol{b}) = \sum_{g=1}^{|G|} \frac{n_g}{n\alpha(\alpha-1)} \left(\left(\frac{\mu(\boldsymbol{b}_g)}{\mu(\boldsymbol{b})} \right)^{\alpha} - 1 \right)$$

We define demographic groups based on the available protected characteristics - age and biological sex. We partition the continuous age variable into age groups, as illustrated in Supplementary Figure 7. For both scores, the ideal value is 0 and higher values indicate unfair classification.

We additionally compute the differential fairness bias amplification exhibited by our models[?]. The differential fairness metric is defined from the standpoint of intersectionality, i.e., equally protecting population sub-groups defined by multiple overlapping protected characteristics. Bias amplification measures a predictive model's unfairness compared to any pre-existing bias reflected in the dataset due to inequality in the real-life generative process of the data. Given a set of patient records $\boldsymbol{x}$ and protected groups $(g_i, g_j) \in G \times G$, the (smoothed) differential fairness ε of a classifier C is defined by the relation:

$$e^{-\varepsilon} \leq \frac{\sum_{\boldsymbol{x} \in g_i} C(\boldsymbol{x}) + \alpha}{|g_i| + |R_Y|\alpha} \frac{|g_j| + |R_Y|\alpha}{\sum_{\boldsymbol{x} \in g_j} C(\boldsymbol{x}) + \alpha} \leq e^{\varepsilon}$$

where $|R_Y|\alpha$ is the Dirichlet smoothing concentration parameter (we set $\alpha = 1.0$, assuming no prior information). Then, the bias amplification metric is defined as the difference $\varepsilon_C - \varepsilon_D$ of the differential fairness value for the model C minus the value for the dataset D's ground truth. A negative bias amplification indicates that the predictive model reduces differential unfairness, while a positive value means the estimator is more biased than the original data.

Data Availability

The data that facilitated the experiments of this study are provided by the Northern Care Alliance NHS Trust, but restrictions apply to the availability of this data, which were used under a data sharing agreement with Durham University for the current study, and so are not publicly available.

Acknowledgements

The authors would like to thank Durham University, Biophysical Sciences Institute for supporting this work via the Summer Research Bursary.

Author contributions statement

All authors carried out method and experimentation design. D.G. collected the data. S.B.L. carried out the experiments, summarised the results and prepared figures. All authors analysed and interpreted the results. S.B.L. wrote the manuscript text and all authors reviewed the manuscript.