Development and validation of a dynamic prediction model for unplanned ICU admission and mortality in hospitalized patients

Davide Placido¹ — Hans-Christian Thorsen-Meyer¹-² — Benjamin Skov Kaas-Hansen²-³
— Roc Reguant¹-⁴ - Søren Brunak¹

Affiliations
¹ Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Denmark
² Department of Intensive Care Medicine, Copenhagen University Hospital (Rigshospitalet), Copenhagen, Denmark
³ Section for Biostatistics, Department of Public Health, University of Copenhagen, Denmark
⁴ Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, New South Wales, Sydney, Australia

* Corresponding author. Address: Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Blegdamsvej 3B, 2200 Copenhagen N, Denmark.

ORCID id's
DP: 0000-0002-8327-8838
BSKH: 0000-0003-1023-0371
RR: 0000-0002-0350-3899
HCTM: 0000-0001-7002-3089
SB: 0000-0003-0316-5866

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Frequent assessment of the severity of illness for hospitalized patients is essential in clinical settings to prevent outcomes such as in-hospital mortality and unplanned ICU admission. Classical severity scores have been developed typically using relatively few patient features, especially for intensive care. Recently, deep learning-based models demonstrated better individualized risk assessments compared to classic risk scores such as SOFA and NEWS, thanks to the use of aggregated and more heterogeneous data sources for dynamic risk prediction. We investigated to what extent deep learning methods can capture patterns of longitudinal change in health status using time-stamped data from electronic health records. We used medical history data, biochemical measurements, and the clinical notes from all patients admitted to non-intensive care units in 12 hospitals in Denmark’s Capital Region and Region Zealand during 2011-2016. Data from a total of 852,620 patients and 2,241,849 admissions were used to predict the composite outcome of unplanned ICU transfer and in-hospital death at different time points after admission to general departments. We subsequently examined feature interpretations of the models. The best model used all data modalities with an assessment rate of 6 hours and a prediction window of 14 days, with an AUPRC of 0.287 and AUROC of 0.898. These performances are comparable to the current state of the art and make the model suitable for further prospective validation as a risk assessment tool in a clinical setting.
Introduction

Early warning scores (EWS) are used in the clinic to assess the health status of hospitalized patients. The first EWS were based solely on five physiological parameters followed by many adaptations and improvements (1), (2). VitalPAC was one such early warning score (ViEWS), which introduced modifications to the parameters contributing to the score (3). These modifications were based on clinicians’ knowledge about the relationship between physiological data and adverse clinical outcomes. Further modifications were implemented in the national early warning score, NEWS (4).

These scores are used extensively because their simplicity makes them applicable and easily comparable across different departments and countries. They specifically provide estimates for the risk of adverse outcomes (5) such as cardiac arrest, ICU transfer, and in-hospital mortality. However, they suffer from some limitations. The relatively few data types used limit their predictive power. Relevant features like biochemical measurements, the order of temporality in the data, or non-linear feature interactions are ignored. These risk calculations are made without exploiting all the information that current EHR systems provide (6). Moreover, the scores mostly rely on data that are manually collected by the staff. This often entails a higher number of missing and incorrect data items, as compared to a fully automated approach. Recently it was shown in population-wide data from inpatients of the Capital Region of Denmark, that around 10% of the NEWS records were incomplete and 0.2% had implausible features (7).

To improve prediction accuracy and reduce alarm fatigue, new applications using more data and more sophisticated methodologies have been developed (8). Given the heterogeneous nature of EHR data and the rarity of events such as clinical deterioration, machine learning methods seem well-suited for addressing this task. Deep learning, in particular, has performed well for many similar clinical tasks. For example, Shamout et al. (9) developed a deep learning model that uses vital signs to predict the composite outcome of ICU transfer, cardiac arrest and mortality. Cho et al. (10) developed a deep learning model to predict the composite outcome of cardiac arrest and ICU transfer, while da Silva et al. (11) developed a long-short term memory neural network (LSTM) to predict the worsening of the vital signs using prognostic indexes. A recurrent deep neural network has also been used by Tóth et al. (12) to predict stability of vital signs to avoid unnecessary measurements at nighttime. Similarly, dynamic deep learning models using multiple data sources have been developed for different tasks; Thorsen-Meyer et al. (13) constructed one to predict 90-days mortality after ICU admission, Rajkomar et al. (14) to predict in-hospital death, 30-day unplanned readmission, prolonged length of stay and patient’s final discharge.
diagnoses; Tomašev et al. (15) to predict kidney failure; Lauritsen et al. to predict risk of sepsis and acute critical illness (16) (17).

Although deviating vital signs often precipitate clinical deterioration, their recognition in the general ward requires continuous engagement by health care personnel, which can delay or even preclude their recording. Therefore, we investigated using data collected less frequently in early detection of deterioration, building an end-to-end machine learning pipeline that integrates heterogeneous clinical data to predict the risk of imminent serious clinical deterioration at regular intervals.

To this end, we combined natural language processing algorithms and recurrent neural networks to leverage latent patterns in the data. We subsequently assessed the impact of the single data sources and the so-called tokens extracted from each of them to improve the understanding of the model.

**Figure 1:** General structure of the prediction framework. Given a specific assessment rate (time between two consecutive risk assessments) and prediction window (time window within which the outcome is observed), the risk of clinical deterioration was assessed continuously during each general
admission. All the data up to the time of assessment was used to train the model, which was then evaluated and interpreted on the holdout test data.

Methods

This paper adheres to relevant items in the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis statement (TRIPOD) (18).

Patients and outcome

The data comprise all inpatient admissions, from 2011 to 2016, to 12 public hospitals in the Capital Region of Denmark and Region Zealand. The admissions were pieced together by concatenating consecutive inpatient visits 24 hours apart so that the department transfers were not considered two separate events.

Direct ICU admissions, outpatients and emergency admissions were excluded, as were individuals with disconnected medical record history (either because these patients moved to another country or lacked a stable residence) and minors (age <16 years).

The outcome clinical deterioration was defined as unplanned ICU transfer or in-hospital mortality within the so-called prediction window. Unplanned ICU transfer was defined as any acute admission to an ICU (filtering by the codes NABE and NABB from the Sundhedsvæsenets Klassifikations System (SKS) - the Danish Health Service Classification System) within 24 hours after discharge from a non-ICU ward.

Model

The model architecture (Figure 2) was designed as a scalable network that can be adapted to different data modalities by using one sub-model per data domain; each sub-model consists of an embedding layer (which transforms the categorical variables into vectors), a recurrent neural network (which learns from the sequence of embedding vectors) and a pooling layer (which reduces the vectors’ dimensionality).

The network uses tokens, i.e. sequences of characters grouped together based on their semantics, as input. We chose the entity embedding approach to exploit the heterogeneity and sparseness of the input data, thus mapping the tokens constructed from the categorical features to embeddings in a Euclidean space (i.e. the embedding space) (19). For each sub-model, the vocabulary size $V$ (i.e. the number of unique tokens for the specific data source) was directly used to calculate the size of the embedding representation according to $\text{EmbeddingSize} = 6 \cdot \alpha \cdot \sqrt{V}$, where the embedding coefficient $\alpha$ was estimated by hyperparameter search (14).
Figure 2: Structure of the deep learning model, exemplified by two admissions.

For each sample (i.e., admission at a specific time point) a 2D tensor (matrix) comprises the sequence of embedded tokens until the time of assessment. Each tensor was given as input to a recurrent block (GRU or LSTM, part of the hyperparameters search) and the time dimension was pooled through an attention pooling layer. The flattened tensors were then concatenated and fed to a linear layer with a standard logistic activation function.

We explored two types of recurrent neural networks, Gated Recurrent Unit (GRU) (20) and Long Short-Term Memory (LSTM) networks (21). Rather than max or average pooling – which retains neither positional nor intensity information – we used an attention-based pooling (22) which employs a weighted mechanism to retain the most relevant parts of a sequence. The attention-based pooling layer was used to aggregate the temporal dimension of the recurrent layer output, which was then concatenated across the different data sources. The final layer was a linear (= dense) layer with a standard logistic activation function hence mapping the output into the interval [0, 1] to yield valid predicted probabilities. The features without temporal components (age, number of previous hospital admissions and sex) were fed into the model by concatenating them to the output of the pooling layer.

Data sources and processing

The input data comprised medical disease history, biochemical measurements, clinical notes and demographics (age, sex and number of previous admissions). The medical disease history prior to the start of the EHR data was also used (as well as after to handle false negative outcomes not covered after the EHR data period). A schematic of the time spans for the different data modalities is depicted in Figure S1. The disease history was extracted from the Danish National Patient Registry (DNPR) for the period between 1977 and 2018 (23). DNPR is nation-wide registry which covers essentially all the hospital encounters in Denmark. The disease codes use Danish adaptations from SKS of the International Classification of Diseases (ICD) version 8 up to 1993, and ICD-
10 from 1994 and onwards. Only diagnosis codes recorded prior to the admission date were included in the predictive schemes.

Biochemical values were extracted from the Clinical Laboratory Information System (LABKA) and the Clinical Chemistry Laboratory System (BCC) databases (24) for the period overlapping the hospital admissions (2011-2016). These databases collect all the biochemical tests performed by the Danish hospital laboratories in the Capital Region and Region Zealand, respectively. The biochemistry tokens were constructed using the name of the biochemical component, the specimen (blood, plasma etc.), the unit and the quantile of the value of the measurement (to yield a vocabulary with a reasonable size); for example, HEMOGLOBIN_B_mmol/L@8-8.5. The quantile binning was included as a hyperparameter.

The clinical notes were extracted from the EHR data for the admissions from 2011 to 2016. The free text required some preprocessing before tokenization, such as removing punctuation, names, stop words (25), negations and signatures of the clinicians. Due to the large number of terms coming from the medical notes, the embedding of this specific data source was trained separately on the full corpus using fastText (26), to reduce the number of parameters to update during training.

Table 1. Summary statistics for the Development and Test sets. Statistics are calculated at the admission level.

<table>
<thead>
<tr>
<th></th>
<th>Development Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>682041</td>
<td>170579</td>
</tr>
<tr>
<td>Number of Admissions</td>
<td>1791705</td>
<td>450144</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Age at prediction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Discharge</td>
<td>60 [40-73]</td>
<td>60 [40-73]</td>
</tr>
<tr>
<td>- ICU Transfer</td>
<td>67 [56-75]</td>
<td>67 [56-75]</td>
</tr>
<tr>
<td>- In-hospital Death</td>
<td>78 [68-86]</td>
<td>78 [69-86]</td>
</tr>
<tr>
<td><strong>Sex:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Male</td>
<td>798185 (45%)</td>
<td>199732 (44%)</td>
</tr>
<tr>
<td>- Female</td>
<td>993520 (55%)</td>
<td>250412 (56%)</td>
</tr>
<tr>
<td><strong>Outcome:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Discharge</td>
<td>1752329 (97.8%)</td>
<td>440221 (97.8%)</td>
</tr>
<tr>
<td>- ICU Transfer</td>
<td>13129 (0.7%)</td>
<td>3305 (0.7%)</td>
</tr>
<tr>
<td>- In-hospital Death</td>
<td>26247 (1.5%)</td>
<td>6618 (1.5%)</td>
</tr>
<tr>
<td><strong>Number of previous Admissions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Discharge</td>
<td>1 [0-4]</td>
<td>1 [0-4]</td>
</tr>
<tr>
<td>- ICU Transfer</td>
<td>2 [0-6]</td>
<td>2 [0-6]</td>
</tr>
<tr>
<td>- In-hospital Death</td>
<td>3 [1-7]</td>
<td>3 [1-6]</td>
</tr>
<tr>
<td><strong>Length of stay before outcome (Hours):</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- ICU Transfer</td>
<td>35 [8-121]</td>
<td>35 [8-132]</td>
</tr>
<tr>
<td>- In-hospital Death</td>
<td>126 [43-285]</td>
<td>122 [43-279]</td>
</tr>
<tr>
<td><strong>Type ICU Transfer:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Surgical</td>
<td>2358 (0.18%)</td>
<td>639 (0.20%)</td>
</tr>
<tr>
<td>- Medical</td>
<td>10771 (0.82%)</td>
<td>2666 (0.80%)</td>
</tr>
</tbody>
</table>

**Training and Evaluation**

We randomly split the dataset into a development set (80%) for model creation and an independent holdout test set (20%) for model evaluation. The split of the dataset was done at the patient level and assigned admissions of the same patient to the same set to avoid leaking information between the sets (27). The development set was further split into a training set (80%, 64% of total) and validation set (20%, 16% of total) to counter overfitting and to calibrate the model before testing it on the holdout set.

Three submodels were trained separately to find the best submodel-specific hyperparameters; we used Optuna’s multivariate TPEsampler, based on the Three-structured Parzen Estimator (TPE) algorithm to search the hyperparameter spaces (28,29). While searching the hyperparameter space, we fixed the prediction window at 24 hours and the assessment rate at 12 hours, to facilitate the comparison across the
different data types. Table S1 shows the hyperparameters explored for each submodel; Figures S2-S4 illustrate the hyperparameter searches. For each experiment, the loss on the validation set was used as training metric to select the model at the best epoch. The performance metric used to select the best model for each search was the area under the precision-recall curve (AUPRC). Although the area under the receiver operating curve (AUROC) score is a more common metric for classification tasks, in this case it would not be sufficient to appreciate the real ability of the model to discriminate between the two classes due to their considerable imbalance (30). AUPRC on the other hand is much more robust to class imbalances. We used 200 bootstraps samples (31) to construct the 95% confidence interval for the metrics shown in figures and tables.

Using the best submodel architectures, we trained and evaluated an ensemble model for different prediction windows (1, 2, 7, 14 days) and assessment rates (every 6, 12 and 24 hours). Finally, we applied post-hoc isotonic calibration to align the final predicted risk with the actual outcome incidence, using the data in the validation set (32). The model re-calibration was achieved by fitting an isotonic regression using the output of the model prior to calibration as regressor and the actual label of the samples as response variable. We used the validation set to fit the isotonic regression and kept the test set untouched. The fitted isotonic model was used to adjust the output of the uncalibrated model to get calibrated predictions.

All the results reported are generated using the re-calibrated model on the test set. To control for biases driven by age or sex, performances were also evaluated at each time of assessment for the different subgroups.

**Interpretation**

The impact of the different tokens on the model outcome was calculated using the GradientShap algorithm (33) from the Captum library (34). Given an input feature of a single risk assessment, its Shap value is correlated to how much (and in what direction) that feature pulls the individual-level prediction away from the population-level mean risk. Importantly, Shap values do not represent the effect of a single feature on the model outcome but rather the effect of that feature in the context of a coalition of features. Shap values were calculated for the best model after isotonic re-calibration.

**Results**

The model was trained on 682,041 unique patients and 1,791,705 admissions and evaluated on 170,579 patients and 450,144 admissions as described in Methods. In both parts, 1.5% of the admissions resulted in in-hospital mortality and 0.7% in ICU transfer. 2,583 tokens from the ICD code data type (medical history), 2,421 tokens from the biochemical measurements and 403,869 tokens from the medical notes were used as
input. We investigated the inclusion of each data type separately as well as jointly in the same model.

When we explored how the prediction window and assessment rate affect the performances, the most performant model was the one trained on all the data sources using a prediction window of 14 days and an assessment rate of 6 hours with an AUROC of 0.904 [0.903-0.905] and AUPRC 0.285 [0.283-0.287] over all the predictions (Table 2). This model was after isotonic recalibration well-calibrated (Figure S5) with a calibration slope of 0.964 [0.951, 0.98], an intercept of 0.002 [0.001-0.003] and an upper bound risk of 79%.

The performances were similar across sexes, with AUPRCs of 0.288 [0.285-0.290] and 0.280 [0.277-0.283] for males and females, respectively. The performances for the different age groups varied more, with a trend of AUPRC increasing with age (age 16-37: 0.168 [0.157-0.180], age 37-58: 0.252 [0.246-0.257], age 58-79: 0.277 [0.274-0.280], age 79-100: 0.310 [0.307-0.313]).

We observed an increased AUPRC for risk estimates further into the hospital stay with a peak at 7 days into the admission (Figure S6-S7).

Table 2. Performance on the test set for models using all data types.

<table>
<thead>
<tr>
<th>Frequency of assessment</th>
<th>Prediction window</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Calibration</th>
</tr>
</thead>
<tbody>
<tr>
<td>6h</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d</td>
<td>1d</td>
<td>0.925 [0.923-0.926]</td>
<td>0.152 [0.149-0.155]</td>
<td>1.042 [1.02-1.064]</td>
</tr>
<tr>
<td>2d</td>
<td>0.922 [0.921-0.923]</td>
<td>0.180 [0.177-0.184]</td>
<td>0.988 [0.966-1.007]</td>
<td>0.002 [0.001-0.003]</td>
</tr>
<tr>
<td>7d</td>
<td>0.913 [0.912-0.914]</td>
<td>0.269 [0.267-0.271]</td>
<td>0.996 [0.985-1.008]</td>
<td>0.002 [0.001-0.003]</td>
</tr>
<tr>
<td>14d</td>
<td>0.904 [0.903-0.904]</td>
<td>0.285 [0.283-0.286]</td>
<td>0.964 [0.951-0.98]</td>
<td>0.004 [0.003-0.005]</td>
</tr>
<tr>
<td>12h</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1d</td>
<td>0.916 [0.916-0.920]</td>
<td>0.129 [0.125-0.134]</td>
<td>0.998 [0.96-1.045]</td>
<td>0.001 [0.0-0.002]</td>
</tr>
<tr>
<td>2d</td>
<td>0.916 [0.916-0.917]</td>
<td>0.187 [0.183-0.191]</td>
<td>0.996 [0.977-1.021]</td>
<td>0.002 [0.0-0.003]</td>
</tr>
<tr>
<td>7d</td>
<td>0.898 [0.897-0.899]</td>
<td>0.233 [0.231-0.236]</td>
<td>1.045 [1.026-1.067]</td>
<td>-0.001 [-0.002-0.0]</td>
</tr>
</tbody>
</table>
Figure 3. Performance of the model for prediction of unplanned ICU transfer or in-hospital death. Panel A: precision recall curves for the six models. Panel B: receiver operating characteristic (ROC) curves show the values of true positive rate (= recall = sensitivity) and false positive rate (= 1-specificity) at different thresholds for the six models. Model A: age, sex, number of admissions. Model B: model A + medical history data. Model C: model A + biochemical data. Model D: model A + clinical notes. Models E and F: model A + medical history data, biochemical data and clinical notes. Models A–E use a 24-hour prediction window and a 12-hour assessment rate. Model F uses a 14-day prediction window and a 6-hour assessment rate.

Among the models trained separately using a 24-hour prediction window and 12 hours assessment rate, the most performant was the one trained on the clinical notes with an
AUPRC of 0.125 [0.120-0.130]; this value improved to AUPRC of 0.131 [0.127-0.135] when the model was trained on all the data using the same prediction window and assessment rate.

The best encoding of disease diagnoses was rolling diagnoses up to the third ICD level (e.g. C341M to C34) and a padding size of 9 (Figure S2). The optimal quantile resolution for the biochemical data was deciles (i.e. 10 bins) and a padding size of 28 was the optimal number of lab values to include before time of prediction (Figure S3). The optimal padding size for medical notes was 299, reflecting the larger amount of information usually held by the clinical notes (Figure S4). The complete list of the optimal parameters for each search can be found in the supplementary Table S1.

Feature importance

The importance of the tokens from the different vocabularies can be sorted according to their attributed Shap values (Figure 4). Higher age and male sex were associated with elevated risk of deterioration. The opposite was the case for the number of previous admissions: a higher number of previous admissions was associated with lower risk of deterioration (Figure 4.A-C).

The diagnosis tokens associated with clinical deterioration were often those of acute illnesses such as respiratory failure, neoplasm and pneumonia (Figure 4.D). In contrast, diagnosis codes related to pregnancy, chronic conditions and orthopedics were associated with lower risk.

Low levels of albumin, lymphocytes and sodium were associated with elevated risk of clinical deterioration; the same were high levels of C-reactive protein, leukocytes, sodium, lactate dehydrogenase, potassium and carbamide (Figure 4.E). In contrast, normal levels of leucocytes, hemoglobin, C-reactive protein, sodium and potassium were all associated with low risk of deterioration.

The tokens from clinical notes most strongly associated with clinical deterioration were severely, intensive, respiratory and chronic. Pain, leukocytes, control, normal, home, discharge(d) and on the other hand, were all associated with low risk (Figure 4.F).
Figure 4: Contribution of the different tokens to the outcome of clinical deterioration. Shap values were estimated using the best model (model using age, sex, number of admissions, medical history data, biochemical data and clinical notes with a 14-day prediction window and 6-hours assessment rate). Panels A-C: the distribution of Shap values for each feature without temporal components, with the color scale representing the feature value. Panels D-F: the distribution of Shap values for top-15 tokens (in up- and downward directions, respectively). Overlaid boxes-and-whiskers show medians, quartiles and 1.5 x quartiles.
The main aim of this study was to explore whether data types registered routinely in general departments are predictive of clinical deterioration, ultimately to assess if these suffice for this task. A solution based on data collected routinely might circumvent a common weakness of EWS, i.e. they depend on data that require clinical engagement (e.g., vital signs). Consequently, offering a viable alternative for risk stratification with minimum additional manual data collection effort is preferable.

Exploiting the combined power of entity embedding of tokens from electronic health records and the ability of recurrent neural networks to learn temporal patterns from such data, we built a performant (AUROC and AUPRC up to 0.90 and 0.28, respectively) and well-calibrated deep learning model for predicting the risk of clinical deterioration. Specifically, leveraging medical history data (up to 40 years) from a national register along with in-hospital biochemical data and clinical notes, we trained the model dynamically, meaning that the same model can handle different time points for the same admission.

Direct comparison of our model to the performances of $h$ is not possible, since the vital signs used for the calculation of such scores are not collected in our EHR dataset. Comparison of the performance of our model to other work is also non-trivial: AUROC is the most common metric for risk stratification and usually the one used to compare performance across studies. Nevertheless, it is unsuited for imbalanced prediction problems (such as the one defined here) because it disregards the prevalence of the outcome of interest (30). Although AUPRC does account for prevalence, it is not always reported in the classic studies on EWS scores. Moreover, direct comparison of the AUPRCs requires equal (or at least similar) prevalence of the outcome across studies; this is problematic because the prevalence tends to vary between cohorts, and, more importantly, so do the criteria used to define it. For example, Watkinson et al. (35) defined the outcome as the composite of ICU transfer, in-hospital death and cardiac arrest with a prediction window of 24 hours (AUROC of 0.868 [0.864–0.872]). Dziadzko et al. (36) defined the outcome as in-hospital death or respiratory failure, with a prediction window of 48 hours (AUROC 0.87 [0.85–0.88] and 0.90 [0.84–0.95] in 2013 and 2017, respectively). Malycha et al. (37) defined the outcome as in-hospital death and ICU transfer within 24 h following assessment from patients admitted longer than 24 hours (AUROC 0.823 [0.819–0.824]). Cho et al. (10) instead defined the outcome as cardiac arrest and unexpected ICU admission, occurring within 0.5-25 hours from the assessment (AUROC 0.865 [CI not reported]). The performance difference of models A-E (Figure3) is driven by information learnt from different data types. The performance does not increase linearly with the addition of new data types, suggesting substantial overlap in the latent information of diagnosis codes,
biochemical data and clinical notes. This corresponds well to what one would expect, e.g. a clinical note may very well cite lab values (and likely those of greatest clinical interest) and summarize anamnestic information such as comorbidities also registered in the diagnosis code data.

On the other hand, the change in performance from model E to model F (AUPRC from 0.131 [0.127-0.137] to 0.287 [0.284-0.289]) is likely driven by the different incidence of the outcome when using longer prediction windows, essentially making the prediction task easier because there are more examples to learn from. Also, it is much more difficult to discriminate between patients having a severe outcome in the subsequent 24 or 48 hours, because drivers of short-term mortality/ICU transfer depend not only on the physiological status of the patient but also on many factors beyond what is captured in clinical data, e.g. the coordination of the resources within and between hospital departments. The assessment rates employed are consistent with how often new observations and information are recorded in patient files: new clinical notes and biochemical measurements are usually recorded a couple of times every day and normally at least once. The change in AUPRC and AUROC during the admissions reflects the increasing availability of new records when the patients proceed further into the hospitalization (Figure S6-S7), and that patients tend to deteriorate either early or late in the admission and less in between.

We decided to keep the outcome definition broad rather than predicting specific severe conditions such as sepsis or organ failure, notoriously difficult to operationalize for prediction tasks (38). In contrast, using a general definition of clinical deterioration allowed us to keep the feature space more generic and less dependent on specific illnesses. Indeed, our model seeks not to advice on interventions but to flag patients at risk of (more of less imminent) clinical deterioration so health care staff can intervene in a manner appropriate for the patient in question, hopefully translating into improved prognosis for that patient.

An ensemble structure of the network was preferred over a structure in which the different data types contributed to the creation of the same embedding space. Separate submodels (embedding + recurrent linear layer + pooling) allowed us to tune their architectures; optimal padding sizes and embedding coefficients, for example, differ for diagnosis codes and biochemical data. Its ensemble nature also renders the model scalable for incorporating new data types/domains.

To obtain a reasonably well-defined training cohort, we only included in-patient admissions to general departments, excluding out-patients and acute admissions to the emergency department. The former groups were excluded because they are more unlikely to experience the outcome, the latter because the rapid course of events and
acute physiology recorded in the emergency department would necessitate a different experimental setup.

Interpretation

Although ranking the features that drive the predicted risks up or down is useful as a sanity check of the signals picked up by the model, these estimates are not causal. Indeed, the attribution of each token is determined by its context therefore the same token could be associated with both elevated and diminished risk depending on the other tokens with which it co-occurs. For example, a lab value outside the physiological range (e.g., low creatinine), which would normally drive the risk up, may not affect the predicted risk when co-occurring diagnosis codes counter its contribution (e.g., pregnancy codes).

This is more evident for the free text, where the semantics of a word is always dependent on its context. Overall, the feature attribution does not contain any counterintuitive explanation and the few seemingly questionable interpretations have plausible explanations. For example, we find low eosinophil counts among the lab values associated with elevated risk of clinical deterioration. Clinically, this seems counterintuitive since low values of eosinophils represent the standard and high eosinophil counts are indicative of infections (especially parasitic) and allergic disorders. Eosinophils, however, are part of standard panels for blood differential counts (a count for the different types of white blood cells) and, as such, the very presence of (any) eosinophil count probably reflects a clinical suspicion of infection which caused additional analysis to investigate on the infectious agent. Interpretation of tokens from the clinical notes also provides some examples that at first seem odd but probably do have some contextual bearing. For example, the tokens *wife* (husbru) and *relatives* (pårørende) are strongly associated with clinical deterioration. This is likely because doctors document in the patient file when relatives have been informed or consulted, and this may well be on poor prognoses or even no-resuscitation orders in which case clinical deterioration is almost certain to ensue.

Numerical features like the number of previous admissions should also be contextualized. While higher age is indeed associated with clinical deterioration, having a lot of admissions prior to the one of the assessments does not. This seeming paradoxical result can be probably explained by the ability of the model to integrate data from different EHR domains to recognize patients with chronic conditions who will have more frequent hospital visits but perhaps be less likely to suddenly fall critically ill.

Strengths
This study has some important strengths. First, it is one of the largest of its kind, with a total of 852,620 patients and 2,241,849 admissions taking place over 6 years. Second, we provided a dynamic risk assessment, showing how prediction performance changes over the course of an admission for different prediction windows. This supersedes early warning scores, based on static metrics that do not take sequential information into account but use just a snapshot of the patient’s current status. Third, thanks to the model’s architecture, adding new data sources is relatively easy. New features can be added and removed from the model, adapting the tool to the available resources. However, as expected, the model performs best with as many data as possible included. Finally, while classic EWS depend on complete data to be calculated, missing data in this setup is not a problem since the model has been designed to handle it thanks to the entity embeddings.

Limitations

Like any study this has limitations. First, although we tried to define the outcome robustly, there are some pitfalls to consider for ICU transfer and in-hospital mortality. Patients in very severe conditions may still be discharged by the hospital if the latter is not able to provide any support to the patient. These patients will probably experience clinical deterioration within the prediction window but not within their hospital stay, hence they are labelled as negative but potentially still flagged by the model as high-risk patients; this could inflate the number of false positive patients detected by the model. Second, unplanned ICU transfer was captured only for the patients who were admitted in one of the ICUs in our catchment area. It cannot be excluded that some patients admitted to a general department of one of the hospitals in our catchment area are transferred to ICUs of other Danish regions than included here, even if it is uncommon. Third, model performance may improve with additional data such as genetic data, vital signs and other biomarkers. Thanks to ensemble nature, adding such data in setting where they are available is simple. Finally, the model was trained on data from the Danish healthcare system and the model (due to the entity embeddings) would need to be trained anew if deployed in other geographical or healthcare systems.

Conclusion

Combining entity embeddings and recurrent neural networks we built a highly performant model for flagging patients at risk of clinical deterioration. The model was developed and evaluated in a controlled in-silicon setting. Although the data used were collected prospectively, a proper prospective evaluation would be needed to establish whether its deployment can produce real-world benefits to patients on hard endpoints.
References


29. Falkner S, Klein A, Hutter F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. 2018;


Data and software statement

The software used in the study is based on python v3.8 and pytorch. The software is available online at https://gitfront.io/r/daplaci/uA6jdq4FtTJQ/ClinicalDeterioration/. The authors do not have permission to share data; access to the original data can be obtained from the Danish health authorities.

Funding and permission statement

This study was approved by the Danish Patient Safety Authority (3-3013-1731 and 3–3013–1723), the Danish Data Protection Agency (DT SUND 2016–48, 2016–50, 2017–57 and UCPH 514-0255/18-3000:) and the Danish Health Data Authority (FSEID 00003092, FSEID 00003724, FSEID 00004758 and FSEID 00005191). The Novo Nordisk Foundation (grants NNF17OC0027594 and NNF14CC0001) and the Danish Innovation Found (5184-00102B) supported this study.