A MACHINE LEARNING MODEL FOR PREDICTING DETERIORATION OF COVID-19 INPATIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The COVID-19 pandemic has been spreading worldwide since December 2019, presenting an urgent threat to global health. Due to the limited understanding of disease progression and of the risk factors for the disease, it is a clinical challenge to predict which hospitalized patients will deteriorate. Moreover, several studies suggested that taking early measures for treating patients at risk of deterioration could prevent or lessen condition worsening and the need for mechanical ventilation. We developed a predictive model for early identification of patients at risk for clinical deterioration by analyzing electronic health records of COVID-19 inpatients at the Sheba and Tel-Aviv Medical Centers in Israel. Our model employs machine learning methods and uses routine clinical features. Deterioration was defined as a high NEWS2 score adjusted to COVID-19. In prediction of deterioration within the next 30 hours, the model achieved an area under the ROC curve of 0.84 and area under the precision-recall curve of 0.74. It achieved values of 0.76 and 0.7 respectively in external validation on data from a different hospital.

1 INTRODUCTION

The coronavirus disease 2019 (COVID-19) emerged in China in December 2019, and since then has spread rapidly around the world. In March 2020, the World Health Organization declared the COVID-19 outbreak as a global pandemic (Cucinotta & Vanelli, 2020). As of February 2021, worldwide cases exceeded 111 million and nearly 2.5 million died (Dong et al., 2020). The extent of the disease varies from asymptomatic to severe, characterized by respiratory and/or multi-organ failure and death (Lapostolle et al., 2020; Huang et al., 2020). Healthcare systems worldwide have faced an overwhelming burden of patients with COVID-19. At the same time, there is limited understanding of disease progression, risk factors for deterioration, and the long-term outcomes for those who deteriorate. Moreover, early treatments such as antiviral medications may prevent clinical deterioration in COVID-19 patients (Mathies et al., 2020). Therefore, early warning tools for COVID-19 deterioration are required. Tools that predict deterioration risk in individuals can also improve resource utilization in the clinical facility, by aggregating risk scores of patients for anticipating expected changes in patient load (Bravata et al., 2021).

Prognostic scores for clinical deterioration of patients are widely used in medicine, particularly in critical care. The National Early Warning Score 2 (NEWS2), the quick Sequential Organ Function Assessment (qSOFA), and CURB-65 (RCP, 2017; Asai et al., 2019; Chalmers et al., 2011) are commonly used clinical risk scores, used for early recognition of patients with severe infection. The NEWS2 score incorporates pulse rate, respiratory rate, blood pressure, temperature, oxygen saturation, supplemental oxygen, and level of consciousness or new confusion. Liao et al. (2020) suggested an early warning score for COVID-19 patients termed "modified-NEWS2" (mNEWS2). It adds to the NEWS2 formula the factor age$\geq$65 years, reflecting the observation that increased age is associated with elevated risk for severe illness (Supplementary Table 1).

We developed a machine learning model for early prediction of deterioration of COVID-19 inpatients, presented as mNEWS2 score$\geq$7. The model was developed by analyzing longitudinal electronic health records (EHRs) of COVID-19 inpatients in Sheba Medical Center (Sheba), the largest hospital in Israel. To validate the model's generalizability, we applied it on EHRs of inpatients diagnosed with COVID-19 from the second largest hospital in Israel, the Tel-Aviv Sourasky Medical Center (TASMC).

2 METHODS

Cohort Description. We conducted a retrospective cohort study comprising two datasets. The *development dataset* consisted of all patients admitted to Sheba between March and December 2020 that tested positive for SARS-CoV-2. The *validation dataset* consisted of all patients admitted to TASMC between March and September 2020 who tested positive for SARS-CoV-2. The data used was extracted from longitudinal EHRs and included both time-independent (static) and temporal (dynamic) features, such as demographics, background disease, vital signs and lab measurements (Supplementary Table 2). The temporal data was discretized to hourly intervals and multiple values of a test measured within the same hour were aggregated by mean. We use the term *observation* for the vector of test values on a particular hour. An observation was formed if at least one measurement was recorded in that hour.

After applying the inclusion and exclusion criteria (detailed in the Supplemental Methods A.1), the development set contained 25,105 hourly observations derived from 662 patients; the validation set had 7,737 observations derived from 417 patients. The characteristics of the two datasets are described in Supplementary Table 2.

Observations with a high mNEWS2 score ($\geq$7) recorded in the preceding 7-30 hours were called positive, and the rest were called negative. Observations where no mNEWS2 score is available in the next 30 hours were excluded. Higher mNEWS2 scores were associated with higher mortality and ICU admissions rates in the development dataset (Supplementary Figure 1).

Feature Engineering. We created summary statistics over time windows of varying sizes to capture the temporal behavior of the data. The summary statistics were generated for 21 dynamic features that were reported as risk factors for severe COVID-19 (Gong et al., 2020; Haimovich et al., 2020; Guo et al., 2020; Ji et al., 2020; Liu et al., 2020) (Supplementary Table 3). We defined two time windows covering the last 24 and 72 hours. For each time window, the summary statistics extracted were the mean, difference between the current value and the mean, standard deviation, minimum and maximum values. In addition, we extracted the same summary statistics based on the entire hospitalization so far, with the addition of the linear regression slope (the regression coefficient). To capture recent data patterns, the difference and trend estimate from the last observed value ($(v_2 - v_1)$ and $\frac{(v_2-v_1)}{(t_2-t_1)}$ for values v_1,v_2 recorded in times t_1,t_2 respectively) were generated as well. In addition, to capture the interaction between pairs of variables, we generated features for the ratios of each pair of variables in the risk factors subset (for example, neutrophils to lymphocytes ratio). We also examined engineered features that capture the measurement frequency and anomalous data points, but none of these were used for the final model (see A.1).

Data Imputation. Missing values were observed mainly in lab tests and vital signs. We used linear interpolation for imputing missing data. The remaining missing data were imputed using the *Mice* (Multivariate Imputation by Chained Equation) approach, which models each feature with missing values as a function of other features (Buuren & Groothuis-Oudshoorn, 2010).

Model Development and Feature Selection. We performed a binary classification task for every hourly observation to predict deterioration in the next 7-30 hours. Deterioration was defined as mNEWS2$\geq$7. As deterioration can usually be predicted by a clinician several hours in advance, based on signs and symptoms, observations from the six hours prior to the deterioration event were excluded (Supplementary Figure 2). Once deterioration has occurred, no predictions were made in the next 8 hours, and observations during that period were excluded.

We evaluated ten supervised machine learning models for this prediction task: linear regression (Tibshirani, 1996; Hoerl & Kennard, 1970), logistic regression, naïve Bayes, support vector machine (SVM) (Cortes & Vapnik, 1995), random forest (Breiman, 2001) and several algorithms for gradient boosting of decision trees, including XGBoost (Chen & Guestrin, 2016) and CatBoost (Dorogush et al., 2018). Data standardization was performed prior to model training when needed. We used the computed feature importance scores, calculated by XGBoost, to select the top 100 features for models training (detailed in A.1, Supplementary Table 4).

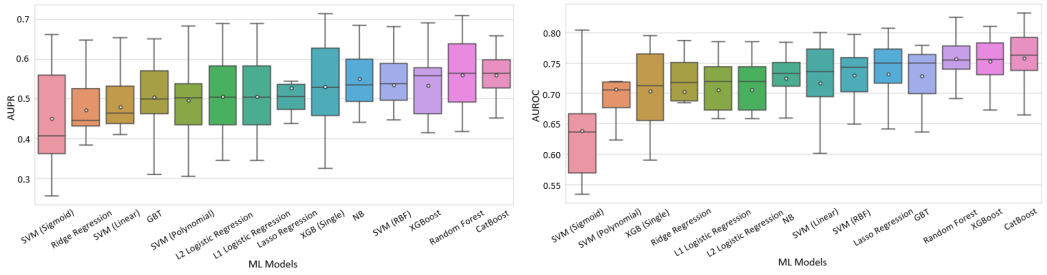
Evaluation Approach. We partitioned the development dataset into 80% *training* and 20% *testing subsets*. To avoid bias resulting from changes in clinical practice over time, the partition was done randomly across the hospitalization dates. To estimate the robustness of the models on different

patients and time periods, we used 10-fold cross-validation over the training set, and measured model performance using the area under the receiver-operator characteristics curve (AUROC) and the area under the precision-recall curve (AUPR). The testing set was used to evaluate the final model performance within the same cohort. Finally, we used the validation dataset (TASMC) for external evaluation.

3 RESULTS

COVID-19 DETERIORATION MODEL. Our models generate a prediction score for the risk of deterioration in each hour that contains a new observation. The models were trained on the training set within the development dataset, consisting of 20,029 hourly observations (e.g. when a new record is available) derived from 530 patients, of which 6,349 (~31%) were labeled positive (mNEWS2≥7 in the next 30 hours). **Figure 1** summarizes the performance of 14 classifiers that were tested in cross-validation on the training set. All predictions refer to events at least seven hours in advance. Classifiers based on an ensemble of decision trees (CatBoost, XGBoost, Random Forest) performed best overall. We chose CatBoost as our final prediction model. Its results on the development testing set are shown in **Figure 2** (a-b). It had good discrimination and achieved AUROC of 0.84 and AUPR of 0.74. To estimate the robustness of the model, we performed a bootstrap procedure with 100 iterations, where, in each iteration, 50% of the testing set was randomly sampled with replacement. The mean and standard deviation of the AUROC and the AUPR over these experiments achieved comparable results to those of the total testing set (**Figure 2a-b**).

To assess the contribution of each feature to the final model prediction, we used SHAP (SHapley Additive exPlanations) values (Lundberg & Lee, 2017). The absolute value indicates the extent of the contribution of the feature, while its sign indicates whether the contribution is positive or negative. The top 20 important features of the final model are summarized in **Figure 2c**, presenting age, arterial oxygen saturation, maximal LDH value and the standard deviation of body temperature as important features for predicting deterioration.



(a) **(b)**

Figure 1: Performance of 14 machine learning models that predict mNEWS2≥7, using 10-fold cross-validation over the training set. (a) AUPR; (b) AUROC. Horizonal line: median. White circle: mean. The models are sorted by the median AUC.

EXTERNAL VALIDATION. The dataset from TASMC was used for external validation of the final model. The results (Supplementary Figure 3) show good performance with AUC 0.76 and AUPR 0.7, albeit less than in the development dataset. A certain reduction in performance is to be expected when validating a predictor on an independent data source. The slight decrease in performance here can be explained, in part, by a lower temporal resolution of the TASMC dataset, as well as by a higher rate of missing values.

4 DISCUSSION

We utilized machine learning models for predicting deterioration event in the next 7-30 hours based on EHR data of adult COVID-19 inpatients. Deterioration was defined as a high COVID-19 early warning score (mNEWS2≥7). On held-out data, the model achieved AUROC of 0.84 and AUPR of 0.74. The model was tested on an independent patient cohort from a different hospital and demon-

strated comparable performance, with only a modest decrease. Using our predictor, we could anticipate deterioration of patients 7-30 hours in advance. Such early warning can enable timely intervention, which was shown to be beneficial (Mathies et al., 2020). Given that the mNEWS2 score is broadly adopted as a yardstick of COVID-19 inpatient status in medical centers around the world, we believe that demonstrating early prediction of high scores could provide valuable insights to physicians and focus their attention towards particular patients that are predicted to be at high risk to deteriorate in the near future.

Previous studies have assessed the utility of machine learning for predicting deterioration in COVID-19 patients (Assaf et al., 2020; Gao et al., 2020). The novelty of our methodology lies in the fact that our model generates repeatedly updated predictions for each patient during the hospitalization, using both baseline and longitudinal data. This enables the identification of patients at risk throughout the hospitalization, while accounting for the temporal dynamics of the disease, allowing adjusted patient therapy and management. In addition, our work emphasizes the importance of including summary statistics of medical and inflammatory markers, such as the standard deviation of body temperature, for predicting the risk of COVID-19 deterioration. Our model can be readily applied with other criteria for deterioration, e.g., mechanical ventilation or other mNEWS2 cutoffs.

Our study has several limitations. First, it is retrospective, and model development was done based on data from a single center, which may decrease its generalizability to external cohorts, especially considering the high variability of COVID-19 outcomes. Second, the mNEWS2 scores present a noisy signal, with frequent changes during the hospitalization. This impairs the score's ability to be used as a robust predictor, compared to other approaches for predicting deterioration, such as initiation of mechanical ventilation or death (Gao et al., 2020; Douville et al., 2020). Third, a deteriorating patient will tend to have more frequent mNEWS2 measurements. This may bias our model and impair its adaptability to a general population of patients (see A.2 regarding this bias).

In conclusion, machine learning-based prognostic tools have great potential for care decisions and resource utilization in hospital wards. In spite of the fact that COVID-19 is a novel disease with high complexity, our model provides useful predictions for risk of deterioration, with good discrimination. Early detection and treatment of COVID-19 patients at high risk of deterioration may lead to improved treatment and reduction in mortality. Further validation of this vision is needed.

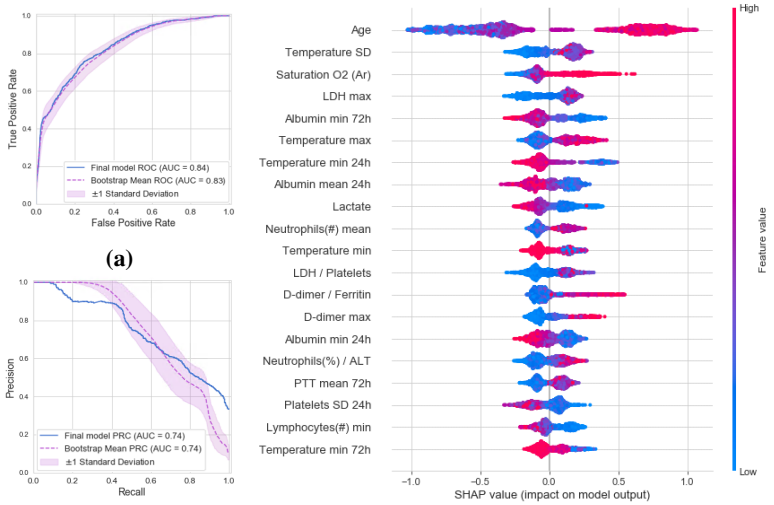


Figure 2: (a-b) Performance of the final model on the testing set. Solid curves were computed from the total set. Dashed curves were computed with a bootstrap procedure. Top: AUROC. Bottom: AUPR. (c) 20 features with highest mean absolute SHAP values. Features (rows) are ordered in decreasing overall importance. The plot for each feature shows the SHAP value for each observation on the x-axis, with color representing the value of the feature from low (blue) to high (red). SD: standard deviation; /: ratio between two features. 24h,72h: time windows which the statistic was computed (if not mentioned, the statistics is calculated on the entire hospitalization period so far).

REFERENCES

Nobuhiro Asai, Hiroki Watanabe, Arufumi Shiota, Hideo Kato, Daisuke Sakanashi, Mao Hagihara, Yusuke Koizumi, Yuka Yamagishi, Hiroyuki Suematsu, and Hiroshige Mikamo. Efficacy and accuracy of qSOFA and SOFA scores as prognostic tools for community-acquired and healthcare-associated pneumonia. *International Journal of Infectious Diseases*, 84:89–96, 2019.

Dan Assaf, Ya'ara Gutman, Yair Neuman, Gad Segal, Sharon Amit, Shiraz Gefen-Halevi, Noya Shilo, Avi Epstein, Ronit Mor-Cohen, Asaf Biber, et al. Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Internal and emergency medicine*, 15(8): 1435–1443, 2020.

Dawn M Bravata, Anthony J Perkins, Laura J Myers, Greg Arling, Ying Zhang, Alan J Zillich, Lindsey Reese, Andrew Dysangco, Rajiv Agarwal, Jennifer Myers, et al. Association of intensive care unit patient load and demand with mortality rates in us department of veterans affairs hospitals during the COVID-19 pandemic. *JAMA network open*, 4(1):e2034266–e2034266, 2021.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

S van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, pp. 1–68, 2010.

James D Chalmers, Pallavi Mandal, Aran Singanayagam, Ahsan R Akram, Gourab Choudhury, Philip M Short, and Adam T Hill. Severity assessment tools to guide ICU admission in community-acquired pneumonia: systematic review and meta-analysis. *Intensive care medicine*, 37(9):1409–1420, 2011.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Domenico Cucinotta and Maurizio Vanelli. Who declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, 91(1):157, 2020.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

Nicholas J Douville, Christopher B Douville, Graciela Mentz, Michael R Mathis, Carlo Pancaro, Kevin K Tremper, and Milo Engoren. Clinically applicable approach for predicting mechanical ventilation in patients with COVID-19. *British Journal of Anaesthesia*, pp. 578–589, 2020.

Yue Gao, Guang-Yao Cai, Wei Fang, Hua-Yi Li, Si-Yuan Wang, Lingxi Chen, Yang Yu, Dan Liu, Sen Xu, Peng-Fei Cui, et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature communications*, 11(1):1–10, 2020.

Jiao Gong, Jingyi Ou, Xueping Qiu, Yusheng Jie, Yaqiong Chen, Lianxiong Yuan, Jing Cao, Mingkai Tan, Wenxiong Xu, Fang Zheng, et al. A tool for early prediction of severe coronavirus disease 2019 (COVID-19): a multicenter study using the risk nomogram in Wuhan and Guangdong, China. *Clinical infectious diseases*, 71(15):833–840, 2020.

Yabing Guo, Yingxia Liu, Jiatao Lu, Rong Fan, Fuchun Zhang, Xueru Yin, Zhihong Liu, Qinglang Zeng, Jing Yuan, Shufang Hu, et al. Development and validation of an early warning score (EWAS) for predicting clinical deterioration in patients with coronavirus disease 2019. *medRxiv*, 2020.

Adrian Haimovich, Neal G Ravindra, Stoytcho Stoytchev, H Patrick Young, Francis P Wilson, David van Dijk, Wade L Schulz, and Richard Andrew Taylor. Development and validation of the COVID-19 severity index (csi): a prognostic tool for early respiratory decompensation. *medRxiv*, 2020.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, Jianping Zhao, Yi Hu, Li Zhang, Guohui Fan, Jiuyang Xu, Xiaoying Gu, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet*, 395(10223):497–506, 2020.

Dong Ji, Dawei Zhang, Jing Xu, Zhu Chen, Tieniu Yang, Peng Zhao, Guofeng Chen, Gregory Cheng, Yudong Wang, Jingfeng Bi, et al. Prediction for progression risk in patients with COVID-19 pneumonia: the call score. *Clinical Infectious Diseases*, 71(6):1393–1399, 2020.

Frédéric Lapostolle, Elodie Schneider, Isabelle Vianu, Guillaume Dollet, Bastien Roche, Julia Berdah, Julie Michel, Laurent Goix, Erick Chanzy, Tomislav Petrovic, et al. Clinical features of 1487 COVID-19 patients with outpatient management in the greater Paris: the COVID-call study. *Internal and emergency medicine*, 15:813–817, 2020.

Xuelian Liao, Bo Wang, and Yan Kang. Novel coronavirus infection during the 2019–2020 epidemic: preparing intensive care units—the experience in Sichuan province, China. *Intensive care medicine*, 46(2):357–360, 2020.

Xiaohui Liu, Si Shi, Jinling Xiao, Hongwei Wang, Liyan Chen, Jianing Li, and Kaiyu Han. Prediction of the severity of corona virus disease 2019 and its adverse clinical outcomes. *Japanese Journal of Infectious Diseases*, pp. JJID–2020, 2020.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Daniel Mathies, Dominic Rauschning, Ulrike Wagner, Frank Mueller, Maja Maibaum, Christin Binnemann, Stephan Waldeck, Katrin Thinnes, Michael Braun, Willi Schmidbauer, et al. A case of SARS-CoV-2 pneumonia with successful antiviral therapy in a 77-year-old man with a heart transplant. *American Journal of Transplantation*, 20(7):1925–1929, 2020.

RCP. National early warning score (NEWS) 2 — RCP London. `https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2`, 2017.

Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A APPENDIX

A.1 SUPPLEMENTAL METHODS

Cohort Description. The study was reviewed and approved by the Sheba Institutional Review Board (number 20-7064) and by the TASMC Institutional Review Board (number 0491-17). The static features were age, sex, weight, BMI and background diseases. The background diseases included hypertension, diabetes, cardiovascular diseases, chronic obstructive pulmonary disease (COPD), chronic kidney disease (CKD), cancer, hepatitis B and human immunodeficienc virus (HIV). The dynamic features include measurement of vital signs (including oxygen saturation), complete blood count (CBC), basic metabolic panel (BMP), blood gases, coagulation panel and lipids panel, including kidney and liver function tests, and inflammatory markers (Supplementary Table 2). Features with more than 40% missing values or with zero variance were excluded. The temporal data was discretized to hourly intervals and patients' values measured within the same hour were aggregated by mean. Each observation contained all of the aggregated features of the patient for that hour.

Inclusion criteria: Adult patients (age$\geq$18) with at least one mNEWS2 score.

Exclusion criteria: Patients who had mNEWS2 score$\geq$7 in the first 12 hours after admission. Patients with no test results for BMP, CBC and coagulation. Patients' observations with $\geq$60% of the feature values missing. Patients' observations from the six hours period prior to the deterioration event. Patients' observations from the eight hours after the deterioration event.

Outlier Removal. To remove grossly incorrect measurements due to manual typos or technical issues, we manually defined with clinicians a range of possible values (including pathological values) per each feature (Supplementary Table 3), and removed values outside this range. In total, 43,507 values were excluded.

Feature Engineering. As imputation masks the information about the measurement frequency, we added features that capture the time since the last non-imputed measurement. While these features indeed improved our performance, the intensity of monitoring of a patient may reflect her medical condition (a deteriorating patient will tend to have more frequent measurements). As we aimed to predict deterioration when is not yet anticipated, we chose not to include these features in the developed model, as they can create bias due to measurement intensity.

Since patients are hospitalized in different stages of their disease progression, we added to the model unsupervised features that aimed to define how much an observation is irregular compared to the other observations. To capture anomalous data points, we examined the impact of adding anomaly scores as features. We applied three anomaly detection approaches, One-Class SVM, Isolation Forest, and local outlier factor (LOF) to each hourly observation. Eventually, none of the anomaly features was included in the final model after the feature selection.

Feature Selection. To handle the high dimensionality of our data after the feature engineering process, we examined two strategies or feature selection. Two strategies for feature selection were examined. The first selected the features with the highest correlation with the target. The second used feature importance as calculated by XGBoost. Specifically, we trained XGBoost on the training dataset and used the computed feature importance scores to select the top 100 features (Supplementary Table 4).

A.2 POTENTIAL MEASUREMENT INTENSITY BIAS

A potential concern is that deteriorating patient will tend to have more frequent mNEWS2 measurements. This may bias our model and impair its adaptability to a general population of patients. To mitigate the bias due to measurement intensity, we chose not to include features that capture measurement frequency, although including them can improve performance. In addition, the training data had a majority of negative observations (~69%), showing that mild and modest conditions are well represented in the data. Furthermore, by summarizing measurements per hour we mask the measurement intensity within the same hour. Future work could examine time discretization over longer time windows and utilization of balancing techniques.

A.3 SUPPLEMENTAL TABLES AND FIGURES

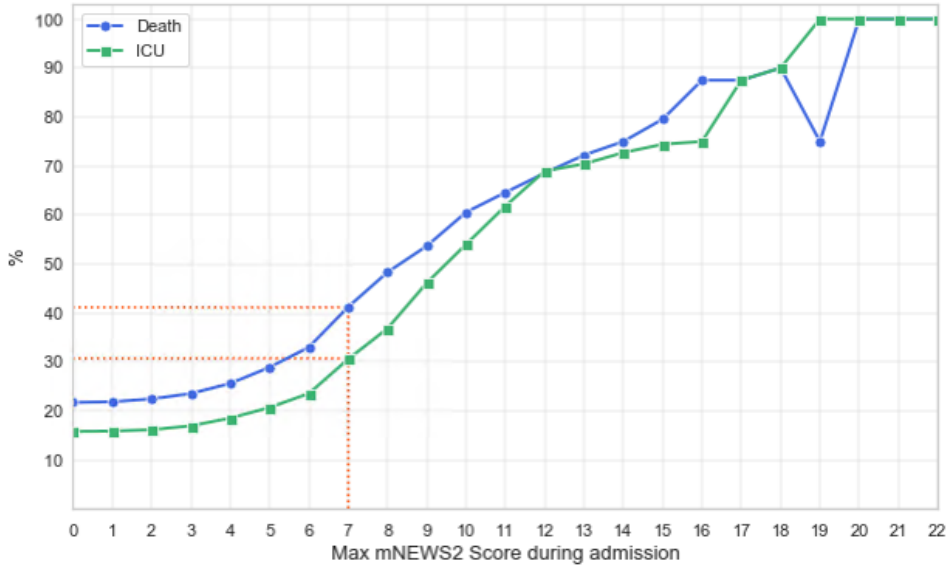


Figure 1: Death and ICU admission rates as a function of the maximal mNEWS2 score during hospitalization in the development dataset.

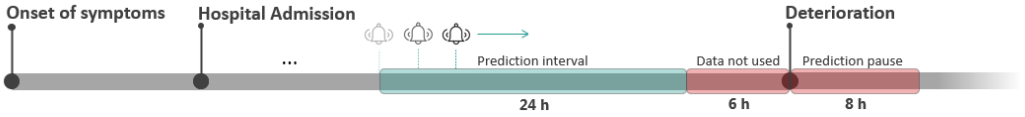
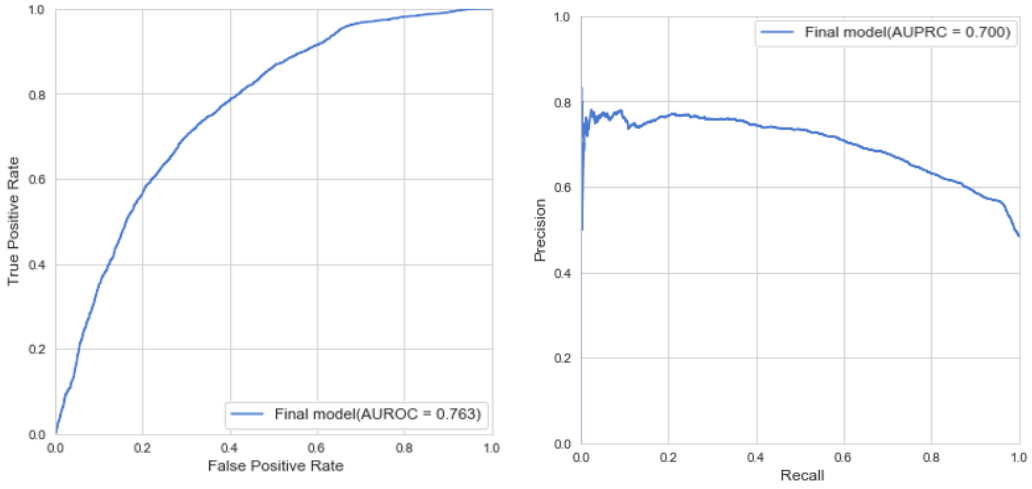


Figure 2: Patient timeline from symptoms to deterioration. Red areas represent blocked prediction periods.



(a) **(b)**

Figure 3: External validation of the final model on the TASMC data. (a) AUROC. (b) AUPRC.

Table 1: The mNEWS2 score. Scores are computed by summing the points for each category. This is an adapted version of the NEWS2 score, with the addition of 3 points for patients with age≥65 (Liao et al., 2020). Scores are computed by summing the points for each category.

Points	**3**	**2**	**1**	**0**	**1**	**2**	**3**
Age				<65			≥65
Respiratory Rate	≤8		9-11	12-20		21-24	≥25
Oxygen Saturation	≤91	92-93	94-95	≥96			
Supplemental Oxygen		Yes		No			
Systolic BP	≤90	91-100	101-110	111-219			≥220
Heart Rate	≤40		41-50	51-90	91-110	111-130	≥131
Consciousness				Alert			Drowsiness Letargy Coma Confusion
Temperature	≤35.0		35.1-36.0	36.1-38.0	38.1-39.0	≥39.1	

Score	**Risk Grading**
0	-
1-4	Low
5-6 or 3 in one parameter	Medium
≥ 7	High

Table 2: Table of Characteristics. Population characteristics for the two datasets used to develop and validate the model. Characteristics of both static features and first measurements of dynamic features are presented. P-values were calculated using Fisher's exact test and T-test for categorical and numerical values, repectively, and Bonferroni corrected for multiple comparisons. 'AR' and 'V' refer to arterial and venous blood, respectively

Variable	Development (Sheba Hospital)		Validation (TASMC)		P-Value
	N (%)	Mean ± SD	N (%)	Mean ± SD	
Overall	662		417		
Age	662 (100.0%)	65.91 ± 16.86	417 (100.0%)	67.16 ± 18.01	1
Gender					1
Male	391 (59.06%)		241 (57.79%)		
Female	271 (40.94%)		176 (42.21%)		
BMI	497 (75.08%)	28.21 ± 7.51	234 (56.12%)	27.37 ± 5.88	1
Weight	507 (76.59%)	81.11 ± 19.76	245 (58.75%)	77.25 ± 18.48	0.5385
Hypertension					<0.0001
Yes	287 (43.35%)		60 (14.39%)		
No	375 (59.65%)		357 (85.61%)		
Diabetes					0.0041
Yes	195 (29.46%)		83 (19.9%)		
No	467 (70.54%)		334 (80.1%)		
Cancer					0.0007
Yes	91 (13.75%)		26 (6.24%)		
No	571 (86.25%)		391 (93.76%)		
Hepatitis B					1
Yes	5 (0.76%)		3 (0.72%)		
No	657 (99.24%)		414 (99.28%)		
CKD					1
Yes	43 (6.5%)		27 (6.47%)		
No	619 (93.5%)		390 (93.53%)		
HIV					1
Yes	1 (0.15%)		1 (0.24%)		
No	661 (99.85%)		416 (99.76%)		
CVD					0.0042
Yes	149 (22.51%)		58 (13.91%)		
No	513 (77.49%)		359 (86.09%)		
COPD					1
Yes	26 (3.93%)		22 (5.28%)		
No	636 (96.07%)		395 (94.72%)		
HBA1C (#)	108 (16.31%)	45.34 ± 15.92	39 (9.35%)	54.11 ± 20.1	0.3497
HBA1C (%)	108 (16.31%)	6.3 ± 1.46	39 (9.35%)	7.1 ± 1.84	0.3629
Albumin	658 (99.4%)	35.81 ± 5.44	377 (90.41%)	38.35 ± 4.85	<0.0001
ALT	660 (99.7%)	33.02 ± 29.06	412 (98.8%)	37.47 ± 72.29	1
AST	661 (99.85%)	46.63 ± 43.84	378 (90.65%)	43.44 ± 35.46	1
BUN	662 (100.0%)	23.42 ± 16.66	417 (100.0%)	22.8 ± 19.15	1
Calcium	662 (100.0%)	8.82 ± 0.64	379 (90.89%)	8.67 ± 0.59	0.0098
CPK	644 (97.28%)	219.57 ± 595.46	380 (91.13%)	232.15 ± 451.02	1
Creatinine	662 (100.0%)	1.18 ± 0.98	417 (100.0%)	1.2 ± 1.36	1
Direct bilirubin	359 (54.23%)	0.28 ± 0.54	385 (92.33%)	0.26 ± 0.34	1
D-dimer	627 (94.71%)	3.22 ± 4.81	371 (88.97%)	2.19 ± 3.88	0.0244
Ferritin	471 (71.15%)	641.34 ± 1094.33	348 (83.45%)	799.82 ± 1280.3	1
Fibrinogen	577 (87.16%)	500.56 ± 174.18	320 (76.74%)	524.01 ± 145.45	1
Glucose	662 (100.0%)	135.34 ± 59.23	417 (100.0%)	128.16 ± 66.65	1
HCO3	597 (90.18%)	24.62 ± 3.78	341 (81.77%)	24.35 ± 4.33	1
HGB	662 (100.0%)	12.67 ± 2.14	417 (100.0%)	13.2 ± 1.94	0.0022
INR	656 (99.09%)	1.13 ± 0.21	416 (99.76%)	1.08 ± 0.21	0.0078
Lactate	600 (90.63%)	2.07 ± 1.07	178 (42.69%)	2.07 ± 1.33	1
LDH	656 (99.55%)	353.77 ± 215.99	367 (88.01%)	565.31 ± 269.78	<0.0001
Lymphocytes (#)	660 (99.7%)	1.18 ± 0.83	417 (100.0%)	1.18 ± 0.95	1
Lymphocytes (%)	662 (100.0%)	18.3 ± 11.77	417 (100.0%)	17.6 ± 12.02	1
Neutrophils (#)	662 (100.0%)	5.59 ± 3.86	417 (100.0%)	6.09 ± 4.27	1
Neutrophils (%)	662 (100.0%)	71.7 ± 14.19	417 (100.0%)	73.12 ± 13.49	1
NRBC	98 (14.8%)	1.24 ± 0.97	416 (99.76%)	0.2 ± 0.38	<0.0001
Osmolality (urine)	25 (3.78%)	368.0 ± 158.68	47 (11.27%)	451.17 ± 186.09	1

PO2 (AR)	56 (8.46%)	74.69 ± 27.7	341 (81.77%)	40.04 ± 36.73	<0.0001
PO2 (V)	596 (90.03%)	36.44 ± 27.54	81 (19.42%)	34.89 ± 35.42	1
PCO2 (AR)	56 (8.46%)	49.83 ± 11.69	342 (82.01%)	42.15 ± 9.08	<0.0001
PCO2 (V)	598 (90.33%)	43.12 ± 8.95	76 (18.23%)	42.67 ± 11.0	1
PH	541 (81.72%)	7.37 ± 0.08	341 (81.77%)	7.38 ± 0.07	1
Platelet	662 (100.0%)	208.46 ± 99.93	417 (100.0%)	202.05 ± 82.31	1
Potassium	661 (99.85%)	4.16 ± 0.61	417 (100.0%)	4.04 ± 0.59	0.0758
PTT	649 (98.04%)	30.88 ± 9.79	416 (99.76%)	32.32 ± 6.42	0.4093
RBC	662 (100.0%)	4.48 ± 0.75	417 (100.0%)	4.48 ± 0.69	1
RDW	662 (100.0%)	14.83 ± 2.23	417 (100.0%)	14.43 ± 1.68	0.0874
Sodium	662 (100.0%)	136.05 ± 5.69	417 (100.0%)	136.0 ± 5.74	1
Saturaion O2 (AR)	592 (89.43%)	57.34 ± 24.21	247 (59.23%)	64.32 ± 26.61	0.0119
Total bilirubin	661 (99.85%)	0.67 ± 0.68	412 (98.8%)	0.6 ± 0.48	1
Triglycerides	367 (55.44%)	162.83 ± 118.33	309 (74.1%)	135.95 ± 67.14	0.0217
Troponin	575 (86.86%)	98.5 ± 950.41	412 (98.8%)	51.92 ± 319.07	1
VB12	312 (47.13%)	610.9 ± 421.27	292 (70.02%)	852.07 ± 511.04	<0.0001
WBC	662 (100.0%)	7.55 ± 4.48	417 (100.0%)	8.85 ± 14.28	1
Temperature	662 (100.0%)	37.016 ± 1.91	417 (100.0%)	37.63 ± 0.92	<0.0001
Pulse	662 (100.0%)	86.23 ± 16.66	417 (100.0%)	87.59 ± 17.23	1
Respiratory Rate	513 (77.49%)	19.84 ± 9.0	113 (27.1%)	20.98 ± 9.74	1
SBP	662 (100.0%)	131.7 ± 24.65	417 (100.0%)	136.94 ± 23.69	0.0295
DBP	662 (100.0%)	75.53 ± 13.23	417 (100.0%)	75.96 ± 15.34	1
Saturation	110 (16.62%)	94.7 ± 5.92	415 (99.52%)	92.75 ± 7.31	0.5151

Table 3: Minimum and maximum accepted values of the dynamic features. Feature engineering was applied for the bolded features. 'AR' and 'V' refer to arterial and venous blood, respectively

Feature	Min	Max	Units	Feature	Min	Max	Units
HBA1C (#)	0	240	mmol/mol	**Lymphocytes (%)**	0.2	100	%
HBA1C (%)	0	24	%	**Neutrophils (#)**	0.1	60	10e3/µL
Albumin	0	100	g/L	**Neutrophils (%)**	0.2	100	%
ALT	0	20000	U/L	NRBC	0	100	%
AST	0	20000	U/L	Osmolality (urine)	50	2000	mosmo/kg
Indirect bilirubin	0	20	mg/dL	**PO2 (AR)**	0	1000	mmHg
Direct bilirubin	0	20	mg/dL	PO2 (V)	0	1000	mmHg
BNP	0	10000	PG/ML	PCO2 (AR)	0	150	mmHg
Respiratory rate	1	100	BPM	PCO2 (V)	0	150	mmHg
BUN	2	200	mg/dL	PH	6.6	7.8	
Calcium	0	20	mg/dL	**Platelet**	0	1000	10e3/µL
CKMB	0	10000	U/L	Potassium	1	10	mmol/L
CPK	0	10000	U/L	**PTT**	5	200	Sec
CRP	0	1000	mg/L	Pulse	10	300	BPM
Creatinine	0	20	mg/dL	RBC	1	8	10e6/µL
DBP	20	240	mmHG	RDW	5	40	%
D-dimer	0	50	FEU mg/L	SBP	40	250	mmHG
Ferritin	0	20000	ng/ml	Sodium	110	200	mmol/L
Fibrinogen	0	1500	mg/dL	Saturaion O2 (AR)	5	100	%
Glucose	0	2000	mg/dL	**Saturation**	5	100	%
HCO3	0	100	mmol/L	Total bilirubin	0	20	mg/dL
HGB	2	25	g/dL	**Temperature**	20	43	C°
INR	0.5	5		Triglycerides	10	2000	mg/dL
Lactate	0.2	15	mmol/L	**Troponin**	1	40000	ng/L
LDH	0	50000	U/L	Vitamin B12	100	2500	pg/ml
Lymphocytes (#)	0	20	10e3/µL	**WBC**	0.2	100	10e3/µL

Table 4: Top 100 features in importance as calculated by XGBoost. SD: standard deviation; /: ratio between two features. 24h,72h: time windows which the statistic was computed (if not mentioned, the statistics is calculated on the entire hospitalization period so far).

Age	Fibrinogen delta mean	Neutrophils (#) min
BMI	Fibrinogen delta mean 24h	Neutrophils (#) min 72h
Lactate	Fibrinogen max 72h	Neutrophils (#) / Glucose
Neutrophils (%)	Fibrinogen mean	Neutrophils (#) / Platelet
Sodium	Glucose mean	Neutrophils (#) trend
Saturaion O2 - arterial blood	Glucose min 72h	Neutrophils (%) max
Albumin mean 24h	Glucose / Troponin	Neutrophils (%) max 24h
Albumin min 24h	Glucose SD	Neutrophils (%) max 72h
Albumin min 72h	LDH max	Neutrophils (%) min
Albumin / PTT	LDH max 72h	Neutrophils (%) min 72h
Albumin SD	LDH mean	Neutrophils (%) / ALT
Albumin SD 72h	LDH mean 72h	Neutrophils (%) / AST
ALT / Fibrinogen	LDH min 72h	Neutrophils (%) / D-dimer
AST min 72h	LDH / Albumin	Platelet delta
AST / Platelet	LDH / ALT	Platelet SD 24h
AST SD 72h	LDH / Platelet	Platelet SD 72h
BUN lr slope	LDH SD 72h	PTT lr slope
BUN delta mean 72h	Lymphocytes (#) min	PTT max 24h
BUN min	Lymphocytes (#) / D-dimer	PTT mean 72h
BUN / ALT	Lymphocytes (#) / Ferritin	PTT min 24h
BUN / Ferritin	Lymphocytes (#) / PTT	Temperature max
BUN / Troponin	Lymphocytes (#) SD 72h	Temperature mean 72h
BUN SD	Lymphocytes (%) max	Temperature min
D-dimer max	Lymphocytes (%) max 24h	Temperature min 24h
D-dimer max 72h	Lymphocytes (%) mean 72h	Temperature min 72h
D-dimer min	Lymphocytes (%) / AST	Temperature SD
D-dimer min 72h	Lymphocytes (%) / D-dimer	Temperature SD 72h
D-dimer / Albumin	Lymphocytes (%) / Fibrinogen	Troponin delta mean
D-dimer / AST	Lymphocytes (%) / PTT	Troponin SD 72h
D-dimer / Ferritin	Lymphocytes (%) SD	WBC min 24h
D-dimer / Fibrinogen	Lymphocytes (%) SD 24h	WBC / D-dimer
D-dimer / Platelet	Neutrophils (#) max	WBC SD
D-dimer SD	Neutrophils (#) max 72h	
Ferritin / Troponin	Neutrophils (#) mean	