EHR-ML: A generalisable pipeline for reproducible clinical outcomes using electronic health records

Yashpal Ramakrishnaiah\textsuperscript{a}, Nenad Macesic\textsuperscript{a,c}, Geoffrey I. Webb\textsuperscript{a,c}, Anton Y. Peleg\textsuperscript{a,c} and Sonika Tyagi\textsuperscript{a,b,*}

\textsuperscript{a}Department of Infectious Diseases, The Alfred Hospital and Central Clinical School, Monash University, Melbourne 3000, VIC, Australia
\textsuperscript{b}School of Computing Technologies, RMIT University, Melbourne 3000, VIC, Australia
\textsuperscript{c}Centre to Impact AMR, Monash University, Melbourne 3000, VIC, Australia

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Digital Health
Electronic Health Records
EHR
Clinical Outcome Prediction
Machine Learning

\textbf{ABSTRACT}

The healthcare landscape is experiencing a transformation with the integration of Artificial Intelligence (AI) into traditional analytic workflows. However, this advancement encounters challenges due to variations in clinical practices, resulting in a crisis of generalisability. Addressing this issue, our proposed solution, EHR-ML, offers an open-source pipeline designed to empower researchers and clinicians. By leveraging institutional Electronic Health Record (EHR) data, EHR-ML facilitates predictive modelling, enabling the generation of clinical insights. EHR-ML stands out for its comprehensive analysis suite, guiding researchers through optimal study design, and its built-in flexibility allowing for construction of robust, customisable models. Notably, EHR-ML integrates a dedicated two-layered ensemble model utilising feature representation learning. Additionally, it includes a feature engineering mechanism to handle intricate temporal signals from physiological measurements. By seamlessly integrating with our quality assurance pipelines, this utility leverages its data standardization and anomaly handling capabilities.

Benchmarking analyses demonstrate EHR-ML’s efficacy, particularly in predicting outcomes like inpatient mortality and the Intensive Care Unit (ICU) Length of Stay (LOS). Models built with EHR-ML outperformed conventional methods, showcasing its generalisability and versatility even in challenging scenarios such as high class-imbalance.

We believe EHR-ML is a critical step towards democratising predictive modelling in healthcare, enabling rapid hypothesis testing and facilitating the generation of biomedical knowledge. Widespread adoption of tools like EHR-ML will unlock the true potential of AI in healthcare, ultimately leading to improved patient care.
List of Figures

1 Elements of the EHR-ML pipeline .................................................. 5
2 Class discrimination analysis ............................................................ 7
3 EHR-ML vs. standalone models ......................................................... 8
4 EHR-ML vs. traditional scoring system .............................................. 9
5 Time-window analysis ................................................................. 11
6 Sample-size analysis ............................................................ 12
7 LOS prediction analysis .......................................................... 14
8 Web-portal ............................................................... 15
S1 Combined QC plots .......................................................... 19
S2 Standardisation analysis .................................................... 20
S3 Class-Ratio analysis .......................................................... 21

List of Tables

1 Patient cohort details ............................................................ 4
2 EHR-ML vs. traditional scoring system ........................................ 10
3 Time-window analysis .......................................................... 11
S1 EHR-ML vs. constituent models ............................................... 20
S2 EHR-ML vs. out of the box models ........................................... 21
S3 Standardisation Analysis ...................................................... 21
S4 Sample-Size Analysis .......................................................... 22
S5 Class-Ratio Analysis .......................................................... 22
S6 LOS Prediction Performance ................................................... 22
S7 LOS Prediction Time Window Analysis ...................................... 23
1. Introduction

Artificial Intelligence (AI) is rapidly transforming healthcare. AI-enabled predictive modelling of patient outcomes [1, 2, 3, 4], can support early disease detection, more targeted therapies, and improved risk stratification. Electronic Health Records (EHRs) paves the way for powerful predictive modelling [5, 6, 7, 8, 9]. However, concern around non-generalizability of research outcomes is a recurring theme in EHR-based predictive modelling [10, 11]. Models trained in one centre may not be applicable in other settings due to differences in clinical practices and data characteristics [12, 13]. Hence, to unlock the true potential of AI in healthcare, site-specific modelling is essential, leveraging localised data within the institutional EHRs.

While numerous studies leverage EHRs in this way, they are impeded by the absence of well-established pre-processing techniques [14, 15, 16], modelling tools [17, 18, 19], and protocols [15, 18, 20]. This fragmented landscape leads to irreproducible results, inconsistent outcomes, needless complexity, and error-prone ad-hoc processes [17, 21, 22]. To establish standardization in this process, various frameworks[23, 24] and guidelines [25, 26, 27, 28, 29] are proposed for conducting and reporting such studies. Additionally, generic toolkits for predictive modeling have been developed to accelerate AI adoption in healthcare[30, 31, 32, 6]. However, these toolkits often suffer from some significant limitations. Restricting the data source to a single rigid input format [31, 32], hinders their applicability, while non-interpretable neural networks raise implementation concerns in healthcare settings [30, 6]. Furthermore, existing solutions lack end-to-end automation, covering data sourcing to model building, posing a roadblock for widespread localized deployments. Moreover, limitations such as restrictions in selecting a target, manual feature selection processes, and non-robust performance metrics compromise their utility. Additionally, generic out-of-the-box models offered by these frameworks often underperform compared to models customized with domain-specific nuances. Furthermore, all these approaches require extensive configuration regarding study design choices, including data window selection, target viability assessment, sample size optimization, and pre-processing steps.

Developing an effective study design to construct robust predictive models from EHR data is challenging. One major hurdle involves determining the appropriate time window for data collection, ranging from early-stage predictions within 12-48 hours [33, 34, 35] to encompassing the entire duration of admission [36], or retrospective analysis [37, 38]. Additionally, the scarcity of examples within certain data classes can impede the modeling of specific clinical targets [39, 40]. For example, while predicting outcomes for hospital stays exceeding seven days may be feasible, it becomes impractical for stays surpassing 30 days due to insufficient data samples. Moreover, healthcare data is often constrained by privacy concerns and the high cost [41, 42]. Understanding the minimum data requirement for reliable modeling is crucial. Another challenge stems from the class imbalance, resulting in skewed outcome distributions, known as representation bias. This imbalance, where the class of interest may have fewer instances (e.g., longer hospital stays or patient deaths), can impede model training and necessitate careful data balancing strategies [43, 44]. Additionally, the varied scales of clinical attributes pose obstacles for machine learning models. Attributes like temperature, measured in degrees Celsius, span a range of 35 to 40, while heart rate, measured in beats per minute, typically falls between 60 and 100. Each attribute operates on distinct units and scales. To mitigate this issue, data harmonization and scaling techniques are employed to standardize all variables, benefiting certain modeling approaches. However, the decision to standardize data adds complexity to the modeling process [45]. Furthermore, EHR measurements are recorded at irregular intervals, making it challenging to design data transformation methods that retain both magnitude and temporal dynamics for machine learning models. Making informed choices about these parameters is crucial for successful modeling but currently relies heavily on empirical guesswork due to a lack of appropriate tool sets for exploring optimal parameter values [14, 46].

In response to these challenges, we introduce EHR-ML, a comprehensive package for the predictive modelling of clinical outcomes using the EHR. This package ensures that every stage, from data acquisition to model construction, adheres to a clearly defined, domain-specific, data-centric, and reproducible protocol, enforcing optimal practices. To demonstrate the effectiveness of EHR-ML, we employed it to forecast clinical outcomes within a selected cohort of patients diagnosed with sepsis, a condition demanding time-sensitive intervention and treatment. Within this cohort, we conducted predictive modeling for two key clinical outcomes: patient mortality and Intensive Care Unit (ICU) Length of Stay (LOS) [47, 48, 49].

To demonstrate its utility, we performed predictive analysis for mortality risk, a pivotal area of focus [50, 51, 52, 53], owing to its significance in patient care and resource management. Prompt identification of mortality risk equips healthcare professionals with crucial insights for patient triage, treatment strategies, efficient resource distribution, and a comprehensive comprehension of the factors affecting patient outcomes. Traditionally, severity scoring systems
like SOFA [54], qSOFA [55], SAPS II [56], and APACHE [57] have played a crucial role in this regard. However, leveraging localised data, machine learning has emerged as a promising alternative outperforming the one-size-fits-all conventional scoring schemes. The next prediction task deals with forecasting another clinical outcome, the ICU-LOS. The LOS prediction is usually approached as a binary prediction problem [58, 59, 60], although some studies adopt continuous regression modelling methods [61, 62]. Notably, the prediction of LOS poses increased complexity compared to mortality prediction [63], as patient distinctions are less pronounced between the classes in this case. By modelling these two diverse outcomes, we aim to showcase the utility, versatility, and simplicity of EHR-ML.

In essence, the goal of EHR-ML is to bridge existing gaps by providing a user-friendly, open-source platform that enables clinicians and researchers to effectively leverage the potential of institutional healthcare data.

## 2. Methods

### 2.1. Data

The development and assessment of EHR-ML utilised two openly accessible, EHR datasets: Medical Information Mart for Intensive Care (MIMIC) IV [64] and eICU [65]. MIMIC IV consists of the data from a single large tertiary teaching hospital, while eICU includes data from a network of critical care units. From these two sources, three distinct cohorts focused on sepsis patients were extracted (refer to Table M1). From each cohort, vital signs and laboratory measurements were extracted for analysis. Next, both datasets were standardised to the OMOP-CDM schema [66, 67] and mapped to standard SNOMED vocabulary [68]. This was achieved through the standardisation module within our previously published EHR-QC tool [69]. Standardisation facilitates consistent data interpretability, leveraging existing tools, interoperability of developed tools, standardised pre-processing, and deduplication of the data.

### 2.2. Machine learning

For machine learning, the chosen data cohort underwent rigorous preprocessing with the EHR-QC quality assurance module (see supplementary figure S1). Initially, vital signs and laboratory measurements for the patient cohort were extracted. Subsequently, a subset of measurements recorded for a high proportion of patients (over 80% in this work) was retained after coverage analysis. This analysis report lists attributes by prevalence in the EHR, aiding in determining a suitable threshold. The measurements that are rare in the EHR lack sufficient numbers for effective modelling, hence removed from the analysis. Essentially, this process retains the most widely recorded measurements (e.g., temperature and heart rate) whereas, rarely occurring measurements (e.g., Intracranial pressure) are removed.

The subsequent phase entails formatting the chosen data to render it compatible with machine learning tasks. EHR data usually contains high-resolution vital signs and laboratory measurements with many recordings within a short period. While this information is valuable in modelling, directly utilising such data can be a challenge with the feature-based machine learning algorithms that necessitate the data to be structured as a two-dimensional matrix. This is addressed by employing a multi-faceted aggregation approach [70] using five functions: minimum, maximum, first value, last value, and mean. These aggregations are applied to vitals and lab measurements, producing 10 feature
Figure 1: A) This panel depicts a typical patient timeline with key timestamps: hospital admission (Day 1), ICU admission (Day A), and discharge (Day D). The, WB and WA parameters define the data window used for obtaining the data from A - WB to A + WA. The relevant clinical outcomes, LOS (LOS = discharge time D - admission time A), and mortality indicating the patient status at discharge, are also shown. B) This panel illustrates the multi-level ensemble architecture utilised by EHR-ML for machine learning. Level 1 consists of four individual models – XGB, LR, LGBM, and MLP – each trained on vital signs and laboratory measurements formed using five different aggregation functions (see methods). Level 2 then combines the learned feature representations to build the final ensemble model. C) This panel depicts datasets with varying positive and negative class ratios, constructed for analysing the impact of class imbalance on model performance. D) This panel showcases datasets of different sizes obtained through sampling, used for analysing the influence of sample size on model performance. For this analysis, both the classes are randomly sampled without any stratification. E) This panel illustrates multiple patient timelines with different data windows generated by varying the WB and WA parameters, used for analysing the influence of data collection window on model performance. F) This panel compares three data representations: raw data without any transformations, data scaled using Standard Scaling (mean of 0 and standard deviation of 1), and data scaled using MinMax Scaling (minimum of 0 and maximum of 1). These variations enable analysis of the impact of data standardisation on model performance.
The quality-assured data encompass all the available data from the patient timeline, whereas there is a need to restrict the data within a certain time frame or time window for modelling purposes. For instance, while using retrospective data for predicting the patient’s risk of mortality post 2 days of admission time, all the futuristic data at the time of prediction (48 hours) needs to be purged.

A patient enters the hospital, marking the start of the analysed timeline denoted as Day 1. If transferred to ICU, the day of ICU admission is marked as Day A and day of discharge marked as Day D. The time-window usually centres around ICU admission (Day A), but can be anchored to any other relevant time point such as the time of positive microbiology culture result. We determine the starting point of our data window with a parameter called "window before" (WB), indicating how much historical data we capture before the anchor time. Conversely, the parameter "window after" (WA) determines the end point of the data window, representing the duration of data considered after the anchor time. The data window thus spans from A - WB to A + WA, covering a total duration of WB + WA. Selecting appropriate values for WA and WB depends on the study design. Increasing WB includes more historical data for modeling, while increasing WA extends the data collection period post anchor point.

The next step is to calculate the target variable for the prediction. In this work, we specifically looked at two clinical outcomes, discharge mortality and ICU LOS. The mortality on discharge is a binary outcome indicating whether the patient survived during the ICU stay. The LOS is calculated as the difference between discharge and admission times (D - A). In this work, to facilitate binary classification, we framed LOS prediction as the probability of a patient exceeding a set hospital stay duration threshold. Specifically, we set the threshold as 7 and 14 days. This approach allows flexible adaptation of the target variable, enabling predictions for any threshold duration. In fact, EHR-ML provides the flexibility to model any clinical outcome of interest that is either directly present or derived from the EHR.

Internally, EHR-ML leverages a two-layer ensemble architecture for robust clinical outcome prediction (Figure 1-B). At level 1, four distinct models - XGBoost (XGB), Logistic Regression (LR), Light Gradient Boosted Machines (LGBM), and Multilayer Perceptron (MLP) – are individually trained on 10 feature sets. As each feature set is used by all four models, a total of 40 models are built at level 1. The predictions from Level 1 models are used as input to the second layer (Level 2). Using the outputs from level 1 models as inputs for level 2 helps to combine the strengths of individual models. This intermediate data, shown in figure 2 to have better discriminative power, serves as input for the final XGBoost prediction. In this architecture, obtaining the important features of the final model will help in understanding factors affecting the clinical outcome under consideration making it interpretable. Hyperparameters for each model at both layers are individually optimised specifically for the corresponding data input, ensuring optimal performance.

2.3. Analytic evaluation and validation

We evaluated and compared the performance of different models using a comprehensive set of metrics and visualisations. For individual model assessment, standard metrics like accuracy, balanced accuracy, average precision, F1 score, area under the receiver operating characteristic curve (AUROC), and Matthews correlation coefficient-F1 (MCCF1) were computed. To facilitate comparisons between different models, we reported True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), along with a heatmap of the confusion matrix to visualise them. All metrics were calculated over 5-fold cross-validation, and mean values were reported.

We compared the performance of our ensemble model by benchmarking it against the traditional SAPS II scoring system [56]. SAPS II utilises 12 physiological and 3 disease-related variables, assigning individual scores for variable ranges and aggregating them to calculate a final score. This score can then be used to estimate the risk of mortality through a predefined formula. In addition, we built another model, EHR-ML-SAPS-II, utilising EHR-ML ensemble architecture, but restricting the features to only those used in SAPS-II. This allowed for a direct comparison between the ensemble model’s architecture and traditional scoring methods while controlling for data variations.

2.4. Parameter optimisation for study design

EHR-ML encompasses an analysis suite to determine the optimal study design for modelling a specific clinical outcome and guide efficient data collection and preprocessing strategies.

- **Class Ratio Analysis** This analysis explores the fluctuation in model performance as the proportions of positive and negative classes vary (Figure 1-C). Understanding the impact of class imbalance helps identify reliable performance metrics in extreme class ratio scenarios and informs strategies to mitigate the bias. Additionally,
Figure 2: This figure compares the original data (A) to the learned feature representations (B). While the data is high-dimensional and difficult to visualise, t-SNE projects it onto a 2D space for comparison. Each data point is coloured and shaped according to its class (orange and "x" for class 1, blue and "o" for class 0). Comparing the two plots reveals that the learned feature representations lead to a clearer separation between classes. This improved class separability translates to better predictive performance in downstream modelling tasks.

it helps assess whether a particular clinical outcome has adequate representation from the minority class to be considered suitable for modeling.

- **Sample Size Analysis** This analysis serves to pinpoint the ideal data size necessary for a specific predictive task, a critical aspect for both retrospective and prospective studies. The process entails randomly sampling data of various sizes and constructing machine learning models utilizing 5-fold cross-validation. By doing so, it enables the evaluation of current data sufficiency and provides direction for data augmentation in retrospective studies. Furthermore, it offers insights into sample size necessities for prospective studies. (Figure 1-D).

- **Data Window Analysis** By varying the "window before" (WB) and "window after" (WA) parameters (Figure 1-E), EHR-ML finds the optimal window for collecting data relevant to a prediction task. The best WB parameter determines the sufficient extent of historical data needed, while the best WA parameter reveals the optimal time duration after the admission (or a custom anchor point) for obtaining the data to get reliable outcome predictions.

- **Data Standardisation Analysis** This analysis compares the performance of models trained on raw and scaled data (Figure 1-F). It helps to decide if scaling is beneficial and, if so, which scaling strategy provides the best results. While some machine learning models handle rescaling internally, others are sensitive to it, necessitating careful analysis.

Additionally, EHR-ML offers flexibility by allowing these analyses to be applied either to the ensemble model or a standalone machine learning model.

3. Results

3.1. EHR-ML outperforms off-the-Shelf models in a comprehensive evaluation

We compared the EHR-ML pipeline against both constituent models within the ensemble and standalone models constructed with standard Python libraries. First, we trained EHR-ML to predict the risk of mortality in patients diagnosed with sepsis [64]. Subsequently, we performed 5-fold cross-validation and obtained the average AUROC,
MCCF1, Accuracy, Balanced Accuracy, Average Precision, and F1 scores. These metrics were also calculated for the constituent models within EHR-ML, grouped into four categories based on their machine learning algorithm namely XGB, LR, LGBM, and MLP. Figure 3-A presents a spider plot comparing the performance of EHR-ML with the best-performing constituent model under each category. Further, we trained standalone XGB and LR models on the same dataset. Eventually, the performance of each model was evaluated using various metrics, visualised in figure 3-B and further detailed in supplementary tables S2 and S1.

Figure 3: Performance comparison of EHR-ML and individual model constituents across six evaluation metrics, including AUROC, MCCF1, Accuracy, Balanced Accuracy, Average Precision, and F1. A) Panel A compares EHR-ML against the best performing constituent model within each category. The categories include XGB, LR, LGBM, and MLP and there will be 12 individual models under each of them. B) Panel B compares EHR-ML against standalone models trained on the same dataset. Overall, these plots demonstrate the consistent performance advantage of EHR-ML relative to both internal constituent models and external standalone models.

Across both comparative evaluations, EHR-ML consistently demonstrates an improved overall performance against all constituent and standalone models, demonstrating its superior overall performance in clinical prediction tasks. This suggests that the two-layered ensemble approach of EHR-ML successfully leverages the complementary strengths of multiple constituent models to achieve superior performance and generalizability compared to off-the-shelf solutions. Specifically, it benefits as each Level 1 model captures different aspects of the data, leading to a more comprehensive representation upon combining. The well-separated predictions from Level 1 make the final model less susceptible to class imbalance issues. As the second layer operates on probability outputs from the previous layer, the need for data scaling or standardisation is less crucial. Overall, this two-layered architecture allows EHR-ML to extract rich learned feature representations from the data, leverage diverse modelling approaches, and mitigate the impact of class imbalance, ultimately leading to robust and accurate clinical outcome predictions.

3.2. EHR-ML outperforms traditional scoring system (SAPS II)

We further evaluated the performance of EHR-ML trained on all the attributes, referred to here as EHR-ML-FULL, against a traditional severity of illness scoring scheme, SAPS II [56]. In addition, we included EHR-ML-SAPS-II, a model also based on EHR-ML architecture but trained only on SAPS II variables. Figure 4-A presents overlapping ROC curves for SAPS-II, EHR-ML-SAPS-II, and EHR-ML-FULL, allowing for comparison of their predictive power across various thresholds. Further, figures 4-B, 4-C, and 4-D further depict the confusion matrices for these models at a calibrated threshold. Furthermore, the table 2 presents their TP, TN, FP, and FN values. This analysis was performed...
to facilitate the comparison of the strengths and weaknesses of EHR-ML and the SAPS-II in identifying true and false predictions for the two target classes.

Figure 4: The figure depicting a comparative analysis between the three methods - EHR-ML-FULL, the traditional scoring scheme - SAPS II, and the EHR-ML-SAPS-II model. A) This figure shows an overlapping ROC plot for all three models - EHR-ML-FULL, EHR-ML-SAPS-II, and SAPS-II. B-D) The confusion matrix is presented as a heatmap for SAPS-II, EHR-ML trained on SAPS II variables - EHR-ML-SAPS-II, and EHR-ML trained on all the variables - EHR-ML-FULL. In the confusion matrix, the vertical axis shows actual patient survival while the horizontal axis displays predicted outcomes. Both axes use "0" for the negative class (survival), and "1" for the positive class (deceased). Each cell reveals the number and percentage of patients falling into each category, with the total class counts displayed in the diagonal cells. The variations in colour intensity visually highlights the percentage of observations in each cell.
Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>AUROC</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAPS II</td>
<td>0.5907</td>
<td>25.2 %</td>
<td>92.9 %</td>
<td>74.8 %</td>
<td>7.1 %</td>
</tr>
<tr>
<td>EHR-ML-SAPS-II</td>
<td>0.7040</td>
<td>48.1 %</td>
<td>92.7 %</td>
<td>51.7 %</td>
<td>7.3 %</td>
</tr>
<tr>
<td>EHR-ML-FULL</td>
<td>0.7217</td>
<td>51.4 %</td>
<td>93.0 %</td>
<td>48.6 %</td>
<td>7.0 %</td>
</tr>
</tbody>
</table>

Figure 4-A reveals the superiority of both EHR-ML variants in predicting mortality risk as compared to conventional SAPS II scoring scheme across various thresholds. Further analysis of the confusion matrices presented in figures 4-B, 4-C, and 4-D reinforces this observation. In all three plots, both EHR-ML variants outperform SAPS II across all quadrants, showing a notable twofold increase in true positives while maintaining similar levels of false positives. This translates to a remarkable improvement, with EHR-ML variants correctly identifying mortality rates twice as often as SAPS II. Moreover, the comparison serves as compelling evidence for the effectiveness of EHR-ML’s two-tier ensemble architecture and the encompassing feature representation learning. The fact that both the models utilise the same data set, highlights the inherent advantage of the EHR-ML’s novel architecture. In conclusion, these findings demonstrate that EHR-ML offers significant improvements over traditional methods like SAPS II in terms of accurately identifying high-risk patients.

3.3. Superior performance without the need for scaling

To investigate the requirement of data scaling, we evaluated the model on three different datasets. The raw dataset contained the original, non-scaled features directly extracted from the EHR. Secondly, the standard-scaled data, utilised a standard scaler to transform each feature to have a mean of 0 and a standard deviation of 1. Third, the min-max scaled data, employed a min-max scaler to transform each feature to have a minimum value of 0 and a maximum value of 1. On these three datasets, we performed a 5-fold cross-validation for each dataset to assess the model’s performance in predicting in-hospital mortality (supplementary table S3). Next, we calculated the AUROC for each fold and subsequently generated a box plot summarising the distribution of these values over 5-fold cross-validation (supplementary figure S2).

The EHR-ML model proposed shows excellent accuracy in predicting mortality risk, and it achieves this without needing data scaling. This is because the model’s initial layer generates feature representations that are inherently uniform, with each subset of features producing a risk score between 0 and 1. These scores act as inputs for the second layer of the model, which combines them to make final predictions. Because the inputs to the final model are already within a consistent range, traditional scaling methods are unnecessary. This is a significant advantage of EHR-ML, as it simplifies the model pipeline and reduces computational overhead. Additionally, the absence of data scaling ensures that the model’s predictions are not influenced by scaling parameters, making them more robust and generalizable across different datasets.

3.4. EHR-ML excels with both narrow and extended data windows

To determine the amount of data required for the model to make reliable predictions, we performed a data window analysis. The analysis involved varying the size of a temporal window around the anchor point by modifying two parameters namely window before and window after. In this analysis, the window before parameter spanned from 0 to 3 days before the anchor point. Similarly, the window after parameter ranged from 1 to 14 days after the anchor point at the end of which the predictions were made. For each combination of the two parameters, we calculated the EHR-ML model’s performance in terms of average AUROC over 5-fold. These values are tabulated in table 3 and plotted as a heatmap in figure 5, providing a visual representation of the impact of the data window in predicting in-hospital mortality.

The figure (Figure 5) reveals the impressive performance of EHR-ML even with limited data availability. Notably, the model achieves a good AUROC of 0.72 on day 1, using data from less than 24 hours post-admission. This is comparable to the performance of SAPS II (0.73) built using data collected for more than the first 24 hours. This
Figure 5: Heatmap displaying AUROC performance across varied data collection windows: Lower boundary initiated 3 days prior to the anchor day (ICU admission) and concluding on the anchor day; upper boundary commencing 1 day post-anchor point and concluding 14 days thereafter. The colour intensity for each cell corresponds to labelled average AUROC values. A legend is provided on the right for reference.

Table 3
The table presents average AUROC values across various time windows. The data collection window’s lower boundary begins 3 days before the anchor day (ICU admission) and concludes on the anchor day. The upper boundary starts 1 day post-anchor point and concludes 14 days afterward.

<table>
<thead>
<tr>
<th>(Days) \ (Days)</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td>1</td>
<td>0.72</td>
<td>0.77</td>
</tr>
<tr>
<td>2</td>
<td>0.73</td>
<td>0.77</td>
</tr>
<tr>
<td>3</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

demonstrates the early predictive power of EHR-ML, potentially enabling timely interventions and improved patient outcomes. Furthermore, as the data window increases, providing more context, EHR-ML’s performance consistently improves, surpassing the benchmark of SAPS II by a significant margin on day 2 at which point there is data of a full 24-hour period. This trend continues as the data window expands, with performance steadily improving due to the model’s ability to capture and utilise temporal information from the time series. This analysis demonstrates EHR-ML’s dual strengths: achieving competitive performance at the very beginning of a patient’s admission and exhibiting continual improvement with longer data availability.
3.5. **EHR-ML achieves high performance with limited data**

To evaluate the impact of sample size on EHR-ML’s performance, we conducted a sample size analysis. We randomly sampled the data from the eICU ICD Cohort, starting with 200 samples and incrementally increasing the size by 100 until reaching 1000. Beyond this point, the sample size increased in steps of 1000, culminating in the full dataset size of 11,146. For each sample size, we performed 5-fold cross-validation and calculated the average and standard error of the AUROC for predicting in-hospital mortality (S4 and figure 6).

![Average AUROC across different sample sizes](image)

**Figure 6:** The figure displays a bar plot presenting the average AUROC obtained through 5-fold cross-validation across various sample sizes. Error bars depict standard errors for each sample size. The sample sizes represent the number of episodes used in the analysis, starting from 200 episodes with increments of 100 until 1000. Beyond this, the sample size increases in steps of 1000 until reaching 10000, with the maximum available samples (11146) included at the end.

Figure 6 reveals a substantial initial performance gain as the sample size increases, reaching a plateau of around 500 samples. This suggests that EHR-ML achieves significant performance even with a relatively small dataset, which is remarkable considering the high dimensionality of the data. This can be attributed to the model’s architecture, which divides features into various subsets. This approach allows EHR-ML to effectively handle high-dimensional data without affecting its predictive capabilities. Furthermore, the figure suggests that performance stabilises after 3,000 samples, with minimal variation observed beyond this point. This indicates that 3,000 samples may represent an optimal sample size for achieving stable and reliable results for this analysis using EHR-ML. This insight can be used to improve research design and resource allocation in using this model for practical applications.

3.6. **Cautionary note on evaluating performance in the face of imbalance**

The next EHR-ML analysis explored the effect of class imbalance on standard performance metrics. Class ratio indicates the proportion of various classes in the dataset, like the ratio of positive to negative observations in binary classification. In EHR-based prediction tasks, encountering highly imbalanced classes is common. This analysis aimed to identify metrics better suited for highly imbalanced data and potentially misleading ones. To achieve this, we created datasets with varying class ratios, ranging from a perfect 50-50 balance to a highly skewed 95-5 ratio. We then calculated the average value of various performance measures for each class ratio and tabulated the results in supplementary table S5. Subsequently, a line plot (supplementary figure S3) presenting the trend lines for these measurements is plotted, providing a visual representation of the behaviour of different metrics under different levels of class imbalance.
The supplementary table S5 and supplementary figure S3 provide valuable insights into the impact of class imbalance on various performance metrics. Increasing class imbalance reveals notable discrepancies in metric behavior. AUROC remains relatively stable, while accuracy can be misleadingly inflated, potentially obscuring issues in minority class identification. Conversely, F1 score and MCC exhibit sensitivity to class imbalance, offering a more nuanced evaluation of model performance across all classes. Additionally, MCCF1 [72], combining F1 and MCC, emerges as a promising metric in highly imbalanced data scenarios. These findings underscore the importance of carefully selecting performance metrics for models trained on imbalanced data. Metrics like accuracy may appear intuitive but can be deceptive in such contexts. Instead, prioritizing metrics like MCCF1 provides a more dependable assessment of model performance across diverse classes.

3.7. Robust LOS prediction with EHR-ML

Another outcome that is considered in this study is predicting the LOS. Specifically, two separate EHR-ML classifiers were trained to predict whether an ICU admission would exceed 7 and 14 days. To that end, two new target attributes were derived from the existing patient admission records, each corresponding to the respective LOS cut-off. Next, the optimal configuration for both the targets we obtained by running benchmarking analysis. Figure 7 presents the results from this analysis.

The results presented in supplementary tables S6, S7, and the figure 7 illustrate performance of the EHR-ML in predicting LOS exceeding 7 and 14 days. The supplementary table S6 reveals impressive performance for both models in predicting LOS. Notably, the high AUROC values consistently exceeding 0.95 and MCCF1 values over 0.8 indicate excellent discrimination between ICU admissions with long and short stays. Figure 7-A further reinforces these findings through the ROC curves for both models exhibiting excellent coverage. Figures 7-B and 7-C offer visual confirmation of the model’s ability to correctly identify both classes with high confidence. The confusion matrices reveal high percentages for both true positives and true negatives. Furthermore, by examining the diverse time windows and their influence on model performance in supplementary table S7, figures 7-D, and 7-E, researchers can make informed decisions regarding the most suitable window for their specific predictions.

3.8. EHR-ML offers a user-friendly interface for clinical outcome prediction from EHR

EHR-ML simplifies model building, performing prediction, running cross-validation evaluation, and performing analysis such as data window analysis, sample size analysis, class ratio analysis, and standardisation analysis. It offers two distinct access points - command line library and a web-portal to promote open research practice. The command line library empowers users with technical expertise to leverage EHR-ML’s functionalities through straightforward installation and execution. This option grants direct control over the analysis process using the command line, enabling customization and integration with existing workflows. In addition, to cater to the users with limited programming experience, the web-portal (refer to figure 8) is made available. It provides an web-based interface for accessing the full spectrum of EHR-ML’s capabilities. Both access points require data to be formatted in a specific manner. EHR-QC [69], our previously developed toolkit, conveniently aids in preprocessing as well as preparing data compatible with EHR-ML. Besides, the simple data representation format required by the utility is easy to construct without relying on any specific tools.

4. Discussion

The use of machine learning for predicting clinical outcomes shows significant potential in improving healthcare decision-making. However, this field faces limitations that hinder its impact. Our survey reveals a prevalent reliance on expert intuition in constructing predictive models, potentially resulting in suboptimal solutions and hinder automation. Additionally, majority of studies rely on one-time, inflexible codebases, limiting their applicability to various clinical outcomes. The lack of a standardized process impedes the comparison of different studies, hindering the generation of reliable knowledge. Moreover, the use of off-the-shelf models overlooks the complexities of health domain time series measurements and the challenge of class imbalance.

The EHR-ML pipeline emerges as a compelling solution to the challenges encountered in clinical outcome prediction. Moreover, its adaptable interface allows for the customization of data windows, anchor points, and target variables, offering precise control over the modeling process. Additionally, it seamlessly integrates with our EHR-QC preprocessing tool [69], facilitating a streamlined and automated workflow, from data sourcing to outcome modelling. Its innovative approach, characterized by a data-driven methodology and a flexible open-source codebase, holds
Figure 7: This figure compares the performance of two models in predicting whether a patient’s LOS in the ICU will exceed set durations of 7 and 14 days. A) The plots in Panel A display ROC curves for both the models. This visualization helps assess the trade-off between sensitivity and specificity across different thresholds. The AUC values are listed for each model in the figure legend along with the corresponding label. B - C) The next two plots in panel B and C show the confusion matrix of the two models. Analysing these values allows us to evaluate the models’ ability to identify both positive and negative cases, as well as the rate of misclassification for each category. In the confusion matrix, the vertical axis represents actual patient LOS status while the horizontal axis corresponds to the predicted outcomes. Both axes use "0" for the negative class (stays below the set duration), and "1" for the positive class (stays over the set duration). Each cell reveals the number and percentage of patients falling into each category, with the total class counts displayed on the diagonal cells. The variation of the colour intensity visually highlights the percentage of observations in each cell. D - E) Panels D and E present heat-maps illustrating the variation in AUROC performance for different data collection windows. Each cell corresponds to an average AUROC value, with colour intensity reflecting the level of performance. The lower boundary of each window starts 1 day before ICU admission (anchor day) and ends on the anchor day, while the upper boundary begins 1 day after the anchor point and concludes 7 days later. A legend for colour interpretation is provided on the right.

promise in addressing these hurdles. Through comprehensive benchmarking analysis, we anticipate that the EHR-ML pipeline will streamline the process of knowledge generation and facilitate process automation. One notable strength of the EHR-ML pipeline is its two-layered ensemble modelling approach which offers a proven machine learning model specifically designed for EHR data. This tailored approach enhances the accuracy and reliability of predictive models in healthcare settings. Additionally, the feature engineering functionalities provided by the EHR-ML excels at extracting and leveraging temporal signals from time-series data, a key challenge in this domain. Furthermore, the performance metrics offered by the EHR-ML pipeline provide comprehensive insights into model performance in real-world scenarios, even when dealing with skewed data distributions. This capability ensures that the models can be effectively evaluated and optimized for practical application in healthcare settings.
Figure 8: This figure provides a screenshot of the EHR-ML web utility interface. This user-friendly platform enables model building, predicting prediction, running cross-validation, and performing several analyses such as data window analysis, sample size analysis, class ratio analysis, and standardisation analysis. The intuitive interface allows users to upload formatted data and readily conduct various analyses.

5. Conclusion

EHR-ML represents a notable stride in democratizing and streamlining clinical outcome prediction with EHR data. By eliminating common obstacles and offering a straightforward route to dependable and replicable analyses, it facilitates accessibility. Available both as a command-line interface and a user-friendly web utility, it empowers researchers across various technical proficiencies to conduct predictive modeling for a wide range of clinical outcomes with ease. The open-source nature of the source code encourages community involvement, not only for utilization but also for active contribution. This contribution aims to propel the field forward, fostering reproducibility, comparability, and data-driven optimization in EHR-based clinical outcome studies.

6. Acknowledgements

AP, NM, GW, and ST acknowledge funding support of Medical Research Future Fund (MRFF) for the SuperbugAI flagship project. YR received Monash Graduate Scholarship for his PhD.

We thank to Jerico Revote from Monash eResearch Centre and William Librata of Alfred Health IT team for their invaluable support to help set up cloud computing infrastructure used in this work.

The authors also extend their sincere appreciation to the open research community responsible for making the following resources available which were instrumental in facilitating the execution of this research: EHR-QC [69], MIMIC IV [73], MIMIC-Extract [70], MIMIC IV to OMOP CDM Conversion (https://github.com/OHDSI/MIMIC), eICU Collaborative Research Database [65], Athena (https://athena.ohdsi.org/), SNOMED (https://www.snomed.org/).

Finally, we acknowledge the Python libraries that are an integral part of EHR-ML toolkit. These include scikit-learn, pandas, numpy, psycopg, scipy, matplotlib, xgboost, and lightgbm.

7. Code availability

The EHR-ML source code is now accessible to researchers for investigative purposes through the following Git repository: https://gitlab.com/superbugai/ehr-ml. Comprehensive documentation for the utility can also be found at https://ehr-ml-tutorials.readthedocs.io.
CRediT authorship contribution statement

Yashpal Ramakrishnaiah: Formal analysis, Investigation, Visualisation, Writing – review and editing, Software development.

Nenad Macesic: Investigation, Funding acquisition, Supervision, Writing – review and editing.

Geoffrey I. Webb: Investigation, Resources, Supervision, Writing – review and editing.

Anton Y. Peleg: Investigation, Funding acquisition, Resources, Supervision, Writing – review.

Sonika Tyagi: Conceptualisation, Formal analysis, Investigation, Funding acquisition, Resources, Supervision, Writing – review and editing.

References


