Early Prediction of All-Cause Clinical Deterioration in General Wards Patients: Development and Validation of a Biomarker-Based Machine Learning Model Derived From Rapid Response Team Activations

Antoine Saab, MEng,† Cynthia Abi Khalil, MSN,‡ Mouin Jammal, MD,§ Melody Saikali, MSc,† and Jean-Baptiste Lamy, PhD**

Objective: The aim of the study is to evaluate the performance of a biomarker-based machine learning (ML) model (not including vital signs) derived from reviewed rapid response team (RRT) activations in predicting all-cause deterioration in general wards patients.

Design: This is a retrospective single-institution study. All consecutive adult patients' cases on noncritical wards identified by RRT calls occurring at least 24 hours after patient admission, between April 2018 and June 2020, were included. The cases were reviewed and labeled for clinical deterioration by a multidisciplinary expert consensus panel. A supervised learning approach was adopted based on a set of biomarkers and demographic data available in the patient's electronic medical record (EMR).

Setting: The setting is a 250-bed tertiary university hospital with a basic EMR, with adult (>18 y) patients on general wards.

Patients: The study analyzed the cases of 514 patients for which the RRT was activated. Rapid response teams were extracted from the hospital telephone log data. Two hundred eighteen clinical deterioration cases were identified in these patients after expert chart review and complemented by 146 "nonevent" cases to build the training and validation data set.

Interventions: None

Measurements and Main Results: The best performance was achieved with the random forests algorithm, with a maximal area under the receiver operating curve of 0.90 and F_1 score of 0.85 obtained at prediction time T_0–6h, slightly decreasing but still acceptable (area under the receiver operating curve, >0.8; F_1 score, >0.75) at T_0–42h. The system outperformed most classical track-and-trigger systems both in terms of prediction performance and prediction horizon.

Conclusions: In hospitals with a basic EMR, a biomarker-based ML model could be used to predict clinical deterioration in general wards patients earlier than classical track-and-trigger systems, thus enabling appropriate clinical interventions for patient safety and improved outcomes.

Key Words: machine learning, clinical deterioration, artificial intelligence, predictive model, clinical decision support

(*J Patient Saf* 2022;18: 578–586)

Delays in medical interventions in clinically deteriorating patients have been found to be associated with increased morbidity and mortality.[1–3] Therefore, early and continuous detection of gradually worsening patient conditions in hospital wards might allow for more rapid treatments and thus improved outcomes.[4]

The most common forms of clinical deterioration are respiratory instability, hemodynamic instability, sepsis, bleeding, cardiac decompensation, and acute hepatic/renal failure.[5] Deteriorating patients often require transfer to a higher level of care (such as intensive care units [ICUs]) and the urgent call for medical and nursing professionals for assessment and interventions.

Studies have documented that clinical signs and symptoms of patient deterioration (such as hypotension, bradycardia, tachypnea, tachycardia, altered level of consciousness, etc) can be detected as early as 6 to 8 hours before the deterioration event or cardiorespiratory arrest.[6]

These findings, derived from the late 1990s, led to the development and wide implementation of specific hospitals early warning systems (EWSs) called "track-and-trigger" systems, which can help predict clinical deterioration. These systems rely on the periodic observation of selected basic clinical signs ("tracking") with predetermined calling or response criteria ("trigger") for requesting the attendance of staff who have specific competencies in the management of acute illness and/or critical care.[7] In practice, most of these systems are based on the regular measurement of vital signs,[8] that would serve to calculate a paper-based or electronic severity score with predetermined thresholds triggering a call for a rapid response team (RRT). This team then evaluates the patient and takes clinical actions to prevent or manage the deterioration. "Track-and-trigger" systems are currently still considered as the criterion standard with regard to detecting and responding to clinical deterioration and have been shown to increase the number of calls to the RRT, decrease the number of cardiac arrests, and improve the response time of emergency medical teams.[9]

However, these track-and-trigger systems have practical limitations. First, the time from detection to actual deterioration is relatively short (0–8 hours), which provides a small window of opportunity for appropriate interventions that could prevent or mitigate the clinical risks. Second, the deterioration prediction score is sensitive to data quality and availability. Thus, any delays, omissions, or errors in the measurement of vital signs, which are all human dependent factors, can potentially affect the performance of the deterioration prediction score. Moreover, automated versions of such track-and-trigger systems cannot be effectively implemented in hospitals with basic EMRs (i.e., staged as 0, 1, or 2 according to the Healthcare Information and Management Systems Society Electronic Medical Record Adoption Model adoption model classification[10]), because they do not include an electronic nursing flowsheet documentation module. It is to be noted that the proportion of hospital with such basic EMRs is significant worldwide, especially in third-world countries.[11]

A new and promising approach described in recent studies[12–16] involves the addition of physiological biomarkers measurements to the traditionally measured vital signs and demographic patient data routinely available in the EMR. Biomarkers are defined as

From the *LIMICS, Université Sorbonne Paris Nord, INSERM, UMR 1142, Bobigny, France; Departments of †Quality and Patient Safety and ‡Nursing Administration, Lebanese Hospital Geitaoui-UMC; and §Department of Internal Medicine, Faculty of Medical Sciences, Saint Joseph University, Beirut, Lebanon.
Correspondence: Antoine E. Saab, MEng, Beirut, Lebanon (e-mail: antoine_saab@outlook.com).
The authors disclose no conflict of interest.
Name of the institution where the work was performed: Lebanese Hospital Geitaoui-UMC, Beirut, Lebanon.
Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.journalpatientsafety.com).

biological characteristics (such as, e.g., the C-reactive protein (CRP), procalcitonin, serum creatinine, etc) that are objectively measured and used as indicators of certain physiopathological processes.[17] This approach is based on the hypothesis that changes in certain biomarkers can precede the onset of clinical signs and symptoms, sometimes as early as 48 to 72 hours[18,19] theoretically permitting an earlier prediction of deterioration than traditional track-and-trigger systems.

Moreover, most of the prediction models were trained according to cases with the following outcome variables: cardiorespiratory arrest/death/unexpected transfer to ICU. Only few studies[20,21] have adopted the activation of the RRT as an outcome for the training and validation of the predictive model,[22,23] although such an outcome encompasses a broader and richer perspective of clinical deteriorations. In fact, a significant percentage (almost half) of RRT deterioration cases end with stabilization of patients on wards,[24,25] a clinical scenario otherwise not used by most systems.

Finally, recent studies have shown that machine learning (ML)–based EWSs can achieve greater accuracy than aggregate-weighted EWSs,[26] thus their increased use in the derivation of new models.

The aim of this study is to elaborate and validate a biomarker-based model (without including vital signs data) based on absolute and differential biomarker values for the prediction of general (all-cause) clinical deterioration, using ML algorithms as a derivation method, and expert-reviewed RRT calls as the main outcome for model training and validation. Our hypothesis is that such a model could predict all-cause clinical deterioration earlier than track-and-trigger systems, without the need to use vital signs and other complex patient data (e.g., diagnosis, clinical notes…), thus allowing such an approach to be used in healthcare settings, which have even the most basic EMR systems. Ultimately, this may provide opportunities to intervene earlier, help allocate resources more effectively, and potentially improve the patients' health outcomes.

MATERIALS AND METHODS

The hospital institutional review board deemed this study as "exempt" from further review, because it does not directly involve human subjects.

Study Design and Setting

We conducted a retrospective single-institution cohort study of all consecutive adult (>18 y) hospitalized patients in noncritical wards for whom an RRT was called after 24 hours of their admission over more than a 2-year period (April 1 2018, through June 30, 2020).

The study took place in a 250-bed tertiary university hospital in Beirut, Lebanon. The hospital's EMR can be considered as basic (stage 1 as per the Healthcare Information and Management Systems Society's Electronic Medical Record Adoption Model). The system contains admissions/discharge/transfer data, basic ancillaries with limited integration (laboratory, radiology, and pharmacy), billing (procedures and consumables), but no electronic nursing or medical documentation, nor computerized physician order entry or clinical decision support applications.

Definitions

We have adopted the following complementary definitions for the clinically deteriorating patient: "one who moves from one clinical state to a worse clinical state which increases their individual risk of morbidity, including organ dysfunction, protracted hospital stay, disability, or death,"[27] and "a dynamic state experienced by a patient compromising hemodynamic stability, marked by physiological decompensation accompanied by subjective or objective findings."[28]

Data Collection

A multidisciplinary expert consensus panel (an internal medicine physician, a nurse, 2 patient safety professionals, and a panel of physicians from specialized disciplines consulted on demand) analyzed all 514 RRT calls that were extracted from the hospital telephone log data. Of these 514, the panel selected the 237 cases where sufficient documentation about the event was found. Sixteen patients for whom an RRT call was initiated within the first 24 hours were excluded. The remaining data set included 221 cases, for which the panel judged if a clinical deterioration occurred after a full review of the patient's medical file. The deterioration was also classified by the panel according to preset deterioration categories that are listed in Supplementary Material, http://links.lww.com/JPS/A501, http://links.lww.com/JPS/A502. Then, after accounting for 3 false alarm calls, the final data set included 218 deterioration cases.

Second, these cases were complemented by 146 "nonevent" patient cases where no deterioration event had occurred during hospitalization, which were randomly chosen from a pool of patients admitted in the same study period, to the same wards and discharged home after a hospital stay between 3 and 7 days (5 days being the median length of stay of patients admitted to the included general wards).

This constructed data set was later split into 3 separate parts that were used respectively for the training, validation, and testing of the model. We used an oversampling algorithm (SMOTE)[29] to balance the data set distribution, after data set splitting. Figure 1 illustrates the data set selection and inclusion steps.

Explanatory Variables (Model Features)

Forty-four explanatory variables (model features) available in the EMR that could potentially be early predictors of the patient deterioration outcome were identified by the expert panel based on a literature review[5,30-42] of the predictors of most common in-hospital clinical deterioration situations. These variables included demographic patient data (e.g., age and sex), laboratory values (absolute value and difference from the previous value, noted Δ), and use of specific medical devices or interventions on the patient (such as bilevel positive airway pressure, mechanical ventilation) but did not include vital signs. The complete list of variables is listed in the Supplementary Digital Content 1, http://links.lww.com/JPS/A501.

Measurement and Prediction Timing

Several time points for prediction were considered to account for the model's time dependency. Time of prediction, T_p, was defined as the time before T_0 at which the prediction was generated, where T_0 is the time of the deterioration event.

For each patient, we selected measurements (values of explanatory variables) at the following prediction time points T_p: T_0–3h, T_0–6h, T_0–12h, T_0–18h, T_0–24h, T_0–30h, T_0–36h, T_0–42h, and T_0–48h. These prediction time points were chosen based on the frequency of patient clinical reevaluation (every 6–8 hours) adopted for noncritical wards in clinical practice recommendations[43] and observed in most hospitals. For nondeteriorating patients, T_0 was set as the time of discharge.

For each prediction time point (T_p), the most recent value relative to T_p of each explanatory variable was measured and documented, all the way up to 3 days (72 hours) before T_p, in line with similar studies.[12] This interval between T_p and T_p–72h will be called the explanatory variables collection (or sampling) window. In fact, this window was chosen to be wide enough to take into consideration the values of different laboratory examinations that are not necessarily ordered by the medical team in the same day nor repeated with the same frequency as per clinical guidelines.[44] At the same time, that same window should be sufficiently limited

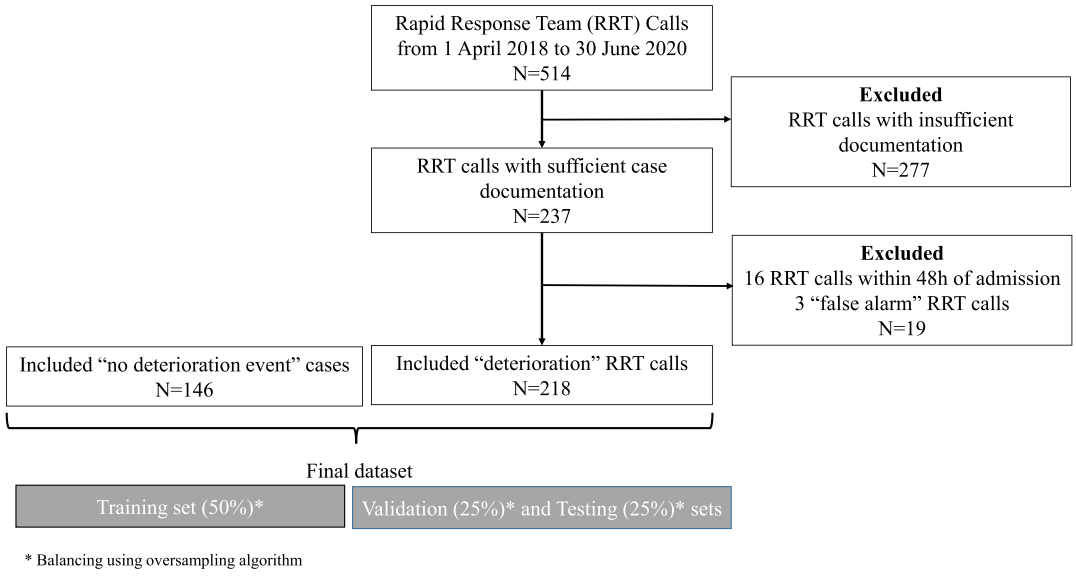


FIGURE 1. Flowchart of cases recruitment and data set construction.

in time (3 days) not to exceed the maximal predictive horizon of physiological biomarkers in the literature[18,19] relative to clinical deterioration (72 hours), hence close enough to the prediction time point so that the contained values of the requested exams can still be associated with the physiological and clinical status of the patient at the time of prediction.

Differential (δ) variables (e.g., ΔC-reactive protein) were defined as the difference between the available value of the variable closest to T_p in time and the available value of the variable furthest from T_p in time, all within the explanatory variables collection window.

Missing values among any explanatory variable in the window were imputed by using the mean value of the same variable over the entire cohort in the same time window. An illustration of the prediction timeline and its associated concepts can be found in Figure 2.

Model Training, Validation, and Testing/Algorithms

The Python programming language was used for developing the scripts to create and analyze the models. A supervised learning approach was adopted using different ML algorithms: random forests (RF), gradient boosting, artificial neural networks (ANN), and logistic regression (LR). We used the implementation from the Sklearn Python module for RF and LR, XGBoost for Gradient Boosting, and Keras for ANN.

Fifty percent of the data set was used as a training set, and the rest of the data set was equally split to be used for validation and testing using a 5-fold cross-validation.

Outcomes and Evaluation Metrics

The area under the receiver operating curve (AUROC) and the F_1 score (defined as the harmonic mean of the precision and recall of the model outcome) were used for reporting the performance results of the different algorithms for each class of deterioration, calculated on the basis of a "one-versus-rest" approach.

To identify the important predictors of the model, variable importance was determined by calculating the relative influence of each explanatory variable on the algorithm classification results using the Python Sklearn library (Python Software Foundation, Python Language Reference, version 3.7).

The model parameters were fine tuned for the different algorithms using only the training and validation data sets (not the

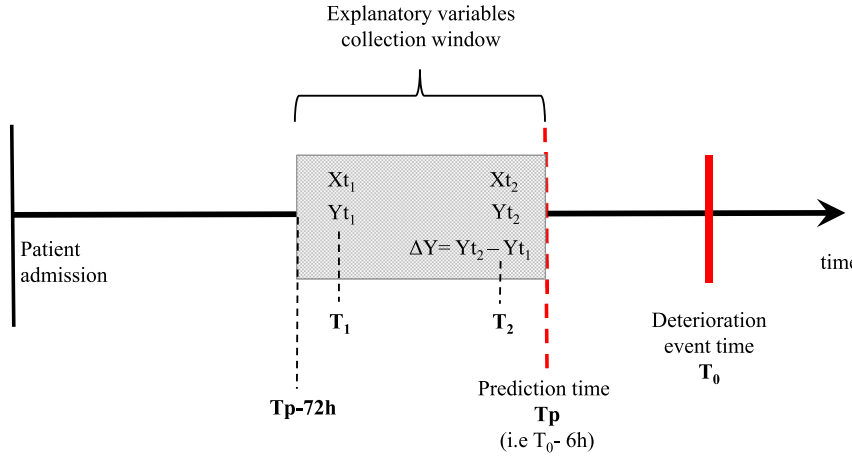


FIGURE 2. Timeline for prediction.

TABLE 1. Distribution of the Clinical Outcomes of the Deterioration Cases Included in the Model

Deterioration Type (Typical Examples)	No. Cases Per Outcome Type					
	Stabilized on Floor	Transfer to ICU	Code Blue	Not for Resuscitation	Total Cases	Percentage
Cardiological (atrial fibrillation, tachyarrythmia, supraventricular tachycardia, cardiac infarct)	38	8			46	21.1%
Pneumonia (pneumonia, aspiration pneumonia, pneumonitis, bronchiolitis)	32	21			52	23.9%
Pulmonary edema/fluid overload (heart failure decompensation, fluid overload)	16	4		1	21	9.6%
Sepsis (sepsis/severe sepsis/septic shock)	25	31	3	2	62	28.4%
Hepatic/pancreatic failure (hepatic encephalopathy)	5	4			9	4.1%
Hypovolemia/hypovolemic shock (hemorrhage)	3	6			9	4.1%
Other (hospital induced/acquired conditions including hypoglycemia, medication errors/adverse effects, etc)	9	8	1		16	7.3%
Grand total	**128**	**82**	**4**	**3**	**218**	**100.0%**

testing data set) and using specific tools in the Python Sklearn model selection library, such as GridSearchCV.

RESULTS

Descriptive Statistics

Patient deterioration events in the study occurred in the following hospital departments: internal medicine (48%), infectious diseases (22%), and medicosurgical (30%).

The deterioration case distribution by diagnosis and clinical outcome distribution (stabilization on floor, transfer to ICU, code blue, not for resuscitation) of the different deterioration cases by class are shown in Table 1.

Model Performance

Performance of the various algorithms was calculated and depicted in Table 2. The best performance was achieved with the RF algorithm, with a maximal AUROC of 0.90 and F_1 score of

TABLE 2. Algorithms Performance Versus Prediction Time

Algorithm/No. Test Cases	Model Parameters	Metrics	T_0– 3h	T_0– 6h	T_0– 12h	T_0– 18h	T_0– 24h	T_0– 30h	T_0– 36h	T_0– 42h	T_0– 48h
RF classifier (n = 108)	(n_estimators = 600, criterion = "entropy," max_depth = 12, min_samples_leaf = 2,min_samples_split = 4)	Precision (deterioration/ no deterioration)	0.81/ 0.80	0.85/ 0.85	0.81/ 0.80	0.85/ 0.79	0.8/ 0.77	0.76/ 0.77	0.74/ 0.78	0.8/ 0.75	0.71/ 0.74
		Recall (deterioration/ no deterioration)	0.79/ 0.80	0.85/ 0.85	0.79/ 0.80	0.77/ 0.87	0.75/ 0.81	0.77/ 0.75	0.79/ 0.72	0.74/ 0.81	0.75/ 0.70
		F_1 score	0.81	0.85	0.81	0.82	0.78	0.76	0.75	0.77	0.73
		AUROC score	0.87	0.9	0.88	0.87	0.88	0.87	0.83	0.82	0.78
Boosting classifier (XGBoost, n = 108)	(max_depth = 12, learning_ rate = 0.01, gamma = 0, min_child_weight = 1, n_estimators = 600)	Precision (deterioration/ no deterioration)	0.75/ 0.76	0.84/ 0.79	0.67/ 0.69	0.74/ 0.69	0.76/ 0.70	0.65/ 0.62	0.76/ 0.70	0.65/ 0.65	0.63/ 0.62
		Recall (deterioration/ no deterioration)	0.77/ 0.74	0.77/ 0.85	0.72/ 0.64	0.66/ 0.77	0.66/ 0.79	0.58/ 0.68	0.66/ 0.79	0.64/ 0.66	0.60/ 0.64
		F_1 score (deterioration/ no deterioration)	0.75	0.81	0.68	0.72	0.73	0.63	0.73	0.65	0.62
		AUROC score	0.85	0.86	0.81	0.83	0.85	0.76	0.79	0.73	0.72
ANN (n = 108)	(architecture 20/8/1, loss = "binary_crossentropy," optimizer = "Adam," metrics = ["accuracy"], BS = 43, EPOCH = 4000)	Precision (deterioration/ no deterioration)	0.74/ 0.74	0.75/ 0.71	0.71/ 0.66	0.76/ 0.73	0.75/ 0.61	0.79/ 0.70	0.62/ 0.69	0.71/ 0.74	0.69/ 0.69
		Recall (deterioration/ no deterioration)	0.74/ 0.74	0.68/ 0.77	0.60/ 0.75	0.72/ 0.77	0.45/ 0.85	0.64/ 0.83	0.75/ 0.55	0.75/ 0.70	0.68/ 0.70
		F_1 score (deterioration/ no deterioration)	0.74	0.73	0.68	0.75	0.65	0.74	0.65	0.73	0.69
		AUROC score	0.78	0.78	0.82	0.79	0.76	0.8	0.72	0.78	0.75
LR (n = 108)	(penalty = "l2," dual = False, tol = 0.0001, C = 1, fit_intercept = False, intercept_scaling = 1, class_weight = "balanced," random_state = None, solver = "lbfgs," max_iter = 30000, warm_start = False, n_jobs = None, l1_ratio = None)	Precision (deterioration/ no deterioration)	0.85/ 0.70	0.90/ 0.71	0.70/ 0.86	0.86/ 0.79	0.88/ 0.74	0.87/ 0.77	0.78/ 0.76	0.80/ 0.74	0.74/ 0.71
		Recall (deterioration/ no deterioration)	0.62/ 0.89	0.62/ 0.93	0.70/ 0.86	0.77/ 0.88	0.68/ 0.91	0.73/ 0.89	0.75/ 0.79	0.71/ 0.82	0.70/ 0.75
		F_1 score (deterioration/ no deterioration)	0.76	0.78	0.78	0.82	0.79	0.81	0.77	0.77	0.72
		AUROC score	0.81	0.85	0.82	0.87	0.86	0.88	0.83	0.81	0.8

0.85 obtained at prediction time T_0–6h. This slightly decreases but is still acceptable at T_0–42h, with an AUROC of 0.82 and an F_1 score of 0.77.

Explanatory Variables' Importance

Explanatory variables' importance for the RF model was calculated and represented in Figure 3, using a "heatmap" representation warm-to-cool color scheme, with the warm colors representing high-value impact of the variable and the cool colors representing a low-value impact.

The most contributing variables to the prediction result (in decreasing order) were the following: CRP, lymphocytes count, sodium minus chloride, sodium differential, alkaline reserve differential, age, blood urea nitrogen differential, potassium differential, and neutrophil-to-lymphocyte ratio. In addition, we illustrated in Supplemental Material 2, http://links.lww.com/JPS/A502, one example (among others) of a logical visualization of the decision-making process of the model using the decision tree algorithm at T_0–12h, showing the previously mentioned variables and the model chosen thresholds.

Benchmark Against Other All-Cause Deterioration Models

Benchmarks to track-and-trigger (vital signs based) deterioration prediction models and to other more hybrid deterioration prediction models (vital signs, laboratory values, patient demographics, diagnosis, etc) from the literature are given in Table 3, both in terms of performance metrics, outcome variables, and best time to prediction.

The prediction model showed an earlier prediction horizon (up to 42 hours) with acceptable performance (AUROC, >0.8), relative to most track-and-trigger systems (6–24 hours), but also most hybrid all-cause deterioration prediction systems (12–48 hours).

The F_1 score (and specifically the positive predictive value) of the model is good (>0.8) and scored better than most track-and-trigger models, which could mean in practice a lower rate of false alarms generated.

DISCUSSION

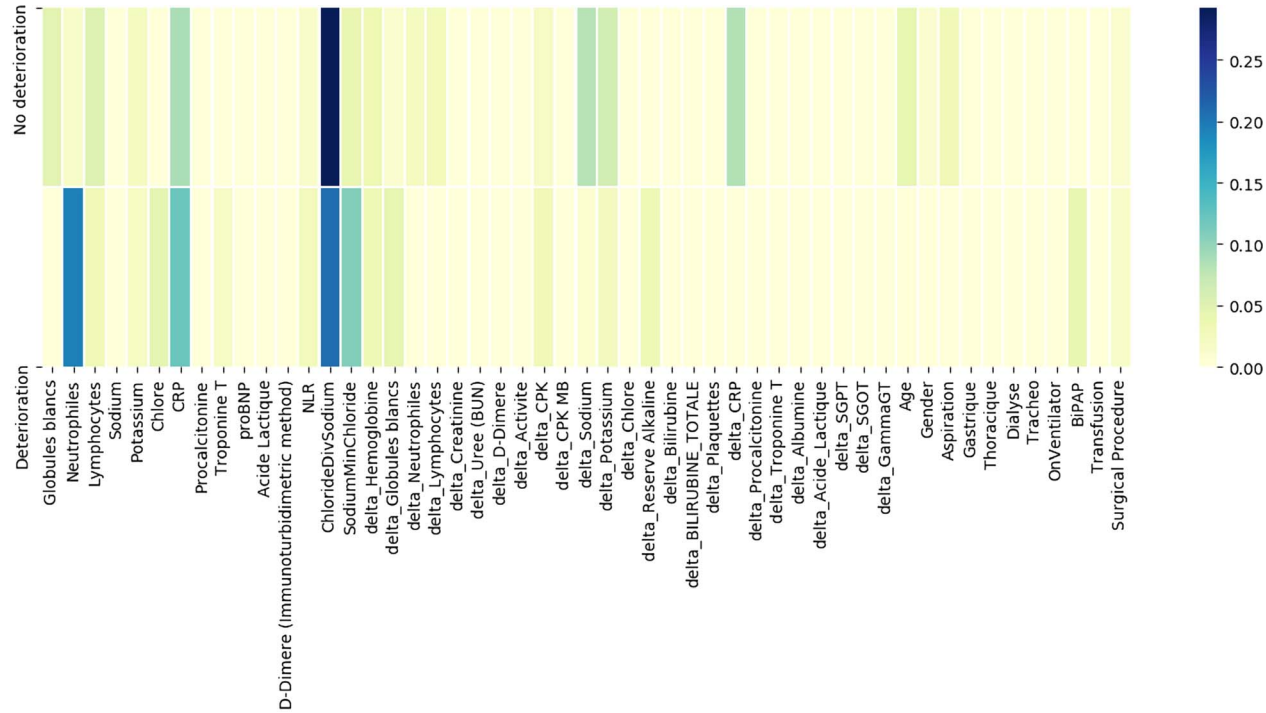
In this retrospective, single-center study, we developed and evaluated an ML model for the prediction of all-cause patient deterioration. The model's explanatory variables were mainly biomarkers values routinely available in basic EMRs, without inclusion of vital signs data.

1)Potential use of the model for predicting clinical deterioration and supporting clinical decision making

If transformed into an automated clinical decision support tool and applied systematically to all hospital inpatients, this model could potentially stratify patients based on their deterioration risk score and proactively alert the healthcare team of patients possibly at high risk of deterioration within the next hours per days. The update or refreshing of the model data prediction result would basically rely on the arrival of new laboratory data, thus on the frequency of blood sample extraction, which in practice can range from 12 to 48 hours for most in-hospital patients.

This prediction is based on the capture of a rich "physiological picture" (mainly through biomarkers), which precedes chronologically the "clinical picture" (captured by track-and-trigger models, through observation of vital signs and clinical exam), hence an earlier prediction of deterioration.

This earlier prediction (up to 42 versus 6–12 hours for track-and-trigger models) can give the healthcare team a window of opportunity to try to stabilize or manage at-risk patients on



(Globules blancs: White Blood Cells, Chlore: Chloride, NLR: Neutrophils to Lymphocytes Ratio, ChlorideDivSodium: Chloride to Sodium ratio, Uree: Urea, Activite: Partial Thromboplastin Time ratio, Reserve Alcaline: blood Bicarbonate, Plaquettes: Platelets)

FIGURE 3. Features importance by deterioration class (RF classifier), prediction at T_0–6h.

TABLE 3. Benchmark Relative to a Number of Recent Studies and Reviews With Similar Scope

Model Category	Study/Model Name	Study Phase	Study Type	Statistical Methods Used for Model Derivation	Prediction Performance	Types of Variables Used	Outcome Measure	Prediction Horizon (Window)
Track-and-trigger models (vital signs based)	Campbell et al[45] (2020)/ Q-ADDS	Prediction model performance benchmark	Retrospective single-center cohort	Clinical consensus based	0.71 (AUC)	Vital signs	Death/unanticipated admission to intensive care	30 h
	Kia et al[46] (2020)/ MEWS++	Prediction model validation	Retrospective single-center cohort	ML algorithms	0.85 (AUROC)	Vital signs	Death/unanticipated admission to intensive care	6 h
	Kirkland et al[20] (2013)	Prediction model validation	Retrospective single-center cohort	Multivariate regression analysis	0.71 (AUROC)	Vital signs, Braden score, fall risk score	RRT activation	2–12 h
	Cho et al[47] (2020)	Automated system performance benchmark	Retrospective single-center cohort	ML algorithms	0.86 (AUC)	Vital signs	Cardiac arrest/ unanticipated admission to intensive care	0.5–24 h
	Gerry et al[23] (2020)		Systematic review	AI and non-AI algorithms	0.55 to 0.96 (C-index)	Vital signs	Death/unanticipated admission to intensive care	24 h/ inpatient stay
	Fu et al[48] (2020)		Systematic review	AI and non-AI algorithms	0.71–0.96 (AUC)	Vital signs	Death/unanticipated admission to intensive care	24 h/ inpatient stay
	Peelen et al[21] (2020)		Systematic review	AI and non-AI algorithms	0.65–0.93 (AUC)	Vital signs	RRT activation, cardiopulmonary resuscitation, unanticipated transfer to an ICU, or death	2–24 h
	Muralitharan et al[26] (2021)		Systematic review	ML algorithms	0.57 to 0.97 (AUC)	Vital signs	Cardiac arrest/death/ unanticipated admission to intensive care	4–24 h
Hybrid deterioration prediction models (vital signs, biomarkers, and patient demographics data)	Jefferey et al[49] (2018)	Prediction model validation	Retrospective single-center cohort	ML algorithms	0.85 (AUROC) 0.27 (F_1 score)	Vital signs, laboratory tests, *ICD-10* diagnosis, demographic data	Cardiopulmonary arrest	48 h
	Churpek et al[16] (2016)/ eCART	Prediction model validation	Retrospective multicenter cohort	ML algorithms	0.77(AUC)	Vital signs, laboratory tests, demographic data	Cardiac arrest/death/ unanticipated admission to intensive care	24 h
	Kipnis et al[12] (2016)/ AAM	Evaluation of implemented system	Retrospective multicenter cohort	Discrete-time LR	0.82 (AUC)	Vital signs, laboratory tests, severity of illness, comorbidity index, demographic data	Unanticipated admission to intensive care	12–24 h
	Pimentel et al[50] (2021)/ HAVEN	Evaluation of implemented system	Retrospective multicenter cohort	ML algorithms	0.90 (AUC)	Vital signs, laboratory tests, comorbidities index, frailty	Cardiac arrest/ unanticipated admission to intensive care	24–48 h
	Blackwell et al[5] (2020)	Prediction model validation	Retrospective single-center cohort	Multivariate regression analysis	0.71–0.84 (AUC) depending on outcome	Vital signs, laboratory tests and continuous 7-lead electrocardiogram signal	Unanticipated admission to intensive care	12 h

ICD-10, International Classification of Diseases – version 10.

general wards, preventing as much as possible their transfer to the ICUs or any further escalation in care. This information can also permit the medical and nursing team to selectively increase surveillance for patients at high risk of deterioration, hence trying to prevent or promptly mitigate expected deterioration events. In the context of a global shortage of health workers, this

information can help in focusing resources on the patients that need those the most.

A complementary use of such a model can be for patient safety professionals in hospitals, who can make use of the prediction data on a daily basis to audit and verify the follow-up and safety actions taken by the healthcare team to manage the deterioration risks, including suitability of the level of care provided to the clinical status of the patient.

2)Model explainability and the road toward clinical validation and clinician adoption

Explainability or the possibility to understand the model's classification logic is an important feature that can facilitate the "clinical interpretation" of the results by the clinicians.

In this study, the deterioration model permits a certain level of "explainability" for most algorithms applied and in particular RF and decision tree, in the sense that it is possible to identify the main variables that influence most the model prediction results, along with their respective weights. Further explainability can be obtained with decision tree algorithm (example in Fig. 3) where a visualization of the decision tree could be obtained, showing the logic behind the classification (Supplementary Digital Material 2, http://links.lww.com/JPS/A502).

Such data insight can help users understand the prediction results and facilitate any future effort to clinically interpret and validate the model by an experienced panel of physicians. This "clinical validation" is an important step toward the practical adoption of the model by clinicians, where the latter are often reluctant to use "black box" models, even when they show good results.

3)Model specificities relative to other predictive models and possible impact on results

While most deterioration models in the literature were derived from cases with specific outcomes of cardiac arrests, death, and unplanned transfer to ICU (Table 3), the model elaborated in this study was trained and validated on deterioration cases linked to RRT activations that were confirmed by a panel of clinical experts. It is to be noted that RRT activation cases depict a broader image of clinical deterioration, because they include an additional outcome in clinical practice, which is the patient stabilization on the floor, amounting to almost half of deterioration cases (Table 1), in addition to the classical previously mentioned outcomes.

Furthermore, almost all of the deterioration models in the literature, which include laboratory variables (such as, for example, those of the LAPS-2 score[12]) use the absolute form of the exam values. To the best of our knowledge,[21] our model is among a few (if not the only one) that use differential (or δ) biomarker variables in deterioration prediction models. It is known, however, that changes in biomarker values (δ) within a specific timeframe can indicate certain underlying pathophysiological changes, such as, for example, in case of bleeding (δ in hemoglobin values) or acute kidney injury (δ in creatinine values).

The analysis of variable importances (Fig. 3) shows that a number of differential variables (e.g., ΔSodium, ΔPotassium, ΔCRP) have a significant weight in the model prediction function.

We believe that the results of the prediction model were impacted to a certain extent by these specificities but also the broad choice of biomarkers that intended to cover multiple deterioration mechanisms that are common to various deterioration etiologies. These mechanisms include but are not limited to respiratory and metabolic acidosis/alkalosis, systemic inflammation, electrolyte imbalance, volume imbalance, and hypoperfusion/ischemia.

Finally, we believe that the exclusion of vital signs data from the model might have in a certain way contributed to an earlier prediction horizon. In fact, in pathophysiological processes leading to clinical deterioration, changes in biomarkers usually occur hours before clinical signs and symptoms. Furthermore, even in hybrid models (where variables comprise vital signs, laboratory data, and other patient data), the importance of biomarkers could have been eclipsed by the direct association (however, late in matter of prediction) between the occurrence of clinical signs (vitals) with the deterioration event outcome. Further research might be needed to better elucidate the relation between the choice of variable type and the impact this has on the prediction horizon of clinical deterioration models.

LIMITATIONS

The study was conducted in a single center, which might have amplified the effect of certain factors on the results, such as the quality of the medical documentation and the specific practice of exam prescriptions for diagnosis and monitoring. An external and a prospective validation of the study model should be a undertaken to understand its performance in a real clinical context, before it can be implemented as a clinical decision support system.

In addition, the number of deterioration events per explanatory variable is relatively small, which might have impacted to a certain extent the performance metrics and the statistics of the variables' importance. This is due to the limited sample size of the study. However, it corresponded to almost 2 years of systematic data collection of deterioration events in our hospital and a thorough and time consuming validation by an expert panel of the cases outcome.

CONCLUSIONS

We have developed and validated an explainable prediction model for inpatient deterioration in general wards, trained on expert validated deterioration events with RRT activation. The model is mainly based on biomarkers, without use of vital sign data. The model performed better than most criterion standard track-and-trigger systems, both in prediction performance and prediction horizon. Such a model can also be suitable for hospitals with limited resources and a basic EMR. Further increase of the data sample could contribute to improving its performance, and the model would gain to be externally and prospectively validated.

ACKNOWLEDGMENTS

The authors thank Dr Charbel Mourad for his kind and critical review of the manuscript.

REFERENCES

1. Cardoso LT, Grion CM, Matsuo T, et al. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Crit Care*. 2011;15:R28.
2. Sankey CB, McAvay G, Siner JM, et al. "Deterioration to door time": an exploratory analysis of delays in escalation of care for hospitalized patients. *J Gen Intern Med*. 2016;31:895–900.
3. Churpek MM, Wendlandt B, Zadravecz FJ, et al. Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. *J Hosp Med*. 2016;11:757–762.
4. Brown H, Terrence J, Vasquez P, et al. Continuous monitoring in an inpatient medical-surgical unit: a controlled clinical trial. *Am J Med*. 2014; 127:226–232.
5. Blackwell JN, Keim-Malpass J, Clark MT, et al. Early detection of in-patient deterioration: one prediction model does not fit all. *Crit Care Explor*. 2020;2:e0116.
6. Rose MA, Hanna LA, Nur SA, et al. Utilization of electronic modified early warning score to engage rapid response team early in clinical deterioration. *J Nurses Prof Dev*. 2015;31:E1–E7.

7. NICE TC for CP at: acutely ill patients in hospital: Recognition of and response to acute illness in adults in hospital. 2007. https://www.nice.org.uk/guidance/CG50. Accessed June 14, 2021.
8. Subbe CP, Kruger M, Rutherford P, et al. Validation of a modified early warning score in medical admissions. *QJM*. 2001;94:521–526.
9. Hall KK, Lim A, Gale B. The use of rapid response teams to reduce failure to rescue events: a systematic review. *J Patient Saf*. 2020;16:S3–S7.
10. HIMSS Analytics–North America. Electronic medical record adoption model. Available at: https://www.himssanalytics.org/emram. Accessed Jun 30, 2021.
11. Kose I, Rayner J, Birinci S, et al. Adoption rates of electronic health records in Turkish hospitals and the relation with hospital sizes. *BMC Health Serv Res*. 2020;20:967.
12. Kipnis P, Turk BJ, Wulf DA, et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. *J Biomed Inform*. 2016;64:10–19.
13. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. *J Hosp Med*. 2016;11:S18–S24.
14. Kamio T, Van T, Masamune K. Use of machine-learning approaches to predict clinical deterioration in critically ill patients: a systematic review. *Int J Med Res Heal Sci*. 2017. Available at: https://www.ijmrhs.com/medical-research/use-of-machinelearning-approaches-to-predict-clinical-deterioration-in-critically-ill-patients-a-systematic-review.pdf. Accessed May 23, 2021.
15. Jeffery AD, Dietrich MS, Fabbri D, et al. Advancing in-hospital clinical deterioration prediction models. *Am J Crit Care*. 2018;27:381–391.
16. Churpek MM, Yuen TC, Winslow C, et al. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med*. 2016;44:368–374.
17. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001;69:89–95.
18. Orfanu A, Aramă V, Popescu C, et al. The dynamical assessment of inflammatory biomarkers in predicting the outcome of septic patients and the response to antimicrobial therapy. *J Crit Care Med (Targu Mures)*. 2020;6:25–31.
19. Kavsak PA, Hill SA, Supapol WB, et al. Biomarkers for predicting serious cardiac outcomes at 72 hours in patients presenting early after chest pain onset with symptoms of acute coronary syndromes. *Clin Chem*. 2012;58:298–302.
20. Kirkland LL, Malinchoc M, O'Byrne M, et al. A clinical deterioration prediction tool for internal medicine patients. *Am J Med Qual*. 2013;28:135–142.
21. Peelen RV, Eddahchouri Y, Koeneman M, et al. Algorithms for prediction of clinical deterioration on the general wards: a scoping review. *J Hosp Med*. 2021;16:612–619.
22. Mann K, Good N, Fatehi F, et al. Predicting patient deterioration: a review of tools in the digital hospital setting. *J Med Internet Res*. 2021;23:e28209. doi:10.2196/28209.
23. Gerry S, Bonnici T, Birks J, et al. Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ*. 2020;369:m1501. doi:10.1136/bmj.m1501.
24. Sørensen EM, Petersen JA. Performance of the efferent limb of a rapid response system: an observational study of medical emergency team calls. *Scand J Trauma Resusc Emerg Med*. 2015;23:69.
25. Tirkkonen J, Tamminen T, Skrifvars MB. Outcome of adult patients attended by rapid response teams: a systematic review of the literature. *Resuscitation*. 2017;112:43–52.
26. Muralitharan S, Nelson W, Di S, et al. Machine learning–based early warning systems for clinical deterioration: systematic scoping review. *J Med Internet Res*. 2021;23:e25187.
27. Jones D, Mitchell I, Hillman K, et al. Defining clinical deterioration. *Resuscitation*. 2013;84:1029–1034.
28. Padilla RM, Mayo AM. Clinical deterioration: a concept analysis. *J Clin Nurs*. 2018;27:1360–1368.
29. Fernández A, García S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J Artif Intell Res*. 2018;61:863–905.
30. Fleuren LM, Klausch TLT, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med*. 2020;46:383–400.
31. Zeiberg D, Prahlad T, Nallamothu BK, et al. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS One*. 2019;14:e0214465.
32. Yang P, Wu T, Yu M, et al. A new method for identifying the acute respiratory distress syndrome disease based on noninvasive physiological parameters. *PLoS One*. 2020;15:e0226962.
33. Le S, Pellegrini E, Green-Saxena A, et al: Supervised machine learning for the early prediction of acute respiratory distress syndrome (ARDS). *J Crit Care* 2020;60:96-102.
34. Ding XF, Li JB, Liang HY, et al. Predictive model for acute respiratory distress syndrome events in ICU patients in China using machine learning algorithms: a secondary analysis of a cohort study. *J Transl Med*. 2019;17:326.
35. Xiong W, Xu M, Zhao Y, et al. Can we predict the prognosis of COPD with a routine blood test? *Int J Chron Obstruct Pulmon Dis*. 2017;12:615–625.
36. Hyland SL, Faltys M, Hüser M, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med*. 2020;26:364–373.
37. Nemati S, Holder A, Razmi F, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46:547–553.
38. Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform*. 2016;4:e28.
39. Tabak YP, Sun X, Nunez CM, et al. Using electronic health record data to develop inpatient mortality predictive model: Acute Laboratory Risk of Mortality Score (ALaRMS). *J Am Med Informatics Assoc*. 2014;21:455–463.
40. Escobar GJ, Greene JD, Scheirer P, et al. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care*. 2008;46:232–239.
41. Karakioulaki M, Stolz D. Biomarkers in pneumonia-beyond procalcitonin. *Int J Mol Sci*. 2019;20.
42. Gori CS, Magrini L, Travaglino F, et al. Role of biomarkers in patients with dyspnea. *Eur Rev Med Pharmacol Sci*. 2011;15:229–240.
43. ACSQH: national consensus statement: essential elements for recognising and responding to acute physiological deterioration (second edition). 2017. Available at: https://www.safetyandquality.gov.au/our-work/recognising-and-responding-deterioration/recognising-and-responding-physiological-deterioration/national-consensus-statement-essential-elements-recognising-and-responding-acute-physiological-deterioration. Accessed July 2, 2021.
44. Ambasta A, Pancic S, Wong BM, et al. Expert recommendations on frequency of utilization of common laboratory tests in medical inpatients: a Canadian consensus study. *J Gen Intern Med*. 2019;34:2786–2795.

45. Campbell V, Conway R, Carey K, et al. Predicting clinical deterioration with Q-ADDS compared to NEWS, Between the Flags, and eCART track and trigger tools. *Resuscitation*. 2020;153:28–34.

46. Kia A, Timsina P, Joshi HN, et al. MEWS++: enhancing the prediction of clinical deterioration in admitted patients through a machine learning model. *J Clin Med*. 2020;9:343.

47. Cho K-J, Kwon O, Kwon J, et al. Detecting patient deterioration using artificial intelligence in a rapid response system. *Crit Care Med*. 2020;48:e285–e289.

48. Fu LH, Schwartz J, Moy A, et al. Development and validation of early warning score system: A systematic literature review. *J Biomed Inform*. 2020;105:103410.

49. Jeffery AD, Dietrich MS, Fabbri D, et al. (2018). Advancing in-hospital clinical deterioration prediction models. *Am J Crit Care*. 2018;2018:381–391.

50. Pimentel MA, Redfern OC, Malycha J, et al. Detecting deteriorating patients in hospital: development and validation of a novel scoring system. *Am J Respir Crit Care Med*. 2021;204:44–52.

Summary Table

What was already known on the topic

- "Track and trigger" systems, based on regular monitoring of patient vital signs values, are still considered "criterion standard" in hospitals, but predict clinical deterioration only up to 12–24 hours maximum.
- Machine learning algorithms have been used recently for the derivation of clinical deterioration models and showed to improve detection performance.
- Most derived clinical deterioration have been trained or derived relative to the outcome of arrest/death/unplanned transfer to intensive care and very few on the broader outcome given by RRT activations.
- Biomarkers' dynamics can predict physiological change up to 72 hours before a deterioration event and could thus be an earlier predictor of deterioration than vital signs.

What this study added to our knowledge

- Using an ML model based mainly on biomarkers (no vital signs inclusion), and trained on RRT deterioration cases can permit a good all-cause deterioration prediction performance and prediction horizon in hospitalized patients on noncritical wards.
- Such models can be used in hospitals with basic EMRs (basic ancillary systems, no integration of vital signs) and could play an important role in helping clinicians stratify the deterioration risk of hospitalized patients and engage interventions accordingly.