A Prediction Model of In-Patient Deteriorations Based on Passive Vital Signs Monitoring Technology

Veronica Maidel[1], Maayan L Yizraeli Davidovich[1], Zvika Shinar[2], Tal Klap[1]

[1]EarlySense Ltd., Ramat Gan, Israel
[2]MindUP, Haifa, Israel

Abstract

Lately, many health systems accelerated their initiatives of advanced remote monitoring systems. Moving to an unattended environment requires overcoming patients' compliance issues and demonstrating the effectiveness of remote monitoring technology. Current Early Warning Scores detection of deterioration, commonly based on spot check EMR data, demonstrates low translational impact from one facility to another. In this study we used vitals collected passively by a sensor, to build a Machine Learning model for timely prediction of deteriorating patients, within 24-hours of their transfer to ICU or death. Time series features, such as trends and vitals' variability were used in conjunction with age & comorbidity data. Evaluating the model yielded an AUROC of 0.81 on data from an inpatient setting, and an AUROC of 0.88 on an independent test set from a COVID-19 unit. The suggested model, based on passive measurement technology, performs equally well as models based on EMR that include nurse inputs. Applying the model on other acute settings (such as a COVID-19 unit) showed similar performance, increasing confidence of its robustness and transferability. The model performance combined with the fact that it does not require human compliance, makes it a good candidate for future testing on home settings.

1. Introduction

Lately, healthcare has been shifting larger segments towards remote patient monitoring, home hospitalizations and unattended settings. This shift raises the importance of having Early Warning Scores (EWS) to detect impending deterioration and allow timely intervention. However, current EWS for detection of deteriorations in adult non-ICU patients demonstrate low precision (positive predictive value (PPV)) [1] and typically have minimal impact when translated from one facility to another [2].

The common EWS (NEWS, MEWS, ViEWS) are based on spot-check EMR data. Other scores, use a combination of continuous monitoring with wearables and intermittent manual monitoring by unit staff [3]. These manually collected data points introduce a nurse-bias (human factor) into the data and the inferred model. This creates a bias in the quality and quantity of the vitals collected and affects the model. This bias adversely affects the translational performance of the EWS, as the 'nurse factor' changes between different settings and guidelines. Thus, when the model is tested outside of the facility where data originated – it performs worse than reported [2].

In this work, we hypothesized that by eliminating data dependency on the 'nurse factor'– we may be able to establish an EWS that is more objective, and less dependent on a specific setting, e.g., specific nurse-to-patient ratio, or specific locations where nurses follow different guidelines.

We used continuous and contact free monitoring systems to record vital signs (EarlySense LTD), and patients' data from a single site, in order to build a model based solely on objective data. We then tested the performance of the model in a completely different setting working under different guidelines to test our hypothesis that the performance will not vary.

2. Methods

We developed a machine learning model to predict in-patient deteriorations that resulted in patient transfer to Intensive Care Unit (ICU) or death. Data from the last 24 hours before an event (transfer to ICU, hospital discharge or death) was used for training and tuning, and predictions were calculated on an hourly level and evaluated on an admission level.

2.1. Data

A total of 38,502 admissions (26,504 patients, 14,899 Female, age: 66.6±18.8, and 11,605 Male, age: 65.5 ±17.3, mean±SD) collected over a period of 3.5 years from Newton Wellesley Hospital (NWH) in Massachusetts, USA, were included in the study, of which, 1,118 admissions ended in a deterioration. Patients were from various hospital unit types: Medical-Surgical, Orthopedic,

Post-operation, and Cardiac units. All patients were monitored with the EarlySense (ES) system, based on a contact-free piezo-electric sensor, placed under the patient's mattress. The ES monitor continuously tracks Respiratory Rate (RR), Heart Rate (HR), motion level, and bed occupancy. Additional data including outcome, age, and a one-time entry of the patient's comorbidity information from the EMR was used.

The datasets in Table 1 were used to train, tune, and test a collection of Gradient Boosting classifiers. Stratified random splitting of the NWH data into training (44%), validation (22%), and testing (33%) sets was done. In addition, we made sure that data from the same admission does not appear in more than one dataset.

Table 1. Datasets used to train, tune, and test the model.

Datasets	Admissions	Deteriorations %
Training set	17,091	2.9%
Validation set	8,549	2.7%
Testing set	12,862	2.9%

2.2. Features

Feature selection was done by visually inspecting plots of the features over time, in the training set, and selecting features that have the best correlation with the outcome, while excluding features that were correlated with each other. This was done to allow better explainability of the model to the end user. A total of 11 features unique to continuous vitals monitoring were selected, and fed into 5 sub-models, with each sub-model relying on 3-4 features. Using a small number of features per sub-model helped prevent over-fitting, while preserving a variety of features.

The features were based on 10-second resolution data. They were calculated every hour with a rolling window size of 1-6 hours, depending on the feature. To deal with missing data, we set conditions for the minimum time in bed and measurements required for a feature to be calculated to keep as much data as possible, while preserving its quality. In addition, a weighted moving average on 4 hours was used on most features to smooth feature values and reduce missing data. The features included: HR and RR six-hour trends, HR and RR variability in last three hours, HR and RR one-hour median, irregularity of HR in the last six hours (may be related to e.g., arrhythmia), and the percentage of age-normalized HR or RR exceeding certain thresholds.

We used age-normalized HR, calculated as: age_normalized_HR = (HR-HRmin)/(HRmax-HRmin), with HRmin=40, and HRmax=(220-age). This normalization is commonly used in the field of sports [4]. Age is an important feature and has a substantial effect on vitals. However, with the relatively small sample size relative to the various ways "to deteriorate" we chose to factor the "age" feature with the vitals.

2.3. Model development

An overview of the model development process is described in figure 1 below, and will be described in the current section.

We performed relabeling of the training set so training could be done on more consistent data. As the deterioration is a process, during the observed 24-hours some data points may include rather stable vitals, and may not suggest deterioration. In addition, some negative labels may be a result of deteriorations that actually occurred, but were treated on time, and therefore may include extreme vitals, indicating deteriorations. Instead of giving the same label for each hourly calculation of the features, based on whether or not the admission ended in deterioration, we

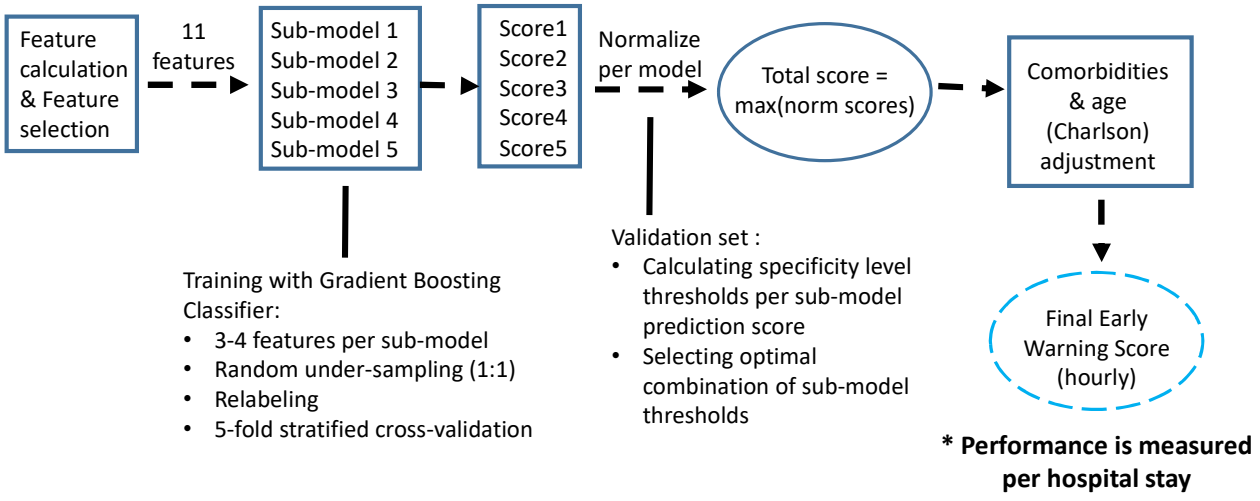

Figure 1. An overview of the ES model, depicting the model development process.

relabeled some of the hourly labels from positive to negative and vice versa. This was done by examining the cumulative probability distributions of the features on the training set with and without deteriorations, and identifying thresholds for relabeling (low and high percentiles).

The ES model is comprised of five sub-models, with each one focused on a different aspect of vitals: age-normalized HR and RR above thresholds, age-normalized HR and RR below thresholds, HR and RR median and HR variability, HR irregular rate, and HR and RR trends. Each sub-model was trained separately with a Gradient Boosting Classifier. During the training stage we performed a 5-fold stratified cross validation, for hyper parameter tuning - making sure that there is no bleed-through of subjects between the training and the testing folds. Random under-sampling (1:1) was performed on the train-fold only, to help prevent over-fitting due to the unbalanced data.

Each sub-model produces its own probability prediction for a deterioration. The validation set was used to calculate several threshold levels on each sub-model's prediction, based on a configurable desired specificity. These specificity threshold levels allow controlling the overall sensitivity/specificity level of the model, and easy tuning of the sub-models so the model is optimized for different clinical settings and outcomes (e.g., COVID-19, post-acute, etc.). The probability of each sub-model was normalized (between 0 and 1) by the desired specificity threshold level for each sub-model, and the maximum prediction score was taken as the total prediction score of the final model.

An additional adjustment for comorbidities and age was then used to adjust the prediction score output of the final model, according to the probability of a deterioration, given the subject age and Charlson Index (calculated from comorbidities) [5]. The probability was calculated based on the training set, and hyper-parameter tuning was done on the validation set, resulting in a parameter of a=0.75 giving optimal results for the weight between the final prediction score and the Charlson-based adjustment.

To determine an operating point for the final model, we checked different thresholds for the prediction score on the validation set, and examined PPV. We chose a point with a relatively high PPV and a high enough sensitivity. Classification was done on an hourly level, and on a 24-hour admission level. An admission was marked "positive" for deterioration if at least one hour of the last 24 hours was marked as "positive".

2.4. Model testing on a separate COVID-19 dataset

Before testing the model on separate data from a COVID-19 unit, tuning of the model for respiratory population from NWH was done. The machine learning model was applied to a respiratory validation subset within the NWH dataset, choosing only patients with respiratory ICD-10 codes. The purpose of this was to optimize the model tuning-parameters to patients with respiratory symptoms (pre-COVID). We calculated all the combinations of sub-models with the specificity threshold levels from the regular model, and chose the combination with the desired specificity level of the overall model performance on the respiratory validation subset. The tuned model was then tested on data from a COVID-19 unit in Sheba Medical Center (Israel) between April 2020 and Sept 2020.

Table 2. Datasets used for tuning the model for respiratory population, and testing on a COVID-19 population.

Datasets	Admissions	Deteriorations %
Respiratory validation subset	1,110	4.4%
COVID-19 test set	131	18.3%

3. Results

As seen in figure 2, visualizing the model predictions versus the HR and RR data over time, can follow deteriorations and changes in the patient status, and thus help with the explainability of the model.

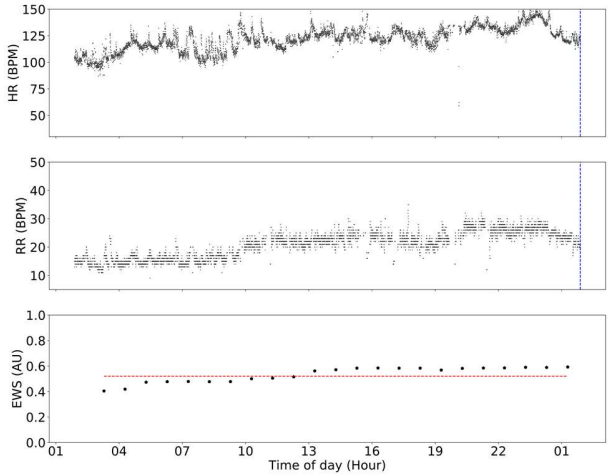

Figure 2. HR (upper panel), RR (middle panel), and hourly EWS (lower panel) for 24 hours before transfer to ICU of a 50-year-old patient. The red horizontal line depicts the threshold for deterioration classification. The blue vertical lines mark the time of transfer to ICU. The model first detected deterioration 13 hours before the event.

An evaluation of the model for hourly predictions of patients' deterioration within 24-hours of an event (death, transfer to ICU or discharge) yielded an Area under the Receiver Operator Curve (AUROC) of 0.81. This is

comparable to an AUROC of 0.78 reported in the literature for an EMR-based a machine learning model with a similar population and deterioration criteria [6]. For a specific operating point that was chosen, we got sensitivity, specificity and PPV of 61%, 84%, and 9%, respectively.

When evaluated on a set of 131 COVID-19 patients at Sheba Medical Center, the suggested model achieved an AUROC of 0.88. For a specific operating point that was chosen, we got sensitivity, specificity and PPV of 67%, 92% and 67%, respectively.

4. Discussion

The suggested model, developed on data from acute settings, based on passive and continuous measurement technology, has achieved similar performance to models developed on EMR systems that include nurse inputs. Applying the model in other acute settings (COVID-19 unit) showed similar (or better) performance, increasing the confidence of its robustness and its transferability to other clinical settings. This supports our hypothesis that training a model on features based on vitals that were automatically measured provides consistency in the performance of the model in different settings.

Our EWS model, based on the ES system presents unique advantages relative to EMR-based EWS. These include, continuous and objective measurements that are not dependent on manual data inputs, such as spot-checks, which may vary due to e.g., different guidelines, personnel skill level, "white coat" effects, etc. In the home or post-acute setting this may be especially important as skilled personnel is not as available. In addition, the system is contact-free and passive, and therefore it does not interrupt the daily routine of the subject or require compliance, once the sensor is placed under the mattress. These advantages suggest that the system may be compatible to home and post-acute settings.

Our EWS model presented low PPV for the NWH data. This may be related to timely interventions that occurred during the course of the deterioration process. This may explain why in the COVID19 unit, where some of the care was given remotely, the PPV was significantly higher. Additionally, some false positive classifications were likely related to other deterioration types (e.g., infections, fever) that occurred.

The model used labels based on ICU transfers and death. However, in the clinical setting predicting other forms of deterioration requiring medical attention may be important. The model training was done on relabeled data, due to the need to transform the binary output label (of transfer to ICU or death) to an hourly label providing consistent training data for deterioration predictions. This improved the model performance. However, it may have added some bias towards deterioration types relating to very high or low HR or RR values, which do not necessarily result in ICU transfer or death, but may require medical attention nonetheless.

For the model to be more relevant for the home or post-acute setting, which are compatible with the ES system characteristics, adding sub-models relating to longer term features (in terms of days) may be relevant. In addition, improving the explainability of the model, by including vitals plots and explainability features, may help medical staff better utilize the model as a decision support system. We are currently exploring both of these directions. Future tuning and testing the model on post-acute and home settings, may help reduce admissions of high-risk individuals, and aid in providing prompt interventions, with minimal human compliance.

References

[1] S. Gerry, T. Bonicci, J. Birks, S. Kirtley, P. S. Virdee, P. J. Watkinson and G. S. Collins, “Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology,” *BMJ, vol. 369, no. m1501, May 2020*.

[2] A. D. Bedoya, M. E. Clement, M. Phelan, R. C. Steorts, C. O’Brien and B. A. Goldstein, “Minimal Impact of Implemented Early Warning Score and Best Practice Alert for Patient Deterioration,” *Crit Care Med., vol. 47, no. 1, pp. 49–55*, *Jan. 2019*.

[3] Clifton L, Clifton DA, Pimentel MAF, Watkinson PJ, Tarassenko L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors,” *IEEE J Biomed Health Inform, vol. 18, no. 3, pp. 722-730*, May 2014.

[4] G. F. Fletcher, P. A. Ades, P. Kligfield, R. Arena, G. J. Balady, V. A. Bittner, et al., "Exercise standards for testing and training: a scientific statement from the American Heart Association," Circulation, vol. 128, no. 8, pp. 873-934, Aug. 2013.

[5] M. Charlson, T. P. Szatrowski, J. Peterson, & J. Gold, "Validation of a combined comorbidity index," J Clin Epidemiol, vol. 47, no. 11, pp. 1245-1251, Nov. 1994.

[6] S. Muralitharan, W. Nelson, S. Di, M. McGillion, P. Devereaux, N. G. Barr, et al., "Machine Learning–Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review," J Med Internet Res, vol. 23, no. 2, e25187, Feb. 2021.

Address for correspondence:

Maayan L Yizraeli Davidovich.
Jabotinsky 7, Ramat Gan, 52520 ISRAEL.
maayan.yd@earlysense.com