Development and Validation of an Early Warning Model for Hospitalized COVID-19 Patients: A Multi-center Retrospective Cohort Study

J. M. Smit (✉ j.smit@erasmusmc.nl)

Dept. of Intensive Care, Erasmus University Medical Center, Rotterdam

J. H. Krijthe

EEMCS, Pattern Recognition & Bio-informatics group, Delft University of Technology, Delft

A. N. Tintu

Dept. of Clinical Chemistry, Erasmus University Medical Center, Rotterdam

H. Endeman

Dept. of Intensive Care, Erasmus University Medical Center, Rotterdam

J. Ludikhuize

Dept. of Intensive Care, Haga Teaching Hospital, The Hague

M. E. Van Genderen

Dept. of Intensive Care, Erasmus University Medical Center, Rotterdam

S. Hassan

Dept. of Clinical Epidemiology, Leiden University Medical Center, Leiden

R. El Moussaoui

Dept. of Internal Medicine, Maasstad Teaching Hospital, Rotterdam

P. E. Westerweel

Dept. of Internal Medicine, Albert Schweitzer Teaching Hospital, Dordrecht

R. J. Goekoop

Dept. of Rheumatology, Haga Teaching Hospital, The Hague

G. Waverijn

Team Business Intelligence, Maasstad Teaching Hospital, Rotterdam

T. Verheijen

Dept. of Rheumatology, Leiden University Medical Center, Leiden

J. G. Den Hollander

Dept. of Internal Medicine, Maasstad Teaching Hospital, Rotterdam

M.G.J. De Boer

Dept. of Infectious Diseases, Leiden University Medical Center, Leiden

D. A.M.P.J. Gommers

Dept. of Intensive Care, Erasmus University Medical Center, Rotterdam

R. Van der Vlies

Team Business Intelligence, Albert Schweitzer Teaching Hospital, Dordrecht

M. Schellings

Dept. Of Clinical Chemistry, MaasstadLab, Maasstad Teaching Hospital, Rotterdam

R. A. Carels

Dept. of Internal Medicine, Ikazia Teaching Hospital, Rotterdam

C. Van Nieuwkoop

Dept. of Internal Medicine, Haga Teaching Hospital, The Hague

M. S. Arbous

Dept. of Intensive Care, Leiden University Medical Center, Leiden

J. Van Bommel

Dept. of Intensive Care, Erasmus University Medical Center, Rotterdam

R. Knevel

Dept. of Rheumatology, Leiden University Medical Center, Leiden

Y. B. De Rijke

Dept. of Clinical Chemistry, Erasmus University Medical Center, Rotterdam

M. J.T. Reinders

EEMCS, Pattern Recognition & Bio-informatics group, Delft University of Technology, Delft

Research Article

Keywords:

Posted Date: March 1st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1380907/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background: Timely identification of deteriorating COVID-19 patients is needed to guide changes in clinical management and admission to intensive care units (ICUs). There is significant concern that widely used early warning scores (EWSs) underestimate illness severity in COVID-19 patients and therefore, we developed an early warning model specifically for COVID-19 patients.

Methods: We collected electronic medical record data to extract predictors and used these to fit a random forest model. To simulate the situation in which the model would have been developed after the first COVID-19 'wave' in the Netherlands and implemented during the second wave, we performed a temporal validation by splitting all included patients into groups admitted before and after August 1, 2020. Furthermore, we propose a method for dynamic model updating to retain the model's predictive performance over time. We evaluated model discrimination and calibration, performed a decision curve analysis, and quantified the importance of predictors using SHapley Additive exPlanations values.

Findings: We included 3 514 COVID-19 patient admissions from six Dutch hospitals between February 2020 until May 2021, and included a total of 18 predictors for model fitting. The model showed a higher discriminative performance in terms of partial area under the receiver operating characteristic curve (0.82 [0.80 to 0.84]) compared to the National Early Warning Score (0.72 [0.69 to 0.74]) and the Modified Early Warning Score (0.67 [0.65 to 0.69]), a greater net benefit over a range of clinically relevant model thresholds, and good calibration (intercept=0.03 [-0.09 to 0.14], slope=0.79 [0.73 to 0.86]).

Interpretation:

This study shows the potential benefit of moving from early warning models for the general inpatient population to models for specific patient groups. The COVID-19-specific early warning model is available online at https://github.com/jimmsmit/COVID-19_EWS and we encourage others to further validate this model independently.

1. Introduction

The COVID-19 pandemic has continued to put pressure on hospital care worldwide. As COVID-19 patients may deteriorate rapidly and unexpectedly, timely identification of deterioration is needed to guide changes in clinical management, e.g., admission to intensive care units (ICUs). Widely used early warning scores (EWSs) based on aggregate-weighted vital signs have been developed for this purpose already, i.e. the Modified Early Warning Score (MEWS),¹ National Early Warning Score (NEWS)^{2,3} and its successor, NEWS2.⁴ A recent systematic review⁵ showed that NEWS2 has been validated for COVID-19 patients in various studies. However till date, only one study⁶ validated it for the purpose it was originally designed for, namely longitudinal monitoring to identify clinical deterioration over a 24-hour interval. Moreover, these existing EWSs were designed for the general inpatient population and do not differentiate between various rates of oxygen delivery. Consequently, there is significant concern that these scores underestimate severity of illness in COVID-19 patients.^{7,8,9,10} Many prognostic models for COVID-19 have been developed,¹¹ but are mostly intended to predict outcomes at the point of hospital admission instead of longitudinal in-patient monitoring. Moreover, most of these use relatively long or unspecified prediction horizons, whereas for the task of early warning, a prediction horizon limited to a few days is recommended.¹²

We aimed to develop an early warning model for longitudinal monitoring of hospitalized COVID-19 patients, based on patient demographics and vital signs, and benchmark it against existing EWSs.

2. Methods

The Medical Ethics Committee at Erasmus MC, Rotterdam, The Netherlands, waived the need for patient informed consent and approved an opt-out procedure for the collection of COVID-19 patient data during the COVID-19 crisis. The study is reported in accordance with the TRIPOD guidelines.¹³

2.1. Study population and data collection

The study was performed in six hospitals in the Netherlands, South Holland province, consisting of two academic hospitals and four teaching hospitals. We collected electronic medical record (EMR) data from patients admitted with COVID-19, defined as a positive real-time reverse transcriptase polymerase chain reaction (RT-PCR) assay for SARS-CoV-2 or a COVID-19 Reporting and Data System (CO-RADS) score¹⁴ ≥ 4 and clinical suspicion without obvious other causes of respiratory distress. The periods of data collection varied per hospital and ranged between February 2020 and May 2021.

2.2. Outcome

We used patient deterioration as a primary outcome, defined as a composite outcome of intensive care unit (ICU) admission or unexpected death on the ward, within 24 hours from the moment of prediction. We qualified each patient death as unexpected unless it occurred after initiation of end-of-life care (EoLC) or a 'do not admit to ICU' order.

2.3. Participants

We handled patients who returned to the same hospital for COVID-19 related matters, after being discharged first, as separate admissions. We validated the model using the observation set definition,¹⁵ i.e., we collected multiple observation sets ('samples') of each patient at different time points, using the most recently observed set of predictors. We collected samples starting at eight hours after hospital admission and added one every 24 hours until discharge, ICU admission, or death. We labeled samples as positive if ICU admission or death occurred within 24 hours from the moment of sampling, and negative otherwise. We excluded patients (1) who were admitted to the ICU straight from home or the emergency department, (2) who were hospitalized shorter than eight hours, and (3) for who EoLC or a 'do not admit to ICU' order was initiated somewhere during hospitalization. We censored patients who were transferred to other hospitals at the moment of transfer. For patients who were still admitted when the data was collected, we censored at 24 hours before the final observed measurement, consequently excluding still admitted patients who stayed shorter than 24 hours.

2.4. Predictors

As recommended by Wynants and colleagues,¹¹ we selected a set of candidate predictors which were identified as clinically important in COVID-19 patients in the literature (supplemental table 1). Additionally, to effectively model the degree of supplemental oxygen (O_2) a patient required, we added O_2 both as a binary (yes/no) and continuous (L/min) predictor. To measure O_2 relative to the patient's oxygenation, we added the SpO_2 to O_2 ratio (SpO_2/O_2). We added changes (Δ s) in frequently measured vital signs to model their dynamics. Finally, to correct for time dependency of some included predictors and model the effect of duration of the hospitalization on the prior deterioration risk, we added the current length of stay on the ward as a predictor. We excluded predictors with entry densities (i.e. fractions of non-empty daily measurements) less than 50% within the development set. More details on the definitions of the candidate predictors can be found in appendix A.

2.5. Missing data

We imputed the two categorical predictors, i.e. sex and AVPU (Alert, Verbal, Pain, Unresponsive),¹⁶ for each sample separately by fitting logistic regression models with sex or AVPU as outcomes and using the remaining data as predictors. To impute the missing values among the remaining predictors, we used the 'IterativeImputer' function offered by scikit-learn in Python,¹⁷ which imputes each predictor with missing values based on the other predictors with Bayesian ridge regression in an iterated round-robin fashion (appendix B).

2.6. Model development

We fitted a random forest (RF) model to discriminate between positive and negative samples. To examine the added value of the inclusion of non-linear predictor outcome relations by the RF model, we also fitted a logistic regression (LR) model with L2 regularization. First, we normalized the samples by centering each predictor and scaling them by the standard deviation (based on the development set). After imputation, we optimized the 'maximum tree depth' and 'max features' hyperparameters of the RF model and the regularization strength (λ) of the LR model using an exhaustive grid search in a stratified 10-fold cross-validation procedure within the development set (optimizing the area under the receiver operating characteristic curve). Supplemental table 2 shows the hyperparameter grids that were searched. Finally, we fitted the models with the optimized hyperparameters using the development set and validated them using the test set.

2.7. Model validation

We validated the models temporally, in which we evaluated two model implementation strategies: a static and a dynamic strategy. First (static strategy), to simulate the situation in which the models would have been developed after the first COVID-19 'wave' in the Netherlands and implemented during the second wave, we split the data for patients admitted before and after August 1, 2020, forming the development set and test set, respectively. We fitted an RF and LR model using the development set and validated these using the test set, as described in section 2.6 (supplemental Fig. 1) and refer to these as the 'static' models.

Second (dynamic strategy), as changes over time may lead to degraded model performance, we simulated the situation in which models would have been developed after the first wave, implemented, and updated each month during the second wave. Therefore, each month

from August 2020 until May 2021, we updated the static models using patient data that would have been available up to that point and validated these using data of the next month (Fig. 1a). The model updating we implemented was twofold: model fitting and hospital-specific recalibration. The latter was performed with a mapping function (i.e. a calibrator), which we fitted using isotonic regression¹⁸ and which re-maps the predictions of the fitted model (appendix C). Each month, for each hospital, we fitted models using all available data up to that month of the five other hospitals and recalibrated the models using all available data up to that month of the hospital itself. We refer to these as the ‘dynamic’ models. Thus, to validate the model each month, models are updated solely based on data that would have been available up to that month, avoiding any leakage from the development set to the test set. An evaluation of other dynamic model updating strategies can be found in appendix G.

Additionally, we validated an RF and LR model in a more classical way, i.e. retrospectively, using a ‘leave-one-hospital-out’ cross-validation procedure. That is, in each iteration, all patients from five hospitals across the full study period formed the development set and all patients from the remaining hospital formed the test set (supplemental Fig. 2).

2.8. Evaluation metrics

Model discrimination for medical prediction models is often quantified by the area under the receiver operating characteristic curve (AUC). In this setting, we consider a false positive rate (FPR) > 0.33 as clinically undesirable as it will lead to alert fatigue.¹⁹ Moreover, other NEWS validation studies^{20,21} have shown that the recommended triggers (i.e. NEWS = 5 or NEWS = 7²²) appear in the receiver operating characteristic curve for FPRs < 0.33. Therefore, we consider the partial AUC²³ (pAUC, appendix F.2) between 0 and 0.33 FPR as our primary endpoint and the (complete) AUC as a secondary endpoint. Also, the positive predictive value (PPV) is suggested as a useful metric to evaluate the clinical usability of EWSs.^{15,24,20} Therefore, we evaluated the area under the precision-recall curve²⁵ (AUCPR, appendix F.3). We assessed 95% confidence intervals (CIs) to calculate uncertainties around the different metrics using bootstrap percentile confidence intervals²⁶ (with 1000 bootstrap replications stratified for positive and negative samples) and tested the statistical significance of the improvements in discriminative performance between models as described in appendix F.4.

We performed a decision curve analysis (DCA)²⁷ to quantify the clinical utility of the models in terms of net benefit (NB, appendix F.5). We considered early detection of a deteriorating patient as at least four times more important than preventing an unnecessary response (false alarm) and therefore, we plotted the DCA results up to 0.2 deterioration probability. The NB is normalized as the fraction of the maximum NB.

Following the calibration hierarchy defined by Van Calster and colleagues,²⁸ we evaluated model calibration in the ‘weak’ sense by calculating calibration intercepts and slopes²⁹ (appendix F.6) and in the ‘moderate’ sense by plotting smoothed calibration curves.³⁰

Each metric was evaluated using the complete test set (i.e. the overall performance) and using the test subsets from the individual hospitals.

2.9. Explainable predictions

To obtain interpretability for the developed models, we calculated the impact of individual predictors on risk output by SHapley Additive exPlanations (SHAP) values. A SHAP value is a model-agnostic representation of predictor importance, where the impact of each predictor is represented using Shapley values inspired by cooperative game theory.³¹ We calculated SHAP values based on RF and LR models fitted on the complete dataset.

2.10. Comparison with existing Early Warning Scores

To benchmark the models against existing EWSs, we calculated the MEWS¹ and the NEWS² for each sample. We validated both scores in the same fashion (using the same imputation) as the RF and LR models. For the DCA, we transformed the discrete scores into probabilities by fitting two calibrators based on the development data using isotonic regression,¹⁸ with respectively the MEWS or the NEWS as the only predictor.

2.11. Sensitivity analyses

To examine the added value of predictive modeling compared to aggregate-weighted early warning scores, we compared the performance of the MEWS and NEWS with RF and LR models fitted only with the predictors required to calculate the MEWS (i.e. heart rate, respiratory rate, systolic blood pressure, temperature, AVPU) and NEWS (which adds supplemental O_2 (yes/no) and SpO_2). To examine the influence of the included predictors on the model performance, we fitted models using more (by allowing more missingness) and fewer predictors (by

selecting only the top five most important predictors based on mean SHAP magnitude). Finally, we examined the influence of the imputation strategy on the model performance by validating the model for 50 imputation rounds.

3. Results

We included 3 514 COVID-19 patients admissions in six Dutch hospitals within varying time windows ranging between February 2020 until May 2021 (supplemental Fig. 3). Table 1 shows the pathway and population characteristics for all included admissions. Pathway and population characteristics separately for admissions before and after August 1 can be found in supplemental tables 5 and 6 and for the individual hospitals in supplemental tables 7–12. ICU admission occurred in 539 (15.3%), unexpected death in two ($< 0.1\%$), and hospital transfer in 485 (13.8%) admissions (Fig. 1b). Supplemental table 3 shows the occurrence of different patient outcomes across the different hospitals. Occurrence of ICU admission over the whole study period was notably low in hospitals B and C (11.4% and 10.7%) compared to the other hospitals (ranging between 16.3% and 19.4%). We identified 47 candidate predictors of which, after exclusion due to missingness, 18 were included (supplemental table 1). To examine the role of included predictors in relation to the outcome, we plotted cumulative predictor distributions for positive and negative samples (supplemental Fig. 4). Here, O_2 , SpO_2/O_2 , respiratory rate and temperature show notable differences between the positive and negative sample distributions. We plotted the correlations between the predictors (before imputation) in a clustered heatmap (supplemental Fig. 5). The vital signs with their dynamic counterparts (Δ s) showed strong positive correlations, as well as temperature with heart rate. O_2 and respiratory rate showed strong negative correlations with SpO_2/O_2 .

Table 1

Pathway and population characteristics. DA = Discharged alive, ICU=Intensive care unit, SA = Still admitted, IQR = Interquartile range, SD = standard deviation, LOS = Length-of-stay, RR = Respiratory rate, SBP = Systolic blood pressure, T = Temperature, HR = Heart rate.

	DA (N = 2472)	ICU(N = 539)	Died (N = 2)	Transfer (N = 485)	SA (N = 16)	Total (N = 3514)
Sex male, %	55.3	64.7	100.0	56.9	56.2	57.0
female, %	43.1	34.9	0.0	39.8	43.8	41.3
unknown, %	1.6	0.4	0.0	3.3	0.0	1.6
Age, years med (IQR)	61.0 (51.0–70.0)	63.0 (55.0–70.0)	76.0 (74.5–77.5)	60.0 (53.2–69.0)	66.5 (55.0–75.5)	61.0 (52.0–70.0)
mean (SD)	59.6 (14.2)	61.5 (11.7)	76.0 (3.0)	59.8 (11.9)	64.5 (11.6)	60.0 (13.5)
Ward LOS, days med (IQR)	3.7 (1.9–6.4)	2.3 (1.1–3.9)	7.6 (7.2–7.9)	1.1 (0.8–2.0)	4.7 (1.2–14.5)	2.9 (1.5–5.5)
mean (SD)	5.3 (6.8)	3.4 (4.4)	7.6 (0.7)	1.8 (2.3)	8.6 (8.9)	4.5 (6.2)
RR, breaths/min med (IQR)	18.0 (16.0–22.0)	22.0 (19.0–26.0)	20.0 (20.0–20.0)	20.0 (16.8–24.0)	18.0 (16.0–23.5)	20.0 (16.0–24.0)
mean (SD)	19.3 (5.0)	22.9 (6.0)	20.0 (0.0)	20.8 (5.1)	20.7 (6.2)	20.1 (5.4)
SpO2, % med (IQR)	96.0 (95.0–98.0)	95.0 (94.0–97.0)	95.5 (95.2–95.8)	95.0 (94.0–97.0)	95.0 (93.5–97.0)	96.0 (94.2–97.0)
mean (SD)	96.0 (3.5)	95.1 (4.8)	95.5 (0.5)	95.5 (2.2)	94.5 (3.4)	95.8 (3.6)
SBP, mmHg med (IQR)	125.0 (113.0–137.0)	125.0 (114.0–137.0)	137.5 (136.8–138.2)	123.0 (113.0–133.5)	119.0 (107.2–126.8)	124.0 (113.0–136.0)
mean (SD)	126.4 (18.8)	127.4 (19.3)	137.5 (1.5)	124.1 (16.5)	120.0 (17.3)	126.2 (18.6)
T, °C med (IQR)	37.1 (36.6–37.8)	37.3 (36.7–38.0)	37.0 (37.0–37.1)	37.0 (36.6–37.7)	36.8 (36.2–37.0)	37.1 (36.6–37.8)
mean (SD)	37.2 (0.9)	37.4 (1.0)	37.0 (0.1)	37.2 (0.9)	36.8 (0.7)	37.2 (0.9)
HR, bpm med (IQR)	81.0 (71.0–91.0)	83.0 (73.0–92.0)	91.0 (84.0–98.0)	81.0 (72.0–90.0)	80.0 (73.8–84.8)	81.0 (72.0–91.0)
mean (SD)	82.0 (15.2)	83.5 (15.2)	91.0 (14.0)	81.6 (13.7)	81.2 (13.6)	82.2 (15.0)
O2, yes/no, %	57.4	76.4	0.0	82.5	68.8	63.8
O2, L/min med (IQR)	3.0 (2.0–4.0)	6.0 (3.0–12.0)	-	4.0 (2.0–5.0)	3.0 (2.0–7.5)	3.0 (2.0–5.0)
mean (SD)	3.6 (3.0)	7.5 (5.4)	-	4.4 (3.0)	5.5 (5.1)	4.5 (3.9)
SpO2/O2, 1/(L/min) med (IQR)	32.7 (23.5–48.5)	15.8 (8.1–31.0)	-	24.2 (18.6–47.0)	30.7 (14.1–48.5)	31.7 (18.6–48.0)
mean (SD)	41.4 (26.5)	23.0 (21.3)	-	32.4 (21.1)	35.0 (24.1)	36.4 (25.7)

We simulated the situation in which models would have been developed after the first wave, implemented, and updated each month during the second wave, i.e. the dynamic models. The overall receiver operating characteristic (ROC) curves (and corresponding pAUCs and AUCs) yielded by these models and the existing EWSs are depicted in Fig. 2a. The dynamic RF model outperformed the NEWS and the MEWS in terms of pAUC with respectively ten and fifteen percentage points. We placed landmarks in the ROC curve of the NEWS that correspond with the recommended triggers for an urgent and emergency response.³ An emergency response triggers a critical care outreach team to respond quickly. Vertical differences between the ROC curves represent potential increase in model sensitivity, hence improving the early detection of deteriorating patients. The horizontal differences represent the potential reduction in false alarms and therefore unnecessary urgent and emergency responses. Also in terms of AUC and AUCPR, the dynamic models outperformed the existing EWSs (supplemental table 4). We also simulated the situation in which models would have been implemented after the first wave without any updating, i.e. the static models. These yielded very similar discriminative performance compared to the dynamic models, with a pAUC

of 0.81 [0.79 to 0.83] and 0.80 [0.78 to 0.82], respectively for the RF and LR model. Finally, we validated the model retrospectively, for which the results are summarized in appendix H.

Figure 2b shows the pAUCs yielded by the dynamic models and the existing EWSs in the individual hospitals. The dynamic models outperformed the existing EWSs in most of the hospitals. The static models yielded similar results (supplemental Fig. 6). Hospital-specific results in terms of AUCPR and AUC are depicted in supplemental Fig. 7.

The overall calibration curves for the RF and LR models are shown in Figs. 3a and 3b, respectively, including the corresponding calibration intercepts and slopes. The dynamic models show improved calibration curves compared to the static models. The dynamic RF model yielded a slightly better calibration curve than the dynamic LR model. Hospital-specific calibration curves can be found in supplemental Fig. 9, with calibration intercepts ranging from -0.37 to 0.49 and -0.24 to 0.38 and calibration slopes ranging from 0.55 to 1.11 and 0.69 to 1.42, respectively for the dynamic RF and LR models.

The distribution of SHAP values of the included predictors is shown in Fig. 4. The top five ranked predictors were SpO_2/O_2 , respiratory rate, temperature, ward length-of-stay and O_2 (L/min). The distribution of SHAP values of the LR model shows similar ranking of importance and is shown, together with fitted model parameters, in supplemental Fig. 10.

The LR and RF models fitted with the MEWS predictors outperformed the MEWS, whereas the models fitted with the NEWS predictors yielded similar discriminative performance as the NEWS (appendix I.1). The models fitted with only the top five most important predictors yielded similar performance in discrimination and calibration compared to the models fitted with the originally 18 included predictors (appendix I.2). To further examine why the dynamic models outperform the NEWS, we also fitted an RF and LR model with the NEWS predictors supplemented with the top ranked predictor (SpO_2/O_2). Again, these yielded similar model discrimination and calibration. Models fitted with more predictors (allowing predictors with more missingness) resulted in similar discrimination, but slightly worse calibration. Finally, model discrimination and calibration showed little variation over 50 imputation rounds (appendix I.3).

4. Discussion

We introduced COVID-19-specific early warning models which showed improved performance compared to the existing early warning scores. The dynamic RF model yielded slightly higher NB compared to the dynamic LR model, despite comparable discriminative performance, which is likely due to a slightly better model calibration. The implementation of dynamic model updating showed to be effective to retain good model calibration over time. When implemented, this model could improve the early detection of deteriorating COVID-19 patients and save workload for healthcare workers by reducing the number of false alarms. Importantly, the improved performance relative to the NEWS is explained by the introduction of SpO_2/O_2 as a predictor, which models the need for oxygen quantitatively instead of qualitatively like in the NEWS.

The probability for clinical deterioration within 24 hours predicted by the model enables clinicians to assess deterioration risk in COVID-19 patients and respond appropriately. In contrast to aggregate-weighted EWSs, it cannot be calculated by hand and requires implementation in the EMR.

The choice for clinical outcomes varies significantly in early warning literature.²⁴ The NEWS was originally intended to identify patients at risk of early cardiac arrest, unplanned ICU admission, and death.² As cardiac arrest unavoidably leads to either death or ICU admission, we did not include it as a separate outcome. Baker and colleagues,⁶ who validated NEWS2 longitudinally in COVID-19 patients, included the initiation of different non-invasive respiratory support methods (CPAP, BiPAP and HFNC) as an outcome too. In the Netherlands, these support methods are also offered to COVID-19 patients outside the ICU and therefore, we did not include these as outcomes. Also, Baker and colleagues included the initiation of EoLC as an outcome. In the Netherlands, EoLC may be initiated before a patient actually deteriorates based on, for instance, high age or significant comorbidities. Hence, a patient for whom EoLC is initiated may have a very different clinical presentation compared to a patient with full care just before ICU admission or unexpected death. Therefore, including the initiation of EoLC as an outcome may decrease the model's predictive performance for ICU admission or unexpected death.

Based on this, we excluded patients where EoLC was initiated and those with a 'do not admit to ICU' order. In a recent NEWS validation study,³² Haegdorens and colleagues also excluded patients with a 'do not perform cardiopulmonary resuscitation' order. However, because these patients may still be admitted to the ICU, we did not exclude them.

Different factors could have influenced the generalizability of the model we presented. For instance, improved patient outcomes after the introduction of Dexamethasone after the RECOVERY trial³³ halfway July, 2020, could have caused the models developed before August 2020 to overestimate deterioration when implemented afterwards. Furthermore, given the low prevalence of unexpected death ($< 0.1\%$), the model we presented is predominantly predicting ICU admission, for which there is no universal guideline. On the contrary, whether a patient is admitted to the ICU depends on many factors, such as the clinician who judges the patient or the availability of beds. Therefore, the model is fitted for a context in which a certain policy around ICU admission exists. Hence, if these policies vary among hospitals or over time, the model may fail to generalize. This could be an explanation for the overestimation observed in hospital C, which follows a relatively conservative ICU admission policy due to the absence of medium or high care units. Despite of these factors, our empirical results show that the dynamically updated model generalizes well to future patients and different hospitals.

This study has limitations. Some patients may not have experienced a serious event due to a prompt medical review followed by an intervention. Hence, we may have labeled some samples as negative whereas a patient actually showed signs of deterioration. Furthermore, patients who are transferred to another hospital are (at the moment of transfer) typically less likely to deteriorate. Therefore, informative censoring may have introduced a bias. We treated repeated observations from individual patients as independent samples, potentially underestimating uncertainties around the model performances. We neither validated the NEWS2⁴ nor the alternative version of the MEWS³⁴ that is mostly used in the Netherlands, as we did not have the necessary data. Finally, validation is ideally performed by independent researchers. We have therefore made the model available online and we strongly encourage others to perform further external and temporal validation.

In conclusion, we have shown that a COVID-19-specific early warning model for longitudinal monitoring to identify clinical deterioration shows improved discrimination and net benefit compared to existing EWSs. We advocate further study and development of such patient-group specific EWSs as well as their evaluation in clinical practice.

Declarations

Author's contributions

JS: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Visualization, Writing - original draft, JK: Supervision, Writing review editing, AT: Supervision, Writing - review editing, HE: Supervision, Writing - review editing, JL: Supervision, Writing - review editing, MvG: Writing review editing, SH: Writing - review editing, REM: Writing - review editing, PW: Writing - review editing, RG: Writing - review editing, GW: Data Curation, Writing - review editing, TV: Writing - review editing, JdH: Writing review editing, MdB: Writing - review editing, DG: Writing - review editing, RvdV: Data Curation, Writing - review editing, MS: Writing - review editing, GC: Writing - review editing, CvN: Writing - review editing, MA: Writing - review editing, JvB: Supervision, Writing - review editing, RK: Supervision, Data curation, Writing - review editing, YdR: Supervision, Writing - review editing, MR: Supervision, Writing - review editing.

Funding

Not applicable.

Availability of data and materials

The data that support the findings of this study of each center were available for the authors affiliated to the corresponding center. All data that support the findings of this study were available for the first author in a pseudonymized form. The random forest model fitted on the complete dataset, as well as the code and other dependencies (i.e. for data normalization and imputation) required to perform further independent validation is available online at: https://github.com/jimmsmit/COVID-19_EWS

Ethics approval

The Daily Board of the Medical Ethics Committee Erasmus MC of Rotterdam, The Netherlands, approved this study (reference: MEC-2020-0361), waived the need for patient informed consent and approved of an opt-out procedure for the collection of COVID-19 patient data during the COVID-19 crisis.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Subbe, C. P., Kruger, M., Rutherford, P. & Gemmel, L. Validation of a modified Early Warning Score in medical admissions. *QJM: An International Journal of Medicine* **94**, 521–526 (2001). URL <https://doi.org/10.1093/qjmed/94.521>.
2. Smith, G. B., Prytherch, D. R., Meredith, P., Schmidt, P. E. & Featherstone, P. I. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* **84**, 465–470 (2013).
3. National Early Warning Score (NEWS): Standardising the assessment of acute-illness severity in the NHS. Report of a working party. Tech. Rep., London (2012).
4. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. Updated report of a working party. Tech. Rep., London (2017).
5. Zhang, K. *et al.* The Prognostic Accuracy of National Early Warning Score 2 on Predicting Clinical Deterioration for Patients With COVID-19: A Systematic Review and Meta-Analysis. (2021).
6. Baker, K. F. *et al.* National Early Warning Score 2 (NEWS2) to identify inpatient COVID-19 deterioration: a retrospective analysis. *Clinical medicine (London, England)* **21**, 84–89 (2021).
7. Pimentel, M. A. F. *et al.* Trajectories of vital signs in patients with COVID-19. *Resuscitation* **156**, 99–106 (2020). URL <https://www.sciencedirect.com/science/article/pii/S0300957220304408>.
8. Sze, S. *et al.* Letter to the Editor: Variability but not admission or trends in NEWS2 score predicts clinical outcome in elderly hospitalised patients with COVID-19. (2021).
9. Carr, E. *et al.* Evaluation and Improvement of the National Early Warning Score (NEWS2) for COVID-19: a multi-hospital study Adding. *medRxiv* 04.24.20078006 (2020).
10. Coughlan, C., Rahman, S., Honeyford, K. & Costelloe, C. E. Developing useful early warning and prognostic scores for COVID-19. (2021).
11. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ* **369** (2020). URL <https://www.bmj.com/content/369/bmj.m1328>.
12. Gerry, S. *et al.* Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology. *BMJ* **369** (2020). URL <https://www.bmj.com/content/369/bmj.m1501>.
13. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine* **13**, 1 (2015). URL <https://doi.org/10.1186/s12916-014-0241-z>.
14. Prokop, M. *et al.* CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19—Definition and Evaluation. *Radiology* **296**, E97–E104 (2020). URL <https://doi.org/10.1148/radiol.2020201473>.
15. Fang, A. H. S., Lim, W. T. & Balakrishnan, T. Early warning score validation methodologies and performance metrics: a systematic review. *BMC Medical Informatics and Decision Making* **20**, 111 (2020). URL <https://doi.org/10.1186/s12911-020-01144-8>.
16. Romanelli, D. & Farrell, M. W. AVPU Score. (2021).
17. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *Journal of machine learning research* **12**, 2825–2830 (2011).
18. Niculescu-Mizil, A. & Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, 625–632 (Association for Computing Machinery, New York, NY, USA, 2005). URL <https://doi.org/10.1145/1102351.1102430>.
19. Ancker, J. S. *et al.* Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC Medical Informatics and Decision Making* **17**, 36 (2017). URL <https://doi.org/10.1186/s12911-017-0430-8>.

20. Haegdorens, F., Monsieurs, K.G., DeMeester, K.&VanBogaert, P. The optimal threshold for prompt clinical review: An external validation study of the national early warning score. *Journal of clinical nursing* **29**, 4594–4603 (2020).
21. Smith, G. B. *et al.* A Comparison of the Ability of the Physiologic Components of Medical Emergency Team Criteria and the U.K. National Early Warning Score to Discriminate Patients at Risk of a Range of Adverse Clinical Outcomes. *Critical care medicine* **44**, 2171–2181 (2016).
22. Royal College of Physicians of London. *National Early Warning Score (NEWS) : standardising the assessment of acute-illness severity in the NHS* (2012).
23. McClish, D. K. Analyzing a portion of the ROC curve. *Medical decision making : an international journal of the Society for Medical Decision Making* **9**, 190– 195 (1989).
24. Fu, L.-H. *et al.* Development and validation of early warning score system: A systematic literature review. *Journal of biomedical informatics* **105**, 103410 (2020).
25. Boyd, K., Eng, K. H. & Page, C. D. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In Blockeel, H., Kersting, K., Nijssen, S. & Železný, F. (eds.) *Machine Learning and Knowledge Discovery in Databases*, 451–466 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013).
26. Qin, G. & Hotilovac, L. Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research* **17**, 207–221 (2008).
27. Vickers, A. J., van Calster, B. & Steyerberg, E. W. A simple, step-by-step guide to interpreting decision curve analysis. *Diagnostic and Prognostic Research* **3**, 18 (2019). URL <https://doi.org/10.1186/s41512-019-0064-7>.
28. Van Calster, B. *et al.* A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology* **74**, 167–176 (2016).
29. COX, D. R. Two further applications of a model for binary regression. *Biometrika* **45**, 562–565 (1958). URL <https://doi.org/10.1093/biomet/45.3-4.562>.
30. Austin, P. C. & Steyerberg, E. W. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. *Statistics in medicine* **33**, 517–535 (2014).
31. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017). URL <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>.
32. Haegdorens, F. *et al.* The introduction of a rapid response system in acute hospitals: A pragmatic stepped wedge cluster randomised controlled trial. *Resuscitation* **129**, 127–134 (2018). URL <https://www.sciencedirect.com/science/article/pii/S0300957218301795>.
33. Dexamethasone in Hospitalized Patients with Covid-19. *New England Journal of Medicine* **384**, 693–704 (2021). URL <https://doi.org/10.1056/NEJMoA2021436>.
34. Ludikhuizen, J., Smorenburg, S. M., de Rooij, S. E. & de Jonge, E. Identification of deteriorating patients on general wards; measurement of vital parameters and potential effectiveness of the Modified Early Warning Score. *Journal of critical care* **27**, 424.e7–13 (2012).

Figures

Figure 1

(a) Schematic representation of the dynamic model updating procedure. For example, to predict deterioration for patients admitted to hospital A in October 2020, the model is fitted using patient data collected up to that date in the remaining hospitals, and a calibrator is fitted using patient data collected up to that date in hospital A. These two combined result in calibrated predictions. This process is repeated each month, for each hospital, from August 2020 until May 2021.

(b) Flowchart of patient inclusion. ICU=Intensive care unit, ED=Emergency department, EoLC=end-of-life care, LOS=Length-of-stay.

Study design.

Figure 2

(a) Overall ROC curves for the RF and LR models and the NEWS. We placed two landmarks for a NEWS score of 5 and 7, i.e. the recommended trigger thresholds for an urgent and emergency response. We calculated both the pAUC between a false positive rate of 0 and 0.33 (grey area) and the complete AUC. Shaded areas around each point in the ROC curves represent the 95% bootstrap percentile CIs²⁶ (with 1000 bootstrap replications stratified for positive and negative samples).

(b) Hospital-specific pAUCs. The error-bars represent the 95% bootstrap percentile CIs²⁶ (with 1000 bootstrap replications stratified for positive and negative samples). P-values, calculated as described in appendix F.4, are shown for the difference in pAUC between the RF models and NEWS (upper bar), between the RF and LR models (middle bar) and between the LR models and NEWS (lower bar).

(c) Overall decision curve analysis results. The standardized net benefit is plotted over a range of clinically relevant probability thresholds with corresponding odds. The 'Intervention for all' line indicates the NB if a (urgent or emergency) response would always be triggered.

Model discrimination and decision curve analysis.

Figure 3

(a) Overall model calibration of the static and dynamic RF models.

(b) Overall model calibration of the static and dynamic LR models.

Overall model calibration of the static and dynamic RF models (a) and LR models (b). (i) Smoothed flexible calibration curves. (ii) Zoom-in of the calibration curve in the 0-0.2 probability range (grey area). Shaded areas around the curves represent the 95% CIs. (iii) Histogram of the predictions(log scale). Shaded areas around each point in the calibration curves (before smoothing) represent the 95% bootstrap percentile CIs²⁶ (with 1000 bootstrap replications stratified for positive and negative samples).

Figure 4

Distribution of SHapley Additive explanations (SHAP) values of the included predictors (based on mean SHAP magnitude) for the random forest model. For each predictor, each dot represents the impact of that predictor for a single prediction. The colors of the dots correspond with the value for the specific predictor. Thus, pink dots with positive SHAP values indicate that high values of the predictor are associated with a high risk of clinical deterioration. Conversely, blue dots with positive SHAP values indicate that low values of the predictor are associated with a high risk of clinical deterioration.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarymaterial.pdf](#)