Developing a deep learning system to drive the work of the critical care outreach team

Georgina Kennedy*[1], John Rihari-Thomas[2], Mark Dras[3], and Blanca Gallego[1]

[1] *Centre for Big Data Research in Health, University of New South Wales*
[2] *Nursing Research Institute, Australian Catholic University*
[3] *Department of Computing, Macquarie University*

Abstract

Care of patients at risk of deterioration on acute medical and surgical wards requires timely identification, increased monitoring and robust escalation procedures. The critical care outreach role brings specialist-trained critical care nurses and physicians into acute wards to facilitate these processes. Performing this role is challenging, as the breadth of information synthesis required is both high and rapidly updating.

We propose a novel automated 'watch-list' to identify patients at high risk of deterioration, to help prioritise the work of the outreach team.

This system takes data from the electronic medical record in real-time and creates a discrete tokenized trajectory, which is fed into a recurrent neural network model. These models achieve an AUROC of 0.928 for inpatient death and 0.778 for unplanned ICU admission (within 24 hours), which compares favourably with existing early warning scores and is comparable with proof of concept deep learning systems requiring significantly more input data.

*Corresponding Author: georgina.kennedy@unsw.edu.au

1 Background

1.1 Context

For a patient in an acute-care setting, there are many complex and interrelated factors that affect their likely trajectory toward either recovery or deterioration. Prior to significant deterioration events, there are observable patterns in clinical features that indicate this change in acuity [1, 2, 3, 4]. These warning signs may be present as much as 48 hours prior to the adverse outcome [1], however they are often overlooked.

In addition, there is evidence that sub-optimal care (including delayed or missed interventions) in general hospital wards is a key contributing factor to both unplanned ICU admissions and preventable inpatient mortality [5, 6].

These factors have combined to drive the modern desire for tools and processes that can accurately highlight patients at risk of deterioration on the general wards such that interventions can be deployed sooner, improving both patient outcomes and resource utilisation. This commonly takes the form of an early warning score such as NEWS [7], which tracks physiological variables and raises an alert when they fall outside of acceptable limits. It may also include the establishment of a critical care outreach team whose purpose is to integrate critical care skills of advanced assessment into the general care wards, by providing resources to follow those discharged from intensive care unit (ICU) beds to support recovery, and by anticipating deterioration that could potentially be averted in order to reduce unplanned ICU admissions [8, 9].

Critical-care outreach is a challenging role, requiring a rapidly updating awareness of events and patients across the whole hospital. In order to effectively prioritise their distributed workload, critical-care outreach nurses and medical officers (CCON & CCOM) must synthesise information on a broader scale than is required of typical ward staff. A physiological early warning score is intended to provide a trigger for emergency response, however the intention of the outreach role is to identify potential deterioration and act prior to emergency onset, which therefore requires a more nuanced measure of deterioration risk.

There has been much interest in the development of deep learning models derived from electronic medical record (EMR) data. Deep learning techniques are robust to heterogeneous, sparse and messy data, which are defining characteristics of the EMR. EMR data also fit naturally into recurrent neural network (RNN) architectures due to the discrete, episodic, time-series nature of the patient trajectory, which draws robust analogies to models of language. These language models have recently been expanded to account for the variable time intervals present in the patient record [10, 11, 12] by incorporating time-modulation gates or weightings for elapsed time.

Recurrent models have been developed from EMR data with high accuracy for diagnostic, phenotyping and prognostic purposes in diverse clinical domains. In particular, such systems have been demonstrated to perform well when used to predict in-patient mortality and ICU admission [10, 13, 14], which are the

most important end-points for understanding short-term risk in a general patient population.

Even though existing techniques have theoretical suitability and high accuracy, they have not been successfully translated into practice. This is due to limitations of the models themselves (interpretability, standardisation, actionability of recommendations), technical limitations (integration with clinical systems), and the tension between model output and clinician knowledge and intuition [15, 16, 17].

1.2 Aim

The primary aim of this project is to investigate the feasibility of an automatically generated watch-list that provides outreach staff with an ordered list of patients most at risk of short-term deterioration. By analysing all available data in the medical record as it is generated, this list can supplement the clinical judgement of the CCON & CCOM and help them to proactively identify patients in need of early intervention to improve outcomes, avoid unnecessary or ineffective ICU admissions and reduce the risk of unexpected death.

The watch-list does not attempt to form a specific diagnosis nor prognosis but rather produces a priority list that can sit alongside clinical judgment. Users are therefore less tied to strictly explainable inference, requiring only a meaningfully calibrated relative risk. As such, we propose that it is a good candidate for piloting a real deep learning system in the clinical workflow. Preliminary user discussions suggest an openness to augment their work-flow in this way, and a lower barrier for requiring exhaustive model scrutability due to the fact that the existing mental model for this role is so burdensome.

1.3 Contributions

We present here a model of patient deterioration that considers multiple short prediction time-frames, across both mortality and unplanned ICU admission. Most patient risk models generate a prediction for their defined endpoint at a single future time (typically either in-admission or within 24 hours). This single prediction point does not represent the true, timely clinical risk to a given patient, so results in more alerts than reasonably can (or indeed should) be addressed by outreach staff. Using multiple shorter prediction time-frames, across both inpatient mortality and unplanned ICU admission endpoints further skews an already highly imbalanced dataset, bringing with it additional challenges. For this reason we have developed an augmentation process that is suitable for discrete time-series data and a calibration process that allows sensible interpretation of predicted probabilities in highly unbalanced datasets.

The majority of available models are based on patient vital signs, which are not universally available in the clinical record, and therefore cannot be implemented as an automated system. These results demonstrate the potential of a patient deterioration model that is not reliant on vital signs, which establishes a viable alternative in such settings.

2 Results

2.1 Models

In order to predict the likelihood of the events of interest at numerous future time points an LSTM-based deep neural network [18] was trained against all patient history data available at the time of prediction.

2.1.1 Inputs

For this work, we used a dataset of hospital admissions from a metropolitan tertiary hospital in Sydney, Australia. This data was gathered retrospectively and was approved for use by the target institution's Human Research Ethics Committee.

All historical entries in the EMR were converted to discrete token values, based on their event type (admission/discharge, historical diagnosis, pathology results, medication administration, ward movement, surgical procedure or demography). These tokenised events were then concatenated to form a list of discrete values describing the patient's historical trajectory that could be fed into the prediction model.

2.1.2 Targets

Events of interest are defined as in-hospital death and unplanned ICU admission. There is no distinction made as to whether a death occurs in general wards, theatre or in the ICU.

No predictions are generated for patients in the ICU at the time of prediction, as they are under the care of the core ICU team. An ICU admission is classified as planned if it follows immediately from a surgical procedure, as there is similarly no intervention required from outreach staff. Patients admitted directly to ICU are excluded from these models (363 admissions). In order to allow all states to be mutually exclusive, we train separate models for ICU admission and death risk.

Prediction time (t=0hr) is set to 24 hours after a patient is admitted to general medical wards, either directly or via transfer from the emergency department. Prediction endpoints are measured at 12 hourly intervals, up to 4 days into the future ($t = 0 + [12, 24, 36...96]$ hours).

2.2 Data

Input data for these models included 192,883 hospitalisations, belonging to 92,802 adult patients (44.05% female), undergoing 117,658 surgical procedures over the period from June 2008 to June 2016. Patients had between one and 899 visits in the time period. Patients with 100 or more admissions (129 patients - all receiving regularly scheduled dialysis or rehabilitation treatments) were removed from the dataset so that they did not overwhelm the models, leaving a range of 1-99 admissions per patient (mean 2.08 $\pm$ 3.92).

Patients had an average of 3864 ± 7221 included clinical tokens at admission time. For admissions lasting more than 24 hours, 65 ± 40 additional events were captured within the first day.

Admissions had one primary diagnosis and up to 44 associated comorbidities (mean 4.63 ± 4.08). Every admission included by definition at least one ward movement (the ward to which the patient was initially admitted). Detailed summary statistics of the data can be found in Supplementary Materials A.

2.3 Endpoint Rates

Data imbalance is a well known challenge in the development of machine learning models. This is particularly relevant when the minority class is the class of interest, which is frequently the case in models that predict mortality, specific diagnoses or other important clinical end points.

In the source admissions, there was an overall inpatient death rate of 1.53% and unplanned ICU admission rate of 3.22%. These rates change over the course of admission time, however, and drop drastically as the time windows become shorter (see Figure 1). At 24 hours after admission, the rate of death in the next 24 hours is 0.35% and for unplanned ICU admission it is 0.61%.

Unplanned ICU admission rates peak in the first day of admission and remain steady after that. Once an admission lasts more than 12 hours, the death rate becomes much higher. This is likely to represent the low death incidence within day-surgery admissions. From 12 hours onwards, the rate rises more gradually as the less severely ill patients are discharged. As death rates rise, unplanned ICU rates fall, which is indicative of an overall increase in acuity over time despite a decrease in instability.

These changes in patient population at different time points after admission also demonstrate the value of providing multiple future target endpoints, as a single model is unlikely to perform equally well in all instances.

2.4 Reported Metrics

We report here metrics that test the output predictions against three measures:

1. A strictly correct forecast (model predicts endpoint within t hours, and this reflects accurately the presence of this endpoint within t hours).

2. A forecast that is correct with a clinically-relevant tolerance. This tolerance is set to 72 hours (model predicts endpoint within t hours, and this reflects accurately the presence of this endpoint within $t + 72$ hours), to account for patients where similar response from outreach staff may be appropriate, given the desire for early intervention.

3. A forecast that is correct within the target admission (endpoint is predicted within t hours, and this is not necessarily accurate, however the endpoint of interest does occur prior to discharge). This gives a better

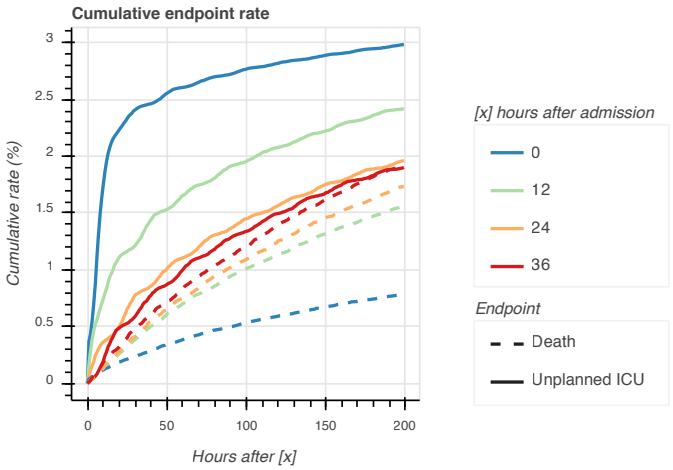

Figure 1: Endpoint rates in source data, relative to the number of patients still admitted at the given prediction point.

sense of the true burden of false positives and false negatives on both patients and outreach staff.

For prediction use-cases with such high degrees of imbalance as those targeted by these models, with far more negative cases than cases of interest, reporting the area under the receiver operator curve (AUC) alone can be highly misleading [19]. Despite this, it remains the most commonly reported statistic of model quality.

For this reason, we also report here the sensitivity and workup to detection ratio (WDR) for every prediction target. Model sensitivity is calculated as true positive predictions divided by all positive cases, or $\frac{TP}{TP+FN}$. WDR is the inverse of the model positive predictive value, and provides the ratio of all positive predictions to all true positive predictions i.e. $\frac{1}{PPV}$, or $\frac{TP+FP}{TP}$.

Sensitivity is the key outcome measure from the perspective of at-risk patients. This is because a false negative corresponds to potential missed interventions and directly impacts their outcomes. WDR is the key metric for outreach staff however, as an increase in the burden of false positives will heavily reduce the usefulness of any predictive model, and may draw clinicians away from truly deteriorating patients. If balanced appropriately, these measures will result in the predictive model with the highest clinical utility.

2.4.1 Example

Figure 2 shows an example of the inputs and prediction targets used to develop these predictive models. This example patient has two historical admissions (1a, 1b) prior to the current admission. Both historical admissions were for planned

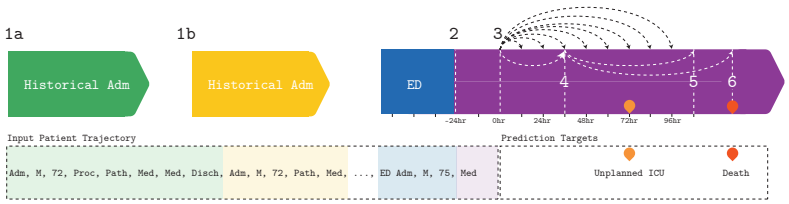

Figure 2: Example patient trajectory

procedures, and include a mix of demographic and clinical tokens.

In the target admission, the patient was admitted via the Emergency Department. Admission time (2) is the time that the patient was transferred to the medical wards. Prediction time (3) is set to 24 hours after admission time (t=0hr). All demographic and clinical tokens up to prediction time are included in the input data. Thus the input trajectory is an ordered list of all tokens occurring in any available historical admission(s), the patient's ED stay, and the first 24 hours of the target admission.

In this example, there is an unplanned ICU admission at t=72hr, and the patient dies outside of the prediction window, but within this admission. At t=36hr (4), neither endpoint has occurred, so a prediction of false is strictly correct. Unplanned ICU admission does occur within 36+72 hours however (5), and therefore a prediction of ICU=true would be correct within the tolerance window and a prediction of death would be correct within the target admission.

2.5 Mortality Prediction

At 24 hours after admission, death within the following 24 hours was predicted with an AUC of 0.928 (see Table 2.5 for all time points). This is higher than the baseline score NEWS [7] (0.89), however as outlined above, this measure alone is unlikely to tell the whole story of model utility. Note also that the NEWS baseline could not be replicated in the source data due to the unavailability of patient vital signs so is compared only to the AUC as reported in the cited study.

Figure 3 demonstrates the discriminative value of this model, i.e. that the output does indeed correspond to prediction of clinically meaningful risk. Although the sensitivity is poor at the earliest time point (due to the enormous class imbalance) later forecasts can be expected to correctly predict between a quarter and a third of patients who will deteriorate rapidly. Sensitivity drops as the tolerance increases to 72 hours. The workup to detection ratio decreases much more rapidly, however, demonstrating that the clinical burden of a false positive in this model is low, and that responding to a patient with even moderate risk is likely to be worthwhile.

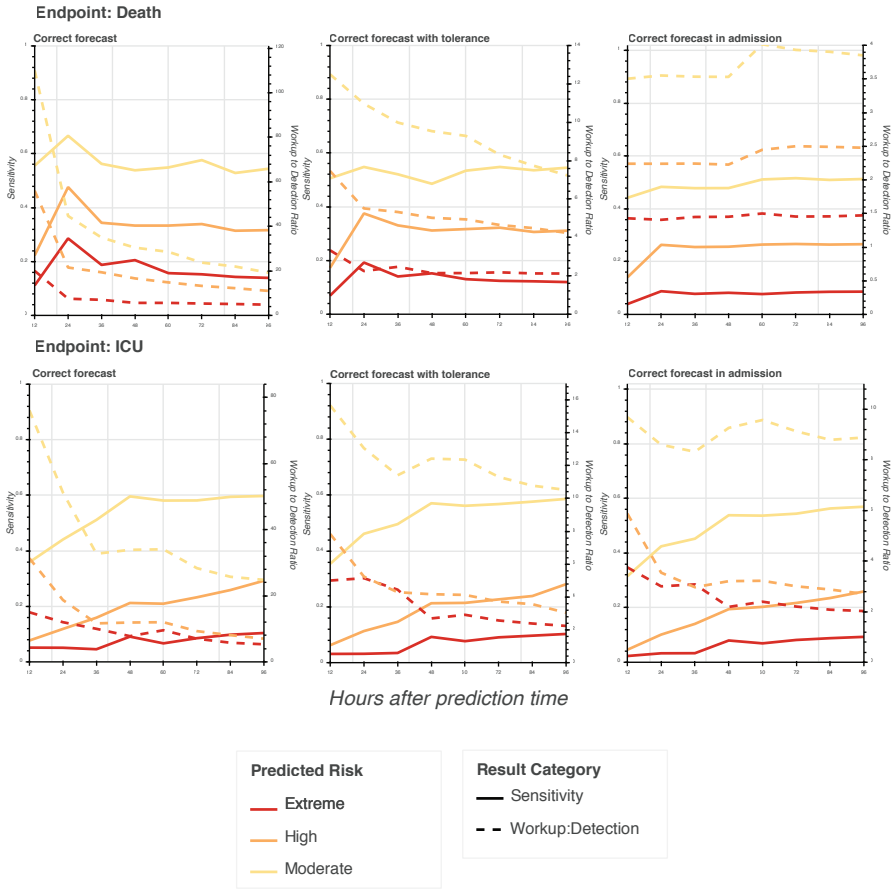

Figure 3: Mortality and unplanned ICU prediction — sensitivity and WDR of death prediction at future time points using data available at 24 hours after admission

2.6 Unplanned ICU admission

There is a significant difference between the AUC of the mortality prediction models and the corresponding unplanned ICU admission models. This is likely to be due to the fact that ICU admission criteria are strongly coupled to vital-sign triggers, and therefore a prediction model that does not include this data will underperform.

Despite this, from Figure 3, it remains possible to predict unplanned ICU admissions within the following 48 hours with a sensitivity of around 20% of all cases, and a corresponding WDR of 1 in 17. When allowing a 72 hour window of tolerance, a WDR of 1 in 12 gives up to 60% sensitivity, and therefore still

[x]	12	24	36	48	60	72	84	96
Death								
Correct forecast	0.918	0.928	0.921	0.915	0.906	0.911	0.902	0.902
Correct forecast with tolerance	0.921	0.917	0.911	0.917	0.903	0.902	0.904	0.901
Target within admission	0.901	0.902	0.903	0.902	0.890	0.890	0.891	0.890
Unplanned ICU admission								
Correct forecast	0.747	0.778	0.777	0.776	0.782	0.776	0.789	0.781
Correct forecast with tolerance	0.754	0.783	0.779	0.774	0.781	0.779	0.789	0.786
Target within admission	0.725	0.757	0.743	0.750	0.757	0.753	0.768	0.767

Table 1: Area under the receiver operating curve for prediction within [x] hours, using data available 24 hours after admission time.

represents a tool with meaningful clinical applications.

2.7 Model Calibration

The raw results produced by this model had poor calibration, despite their good discriminative power, meaning that the probabilities output by the models could not be directly interpreted as the actual probability of the event occurring. There was a very low positive class count (not only proportionally, but also numerically) in the small calibration set. This meant that typical recalibration methods of isotonic regression [20] and Platt scaling [21] were ineffective (see Figure 4).

We observe for this model a significant right skewness in the averaged probabilities, where this model is in fact not confident about the predictions for many of the patients. This is atypical for deep learning models, as they tend to be overconfident, or 'sharp' in their predictions [22]. This indicates that the data augmentation processes are indeed instrumental in providing the desired robustness in detection of temporal relationships, whereby specific feature combinations may be observed and highly weighted from the training set, but this specific relationship is only detected in some of the permuted test samples.

For such short-term deterioration, it is a reasonable expectation that the proportion of patients deemed at low risk will far outweigh those at high risk. We follow the argument in [22] for the use of unevenly spaced bins to generate measures of calibration quality to its logical conclusion and use these unevenly spaced bins to form the basis of the recalibration function itself.

We find that the highest probability that we can assign to precise death forecasts is 40%, deaths within 72 hours of their forecast time have a maximal confidence of 80% and in-admission death has a maximum confidence of 90%. This matches the expectation that clinical trajectories are non-deterministic, particularly over the short term, but as the precise prediction time expands, confidence increases. Thus, a patient who is given a risk score of 10 at the time point 12 hours in the future, will have a 40% likelihood of dying within the next 12 hours, 80% likelihood in the next 12+72 hours and 90% likelihood of dying within this admission.

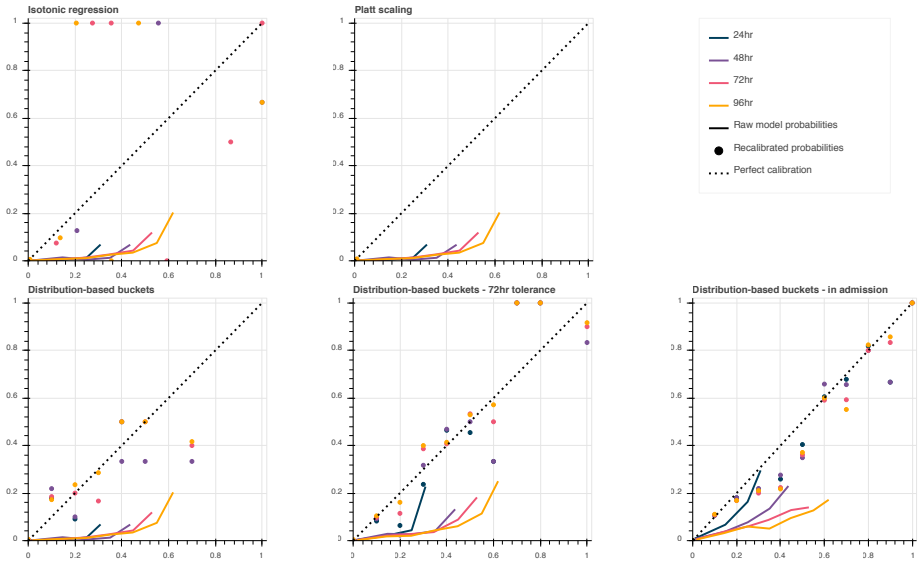

Figure 4: Recalibration techniques for death model predicted at 24 hours after admission. Note that Platt scaling reduces all probabilities to a single point close to the origin.

2.8 Capacity for Intervention

Smith et al [7] states that *'[the] true benefit of any EWS is its ability to recognise patients who are deteriorating but who can have their outcome changed by a timely intervention, rather than to predict those who are destined to die'*. In addition, it is to be expected that at least some false positive predictions will represent a group of patients for whom clinical intervention did indeed avert a real crisis event. It is a known issue with clinical prediction models that accounting for the effects of missed or delivered interventions is not always possible from data alone.

Therefore, in order to understand the limitations of this model in these contexts, we ran the false positive samples with highest predicted risk (predicted death within 36 hours with a probability of 0.6 or higher but discharged alive) and the false negative samples with lowest predicted risk (died within 24 hours but death probability at 96 hours was lower than 0.2) through the LIME Text Explainer module [23]. LIME is an algorithm that provides insights into a 'black-box' model by learning a locally interpretable model that can explain which input data was most relevant to a given prediction.

There is a clear pattern between the factors that contribute strongly to a prediction of high risk versus those contributing strongly to a low risk prediction (see Figure 5). Lab results are generally indicative of a risk increase, where

1.

2.

Figure 5: Word clouds demonstrating the most highly weighted terms for (1) false positive predictions and (2) false negative predictions.

medications and medication-related tokens dominate lower risk predictions.

For false negatives, most of these drug terms represent the highest-frequency tokens in the corpus. Their interpretation therefore is limited to the fact that they are evidence of a sort of regression to the mean, where these patients simply do not have enough distinctive data at the point of prediction to make an accurate risk assessment. Overall, despite having a comparable number of unique tokens, the medication terms each individually tend to have higher frequency than other token types. This holds true even when accounting for the repeated administration of medications, as these tokens on average each appear in more distinct patient trajectories than other event token types (excluding ward movement tokens).

In the list of terms contributing to false positives, there are numerous terms that may indicate that the patient has a complex history or is in a high-risk category, e.g. low white cell count, high blood urea, medication resistance, artificial opening status, sirolemus testing, low lipase. There are also, however, terms that either don't have a sensible interpretation with respect to deterioration risk, e.g. low bilirubin, low blood alcohol content, Nystatin administration, or that are not sufficiently specific to make an informed interpretation of risk e.g. anaemia, sigmoidoscopy procedure, abdominal x-ray. This system is therefore insufficient to provide directed actions or interventions and its use must be limited to the prioritisation of attention.

3 Discussion

3.1 Source Data Limitations

Scores or tools that target imminent patient deterioration typically aim to detect derangement of physiological signs and symptoms. This is based on the observation of predictable patterns of changes in patient vital signs prior to each of the relevant deterioration end points cardiopulmonary arrest, unplanned ICU admissions and death [1, 2, 3, 24, 25].

Although a physiological early warning score (EWS) is used as a manual trigger of emergency response at the target institution [26], due to a lack of availability of vital sign data within the EMR, it is not currently possible to use such a score as the basis for a fully automated watch-list.

This, along with variable importance analyses in logistic regression models such as [27], serve to highlight the importance of vital sign data as the key element underpinning the vast majority of current best practice for prediction of in-patient deterioration. The limitation seen in our data is a realistic one, however, that should be considered for implementation of a fully automated system. It is characteristic of many EMR systems to serve the purposes of hospital administration first, and support clinically relevant data only where this aligns with the requisite billing and logistical goals, and/or where the clinical utility is high enough to justify the additional documentation burden above what can be provided with paper charts. Thus, it is unsurprising to observe in

this data set that all theatre-based procedures are fully available in the clinical record, as they are not only billable, but also require the booking of resources from a central pool, compared with typical bedside procedures and nursing observations that go unrecorded for the inverse reasons.

This limitation in the breadth of input data is significant, however encourages a model that is built primarily around administrative data points, which are likely to be more reliably and consistently available in the EMR.

3.2 Congruence with Current Clinical Practice

The use of rapid response systems is intended to act as a safety net for deteriorating patients via the monitoring of a standardised subset of patient vital signs. It has, however, been argued that this drives nursing practice towards the detection of deterioration that is already well underway, as opposed to highlighting at-risk patients who are yet to go downhill [28]. By removing the reliance on vital signs, this model affords the capacity to move away from detection and into the realm of prediction.

Studies have also found that workloads and hospital work culture affect the likelihood of staff triggering rapid response calls according to the prescribed protocols [29]. Although calling criteria are nominally specified to allow triggering of the rapid response protocol based on clinical intuition alone (even when vital-signs based criteria are not yet met) nursing staff who wish to act upon early signs of deterioration report themselves to be reluctant to do so in the face of potential criticism. This is true despite the fact that nursing intuition can preempt deterioration identified by vital signs alone [30]. A system that is able to provide contextualisation of such minor changes in patient state is therefore well placed to augment existing escalation protocols. See figure 6 for an example implementation as envisaged.

3.3 Baseline

As a baseline, we present in Table 2 a selection of models that have been developed with the goal of detecting the early stages of short-term patient deterioration in a general ward population. Not all of these baselines can be compared directly to the models presented in this work due to the variability of endpoints and prediction times, giving instead an overview of the general targets and performances in existing models.

Note that it is only possible to compare WDR to baselines reported in different populations if a fixed incidence rate is chosen in order to standardise this measure. Where it was possible to make this calculation, the fixed rate was set to 0.35%, which is the death rate within 24 hours in this population, per section 2.3.

The traditional models were identified from a recent review paper that is closely aligned with the target use-case [31] in addition to the NEWS model [7], which is a highly cited and widely implemented early warning score that forms the basis for comparison for many similar works.

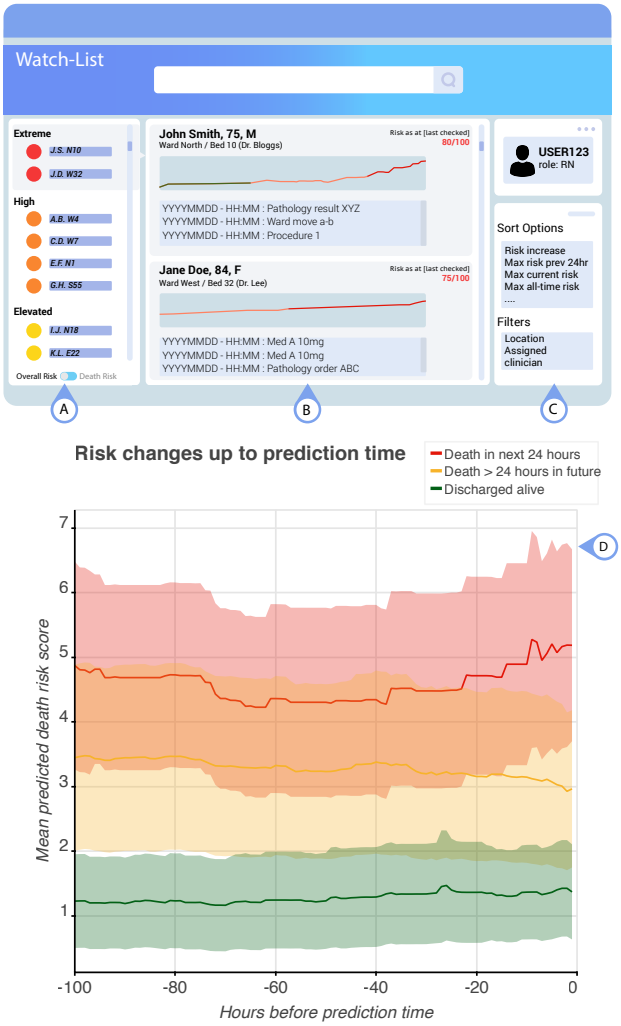

Figure 6: The watch-list as envisaged. (A) Summary panel - see high-level view of cases allowing an immediate overview of the risk profile of the hospital (B) Detail view - see predictions for individual cases over time to understand trends, and highlight events co-occurring with risk changes over time (C) Control panel - sort and summarise by risk category, location, assigned clinician or other relevant parameters (D) It is not possible to give a real example for the detail panel due to privacy restrictions on the source data, so here we present the average risk trend for the first 100 hours of admission for admissions $>$ 100 hours in length in the test set (grouped by actual outcome).

In order to capture potential deep learning baselines, the reference list of two systematic reviews [32, 33] were filtered to identify EMR-based patient deterioration prediction models. General deterioration endpoints not applicable to the CCON/CCOM role were excluded, e.g. readmission, death other than short-term, or studies only applicable to patients already within the ICU. Notably, many deep learning models do not fit our use-case as they either predict only in-patient or longer-term mortality e.g. [34, 35, 10], target a specific morbidity such as congestive heart failure or sepsis e.g. [36, 37] or are developed using data for patients already admitted to the ICU e.g. [38, 39] (largely due to the wide utilisation of the freely-available MIMIC-III database [40]). [10] was retained as the deep learning baseline, as it is closest to meeting the target use-case. Interestingly, this reference uses the NEWS model as a mortality baseline, despite the fact that NEWS was developed to detect 24-hour mortality where the deep learning model predicts in-patient mortality.

This summary of baselines exposes a number of issues with the comparison of such predictive systems. In particular, the precise definition of endpoints is inconsistent. We also note that all mortality endpoints reported here are for in-hospital mortality only, i.e. they are unable to report full mortality as an endpoint due to the lack of data linkage and potential loss to followup. Only Kipnis et al [41] have access to network-level data linkage, but this is not utilised as a primary endpoint. Rajkomar et al [10] go further by redefining readmission to include only readmission to the same institution. The availability of linked data as per [42] would provide additional insight and allow expansion of these models to include identification of patients at the end of life.

3.4 Data Processing

Many clinical prediction scores rely on highly regulated data collection that may not reflect existing clinical processes, thus requiring additional data entry or hand calculation. Our noisy dataset reflects true practice and availability, with pre-processing limited to routines that can be performed with no human input. Within this pre-processing of data, we do not attempt to normalise the labelling of medications and pathology — e.g. different spellings are present for the same test across different panels — instead, allowing contextual embeddings to handle this noise. We do this on the assumption that the more hands off we are in data preparation, the more robust the results will be to changing practice and the lower effort required by the end users. We also do not make any effort to handle multiple recordings at the same time, or detect outliers.

Because we rely on the naturalistic data ecosystem, rather than one requiring abstraction, we assume that we are reducing errors caused by hand calculations or operational error, and robust to errors preexisting within the EMR. The trade-off with this strategy is that we cannot expect these models to achieve generalisation in a new setting without re-training to accommodate local vocabularies and idiosyncrasies of data entry. An external validation study will therefore require translation of the entire model pipeline, rather than transfer and mapping of only the model inputs themselves.

Table 2: Comparison to baseline models

Model	Target Endpoint	Incl/Excl Criteria	Prediction Time	AUROC	Sens.	Spec.	Standardised WDR
		Traditional models					
NEWS [7]	In-hospital death (within 24hr) Unplanned ICU (within 24hr) Cardiac arrest (within 24hr) Combined 24hr deterioration	Ex: Discharged before midnight of admission day; admitted directly to ICU	Time observations taken in medical assessment unit	0.89 0.86 0.72 0.87	- - - -	- - - -	- - - 31.5
Alvarez et al [43]	Resuscitation events and death	Inc: Adult patients admitted to internal medicine ward or ICU. Ex: admitted directly to surgery; DNR order at admission; obstetrics admission; events on first day of admission	Daily prediction	0.85	0.52	0.94	35.6
Churpek et al (a) [44]	Cardiac arrest (in admission) Unplanned ICU (in admission) Cardiac arrest (within 24hr) Unplanned ICU (within 24hr)	Inc: Adult patients with documented vital signs	Every 8 hours	0.88 0.77 0.88 0.76	- 0.54 0.65 -	- 0.90 0.93 -	- 55.6 33.2 -
Churpek et al (b) [27]	Combined 8hr deterioration	Inc: Adult patients with documented vital signs	Every 8 hours	0.80	0.50	0.93	42.8
Kipnis et al [41]	Combined 12hr deterioration	Inc: Adult patients. Ex: out of network transfers; childbirth admissions, 'comfort care only' orders.	Hourly	0.82	0.49	0.92	49.5
Green et al [45]	Combined 24hr deterioration	Inc: All admissions.	At time of vital sign observation	0.80	0.50	0.90	59.9
		Deep learning models					
Rajkomar et al [10]	In-admission death	Inc: Length of stay > 24hr; adult patients	24hr after admission	0.95	-	-	-
CCO watch-list (this work)	Unplanned ICU (within 48hr) In-hospital death (within 24hr)	Inc: Length of stay > 24hr; adult patients, fewer than 100 visits, not admitted directly to ICU	24hr after admission	0.77 0.93	0.50 0.47	0.88 0.97	71.2 21.3

Notes: Where more than one result available for same end-point, result with highest AUROC is reported.
Where more than one prediction time is available, most clinically relevant prediction time for that end-point is reported.
Where multiple cutoff points are available, sensitivity and specificity are reported as per review paper [31].
Workup to detection ratio is only reported where it is possible to standardise this measure to a fixed reference prevalence rate.
Reference rate has been set to 0.35% for all WDR calculations, setting sensitivity in range ~50% per [31].
For NEWS, fixed sensitivity/specificity in target range not available. WDR instead calculated from EWS efficiency curve.

3.5 Calibration Measure

It is not feasible to calculate the Hosmer-Lemeshow statistic of calibration for this model due to the large sample size and excessive degrees of freedom [46, 47]. Alternative calibration statistics were reviewed for their applicability such as [48], however were found to be unsuitable due to their focus on density. This makes sense for many use-cases, where it is valuable to prioritise areas of the calibration curve that represent the majority of samples, however in this situation it is not suitable, as the differences between probabilities at the low end of the risk scale are not clinically meaningful. Instead, the differences in the most sparse regions must be prioritised — outreach staff may be expected to treat patients at 80% risk quite differently to those at 90% risk, despite there being very few patients in those risk categories, where their response will differ very little for patients at 10% risk vs. 20% risk.

This knowledge-based interpretation of the utility of a model's calibration cannot be quantified without some parameters set by target users a priori.

4 Methods

4.1 Data Preparation

Contextual embedding is a technique whereby terms in a text are assigned a representation in a vector space such that their proximity to other terms in some way captures their relatedness – i.e. one would expect similar concepts such as 'eye' and 'sight' to be closer than unrelated terms 'hat' and 'frog'. In an effective embedding space, it may also be possible to capture the nature of a relationship – in an idealised case, 'diabetes' and 'metformin' may have a spatial relationship that is similar to the relationship between 'asthma' and 'seretide' which thus provides an interpretation of 'treats chronically'. This transformation into a space that is of significantly lower dimensionality than simple one-hot representations of the entire vocabulary also allows the neural network equations to become tractable [49].

In order to take advantage of these techniques that were initially developed for natural language processing (NLP) tasks, and as per prior deep learning work with EMR data [50, 51, 52], we converted each entry in the clinical database into token(s) of one of the following types: admission, discharge, pathology result, medication administration, ward movement, surgical procedure.

Pathology results and surgical procedure details contain continuous data types (numerical results, duration respectively), which cannot be handled by a straightforward contextual embedding model. These numerical values are therefore converted to decile results for each test or procedure type respectively.

These tokens are then concatenated for each patient, with their associated time-delta since time of index admission, in order to describe their care trajectory, such as in Figure 2.

All data are inserted into the care trajectory at the time that they become available in the EMR. Ward movements, medication admission, pathol-

ogy result, procedure and theatre movements are incorporated into the EMR in real-time. Some demography data are available at triage time, whilst some demography data are input only at discharge. Diagnosis data are not available in the EMR until some time after the time of discharge due to medical coding procedures. We therefore mask diagnosis codes associated with the target admission and only include historical diagnoses as input data, even for models generated at the time of discharge.

4.1.1 Time-sensitive Concept Embedding

These tokens were then transformed into a lower dimensional embedding space using a modification of the skipgram algorithm [53].

Temporal and relational knowledge was encoded within the embedding by using a sampling function that was weighted inversely proportional to both the time delta between two events, and also whether or not the event occurred in the same admission. In equation 1, s is the distance between the two events by admission (for events in the same admission, *s=0*, for events in the admission immediately prior, *s=1* etc.) and t is the time interval between the two events in hours. This weighting (W) was then used to distribute the likelihood of sampling token pairs for inclusion in the embedding model.

$$W = \frac{1}{(s+1)(t+1e^{-3})^{\frac{1}{100}}} \tag{1}$$

4.1.2 Data Balancing

The targets of this model have a highly imbalanced distribution, which represents a significant challenge in the development of a useful model [54], with imbalances as skewed as 1 event in 160 for unplanned ICU admission and 1 in 180 for death within the shortest time-frames. We use a data augmentation strategy that allows the models to weight the loss functions appropriately and learn a more accurate representation of both the majority and minority classes.

For image-processing tasks, it is typical to flip, rotate, skew, scale and mask portions of the input image in order to create multiple synthetic samples that retain the same class as the source, but allow a network to learn a more robust set of features that are less likely to over-learn idiosyncrasies related strictly to scale and positioning rather than the content of the image itself. Similarly, [55] applies window slicing and window warping strategies to provide synthetic samples from time-series data.

Following from these techniques, we implemented a data augmentation algorithm that can be applied to discrete time-series events such as those present in the EMR.

After copying trajectories and then randomly truncating the copies to 20-100% of their original length (by dropping the oldest events), time-series events were bucketed into 1 hour windows. 1 hour windows were chosen given the likelihood of meaningless time distinctions at any higher resolution based on

an assumption of primarily manual data-entry processes. Events within each of these 1-hour windows were then randomly shuffled and/or masked to create modulated patient trajectories which could be used to augment the input data. Each trajectory not including the target event was randomly augmented 4 times. Trajectories that included the target were augmented at a rate that was inversely proportional to the time to event (thus emphasising indicators of proximal deterioration), producing a balanced dataset. In the validation and test datasets, all trajectories were augmented 30 times.

4.2 Final Models

The final model architecture was made up of three sub models that were trained jointly (Figure 7).

Model 1: A flat set of features was created for each admission (see Table 4.2). These flat features were fed into a dense feed-forward network with a 4 dimensional output branch (Death, ICU, Discharge, Ward) for each of 8 time points (12, 24, 36, 48, 60, 72, 84 & 96 hours in the future). Terminal layer activation was set to Softmax, all prior layers had a LeakyReLU activation. This is the most straight-foward NN architecture, fully connected between layers.

Model 2: The most recent 500 tokens in the patient trajectory were fed into a bi-directional LSTM layer, which then connected to a densely connected network, trained with the same 8 output branches as Model 1. Activations were also set as per Model 1. An LSTM is a recurrent neural network, which is suited to sequential data.

Model 3: The 64 output variables from models 1 and 2 were concatenated into a single vector and used to train a densely connected network, with binary outcomes (i.e., death/ death or ICU/ ICU) at each of the target times.

4.2.1 Training Process

These models were trained jointly, so a single training batch was fed into models 1 and 2, with the resulting gradients back propagated, and then the output of this same batch was fed into model 3 and back propagated before moving onto the next training batch.

The models were trained on all 8 output times (12 to 96 hour forecasts) for 15 epochs and then the loss function was modified to attend to the first 4 output times only and trained for a further 5 epochs in order to improve the temporal discrimination.

A 10% test set (randomly selected) was held out with no processing applied until both ICU and Death model training was completed, with the remaining 90% used in a 5-fold cross validation process. At each fold, the training data was split into 80% training, 5% calibration and 15% validation sets.

Feature	Range	Available (%)	Most common
Age (yrs)	18-114	100	71
Marital Status	8 distinct	96.2	Married/partner
Aboriginal or Torres Strait Islander Ethnicity	7 distinct	98.7	Neither
Insurance Status	94 distinct	100	Medicare - overnight
Postcode of residence	1497 distinct	100	Postcode of hospital
Country of birth	220 distinct	98.8	Australia
Relative admission day	0-2000	100	1703
Admission hour	1-23	100	7 (07:00hr)
Admission day of week	0-7	100	3 (Tuesday)
Admission speciality	47	98.6	Emergency
Last discharge code	9 distinct	99.8	9 (discharge alive)
Days since last stay	0-2000	62	2
Last length of stay (LOS)	0-475	62	0
Historical total LOS	0-592	62	0
Historical average LOS	0-237	62	0
Historical ICU hours total	0-2733	62	0
Historical ICU hours mean	0-1665	62	0

Table 3: Flat demography and historical summary features for each admission

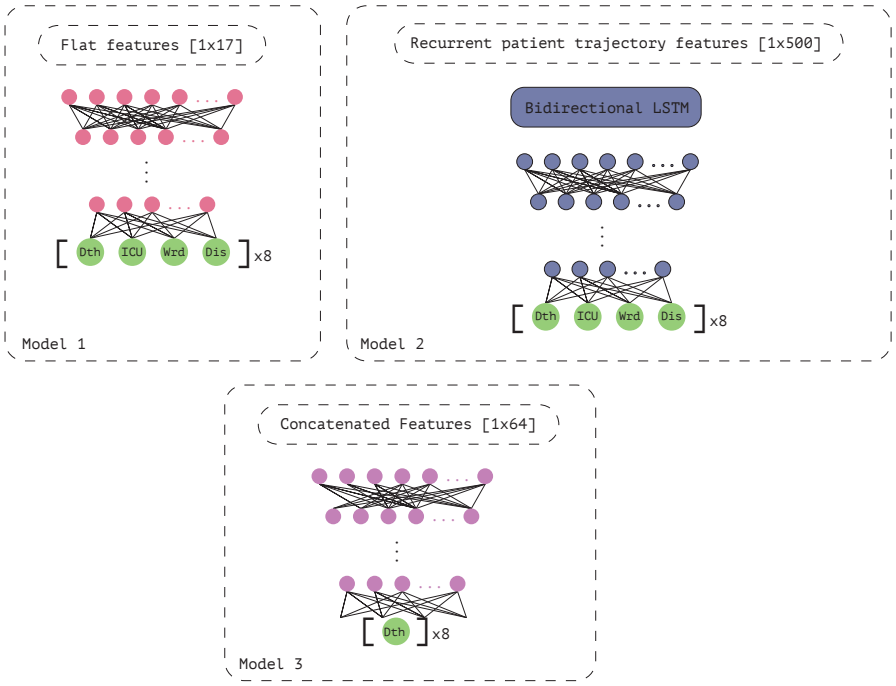

Figure 7: Final model architecture

4.2.2 Process to predict on unseen data

Using the 30 augmented trajectories that were created for each validation and test sample, the final model was sampled 10 times with dropout active (Monte Carlo dropout [56]), generating 300 predictions for each patient.

The models were not constructed in such a way that required forecast risk to be monotonically increasing over time, although this is a logical requirement given that we predict the occurrence of an endpoint at $time \leq t$. Thus, with a forecast risk score of S_t for a given time point, the risk score for $time = t + 1$ was defined as $max(S_t, S_{t+1})$.

4.2.3 Calibration

A reference distribution of risks and uncertainty were produced by generating 300 predictions for each patient in the calibration set as per the validation data. This distribution was then bucketed using a stick-breaking process at the quantiles $[0, 1 - \frac{1}{2}^{0+\alpha}, 1 - \frac{1}{2}^{2+\alpha}, ..., 1 - \frac{1}{2}^{10+\alpha}]$ to generate scoring thresholds that appropriately reflected the far higher proportion of patients in low risk categories. A different α was selected using the accuracy of calibration set predictions for each category (correct, correct+72 hours, correct in admission).

The risk score between 0 and 10 was then generated by comparing the predicted probability against these cutoff thresholds. This technique has similarities to Dirichlet calibration maps [57], however quantile cutoffs were used instead of a parametrized function, as the extreme class imbalance meant that the positive class count in the calibration set was very small.

Statements

Contributions Author G.K. performed all data analysis and programming tasks, and bulk of writing. B.G. contributed project concept and high-level direction across all project phases, J.R-T. provided clinical input to anchor the work in real-world practice and M.D. provided technical guidance.

Competing Interests The authors declare no competing interests.

Data availability The data input for the current study are not publicly available. Due to reasonable privacy and security concerns, the data cannot be distributed to researchers other than those granted access via the Human Research Ethics Committee of the target institution.

Code availability The code for this study is not currently available, however this work is being re-implemented as described in the publicly available MIMIC-III data-set, at which point the code will be released and posted to https://github.com/CBDRH/PaTMan

References

[1] Hillman, K., Bristow, P., Chey, T., Daffurn, K., Jacques, T., Norman, S., Bishop, G., Simmons, G.: Antecedents to hospital deaths. Internal medicine journal **31**(6), 343–348 (2001)

[2] Schein, R.M., Hazday, N., Pena, M., Ruben, B.H., Sprung, C.L.: Clinical antecedents to in-hospital cardiopulmonary arrest. Chest **98**(6), 1388–1392 (1990)

[3] Kause, J., Smith, G., Prytherch, D., Parr, M., Flabouris, A., Hillman, K., *et al.*: A comparison of antecedents to cardiac arrests, deaths and emergency intensive care admissions in Australia and New Zealand, and the United Kingdom—the ACADEMIA study. Resuscitation **62**(3), 275–282 (2004)

[4] Buist, M., Bernard, S., Nguyen, T.V., Moore, G., Anderson, J.: Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study. Resuscitation **62**(2), 137–141 (2004)

[5] McQuillan, P., Pilkington, S., Allan, A., Taylor, B., Short, A., Morgan, G., Nielsen, M., Barrett, D., Smith, G.: Confidential inquiry into quality of care before admission to intensive care. BMJ

[6] Dubois, R.W., Brook, R.H.: Preventable deaths: who, how often, and why? Annals of Internal Medicine **109**(7), 582–589 (1988)

[7] Smith, G.B., Prytherch, D.R., Meredith, P., Schmidt, P.E., Featherstone, P.I.: The ability of the national early warning score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation **84**(4), 465–470 (2013)

[8] Department of Health (United Kingdom): Comprehensive critical care – A review of adult critical care services (2000)

[9] McDonnell, A., Esmonde, L., Morgan, R., Brown, R., Bray, K., Parry, G., Adam, S., Sinclair, R., Harvey, S., Mays, N., *et al.*: The provision of critical care outreach services in England: findings from a national survey. Journal of critical care **22**(3), 212–218 (2007)

[10] Rajkomar, A., Oren, E., Chen, K., Dai, A.M., Hajaj, N., Hardt, M., Liu, P.J., Liu, X., Marcus, J., Sun, M., *et al.*: Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine **1**(1), 18 (2018)

[11] Baytas, I.M., Xiao, C., Zhang, X., Wang, F., Jain, A.K., Zhou, J.: Patient subtyping via time-aware LSTM networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 65–74 (2017). ACM

[12] Wang, L., Wang, H., Song, Y., Wang, Q.: Mcpl-based ft-lstm: Medical representation learning-based clinical prediction model for time series events. IEEE Accesss **7**, 70253–70264 (2019)

[13] Laksana, E., Aczon, M., Ho, L., Carlin, C., Ledbetter, D., Wetzel, R.: The impact of extraneous variables on the performance of recurrent neural network models in clinical tasks. arXiv preprint arXiv:1904.01125 (2019)

[14] Gupta, P., Malhotra, P., Vig, L., Shroff, G.: Transfer learning for clinical time series analysis using recurrent neural networks. arXiv preprint arXiv:1807.01705 (2018)

[15] Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., *et al.*: Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface **15**(141), 20170387 (2018)

[16] Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. Briefings in bioinformatics **19**(6), 1236–1246 (2017)

[17] Kennedy, G., Gallego, B.: Clinical prediction rules: A systematic review of healthcare provider opinions and preferences. International journal of medical informatics (2018)

[18] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[19] Raeder, T., Forman, G., Chawla, N.V.: Learning from imbalanced data: Evaluation matters. In: Holmes, D.E., Jain, L.C. (eds.) Data Mining: Foundations and Intelligent Paradigms, pp. 315–331. Springer, Berlin, Heidelberg (2012)

[20] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 694–699 (2002)

[21] Platt, J., *et al.*: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[22] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1321–1330 (2017). JMLR. org

[23] Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the predictions of any classifier. CoRR **abs/1602.04938** (2016). 1602.04938

[24] Franklin, C., Mathew, J.: Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. Critical care medicine **22**(2), 244–247 (1994)

[25] McGloin, H., Adam, S.K., Singer, M.: Unexpected deaths and referrals to intensive care of patients on general wards. Are some cases potentially avoidable? Journal of the Royal College of Physicians of London **33**(3), 255–259 (1999)

[26] Hughes, C., Pain, C., Braithwaite, J., Hillman, K.: "Between the flags": implementing a rapid response system at scale. BMJ Qual Saf **23**(9), 714–717 (2014)

[27] Churpek, M.M., Yuen, T.C., Winslow, C., Meltzer, D.O., Kattan, M.W., Edelson, D.P.: Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical care medicine **44**(2), 368 (2016)

[28] Osborne, S., Douglas, C., Reid, C., Jones, L., Gardner, G., *et al.*: The primacy of vital signs–acute care nurses' and midwives' use of physical assessment skills: a cross sectional study. International Journal of Nursing Studies **52**(5), 951–962 (2015)

[29] Douglas, C., Osborne, S., Windsor, C., Fox, R., Booker, C., Jones, L., Gardner, G.: Nursing and medical perceptions of a hospital rapid response system. Journal of nursing care quality **31**(2), 1–10 (2016)

[30] Douw, G., Schoonhoven, L., Holwerda, T., van Zanten, A.R., van Achterberg, T., van der Hoeven, J.G.: Nurses' worry or concern and early recognition of deteriorating patients on general wards in acute care hospitals: a systematic review. Critical care **19**(1), 230 (2015)

[31] Linnen, D.T., Escobar, G.J., Hu, X., Scruth, E., Liu, V., Stephens, C.: Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and icu transfer: a systematic review. Journal of hospital medicine **14**(3), 161 (2019)

[32] Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association **25**(10), 1419–1428 (2018)

[33] Osmani, V., Li, L., Danieletto, M., Glicksberg, B., Dudley, J., Mayora, O.: Processing of electronic health records using deep learning: a review. arXiv preprint arXiv:1804.01758 (2018)

[34] Mayampurath, A., Sanchez-Pinto, L.N., Carey, K.A., Venable, L.-R., Churpek, M.: Combining patient visual timelines with deep learning to predict mortality. PloS one **14**(7) (2019)

[35] Avati, A., Jung, K., Harman, S., Downing, L., Ng, A., Shah, N.H.: Improving palliative care with deep learning. BMC medical informatics and decision making **18**(4), 122 (2018)

[36] Cheng, Y., Wang, F., Zhang, P., Hu, J.: Risk prediction with electronic health records: A deep learning approach. In: Proceedings of the 2016 SIAM International Conference on Data Mining, pp. 432–440 (2016). SIAM

[37] Kam, H.J., Kim, H.Y.: Learning representations for the early detection of sepsis with deep neural networks. Computers in biology and medicine **89**, 248–255 (2017)

[38] Harutyunyan, H., Khachatrian, H., Kale, D.C., Ver Steeg, G., Galstyan, A.: Multitask learning and benchmarking with clinical time series data. Scientific data **6**(1), 96 (2019)

[39] Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmarking deep learning models on large healthcare datasets. Journal of biomedical informatics **83**, 112–134 (2018)

[40] Johnson, A.E., Pollard, T.J., Shen, L., Li-wei, H.L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A., Mark, R.G.: MIMIC-III, a freely accessible critical care database. Scientific data **3**, 160035 (2016)

[41] Kipnis, P., Turk, B.J., Wulf, D.A., LaGuardia, J.C., Liu, V., Churpek, M.M., Romero-Brufau, S., Escobar, G.J.: Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. Journal of biomedical informatics **64**, 10–19 (2016)

[42] Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., Gallego, B.: Real-time prediction of mortality, readmission, and length of stay using electronic health record data. Journal of the American Medical Informatics Association **23**(3), 553–561 (2015)

[43] Alvarez, C.A., Clark, C.A., Zhang, S., Halm, E.A., Shannon, J.J., Girod, C.E., Cooper, L., Amarasingham, R.: Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC medical informatics and decision making **13**(1), 28 (2013)

[44] Churpek, M.M., Yuen, T.C., Park, S.Y., Gibbons, R., Edelson, D.P.: Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. Critical care medicine **42**(4), 841 (2014)

[45] Green, M., Lander, H., Snyder, A., Hudson, P., Churpek, M., Edelson, D.: Comparison of the Between the Flags calling criteria to the MEWS, NEWS and the electronic Cardiac Arrest Risk Triage (eCART) score for the identification of deteriorating ward patients. Resuscitation **123**, 86–91 (2018)

[46] Hosmer, D.W., Lemesbow, S.: Goodness of fit tests for the multiple logistic regression model. Communications in statistics-Theory and Methods **9**(10), 1043–1069 (1980)

[47] Kramer, A.A., Zimmerman, J.E.: Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. Critical care medicine **35**(9), 2052–2056 (2007)

[48] Austin, P.C., Steyerberg, E.W.: The integrated calibration index (ici) and related metrics for quantifying the calibration of logistic regression models. Statistics in medicine **38**(21), 4051–4065 (2019)

[49] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)

[50] Xiao, C., Ma, T., Dieng, A.B., Blei, D.M., Wang, F.: Readmission prediction via deep contextual embedding of clinical concepts. PloS one **13**(4), 0195024 (2018)

[51] Choi, E., Xiao, C., Stewart, W., Sun, J.: Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In: Advances in Neural Information Processing Systems, pp. 4547–4557 (2018)

[52] Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference, pp. 301–318 (2016)

[53] Guthrie, D., Allison, B., Liu, W., Guthrie, L., Wilks, Y.: A closer look at skip-gram modelling. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06). European Language Resources Association (ELRA), Genoa, Italy (2006)

[54] Branco, P., Torgo, L., Ribeiro, R.P.: A survey of predictive modeling on imbalanced domains. ACM Computing Surveys (CSUR) **49**(2), 31 (2016)

[55] Le Guennec, A., Malinowski, S., Tavenard, R.: Data augmentation for time series classification using convolutional neural networks. (2016)

[56] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, pp. 1050–1059 (2016)

[57] Kull, M., Nieto, M.P., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In: Advances in Neural Information Processing Systems, pp. 12316–12326 (2019)