


**Please cite the Published Version**

Chang, K, Jia, T, Zhou, YN, Shu, ZX, Liu, JF, Sun, J, Zheng, QG, Tian, HY, Xia, JN, Yang, K, Wang, N, Sun, HL, Wang, XY, Yan, DY, Clark, T, Liu, BY, Li, XD, Peng, YH  and Zhou, XZ (2023) Validation and refinement of two interpretable models for coronavirus disease 2019 prognosis prediction. World Journal of Traditional Chinese Medicine, 9 (2). pp. 191-200. ISSN 2311-8571

**DOI:** <https://doi.org/10.4103/2311-8571.372326>

**Publisher:** Medknow

**Version:** Published Version

**Downloaded from:** <https://e-space.mmu.ac.uk/635034/>

**Usage rights:**  [Creative Commons: Attribution-Noncommercial-Share Alike 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/)

**Additional Information:** This is an open access article which first appeared in World Journal of Traditional Chinese Medicine

**Enquiries:**

If you have questions about this document, contact [openresearch@mmu.ac.uk](mailto:openresearch@mmu.ac.uk). Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

# Validation and Refinement of Two Interpretable Models for Coronavirus Disease 2019 Prognosis Prediction

Kai Chang<sup>a,b</sup>, Ting Jia<sup>a</sup>, Ya-Na Zhou<sup>c</sup>, Zi-Xin Shu<sup>a</sup>, Ji-Fen Liu<sup>c</sup>, Jing Sun<sup>c</sup>, Qi-Guang Zheng<sup>a</sup>, Hao-Yu Tian<sup>a</sup>, Jia-Nan Xia<sup>a</sup>, Kuo Yang<sup>a</sup>, Ning Wang<sup>a</sup>, Hai-Long Sun<sup>a</sup>, Xin-Yan Wang<sup>a</sup>, Deng-Ying Yan<sup>a</sup>, Taane G. Clark<sup>d,e</sup>, Bao-Yan Liu<sup>f</sup>, Xiao-Dong Li<sup>g</sup>, Yong-Hong Peng<sup>h</sup>, Xue-Zhong Zhou<sup>a</sup>

<sup>a</sup>Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, <sup>b</sup>College of Information Engineering, Hubei University of Chinese Medicine, Wuhan, Hubei, China, <sup>c</sup>Department of Data Center, Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, Hubei, China, <sup>d</sup>Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, England, <sup>e</sup>Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, England, <sup>f</sup>Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing, China, <sup>g</sup>Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, England

## Abstract

**Objective:** To validate two proposed coronavirus disease 2019 (COVID-19) prognosis models, analyze the characteristics of different models, consider the performance of models in predicting different outcomes, and provide new insights into the development and use of artificial intelligence (AI) predictive models in clinical decision-making for COVID-19 and other diseases. **Materials and Methods:** We compared two proposed prediction models for COVID-19 prognosis that use a decision tree and logistic regression modeling. We evaluated the effectiveness of different model-building strategies using laboratory tests and/or clinical record data, their sensitivity and robustness to the timings of records used and the presence of missing data, and their predictive performance and capabilities in single-site and multicenter settings. **Results:** The predictive accuracies of the two models after retraining were improved to 93.2% and 93.9%, compared with that of the models directly used, with accuracies of 84.3% and 87.9%, indicating that the prediction models could not be used directly and require retraining based on actual data. In addition, based on the prediction model, new features obtained by model comparison and literature evidence were transferred to integrate the new models with better performance. **Conclusions:** Comparing the characteristics and differences of datasets used in model training, effective model verification, and a fusion of models is necessary in improving the performance of AI models.

**Keywords:** Coronavirus disease 2019, decision tree, interpretable models, logistic regression, prognosis prediction

## INTRODUCTION

According to the World Health Organization, there have been over 240 million confirmed coronavirus disease 2019 (COVID-19) cases and 4.8 million deaths worldwide as of October 19, 2021. The person-to-person transmission and serious consequences, including pneumonia and death, have emerged as the enormous threats to human health.<sup>[1,2]</sup> Many countries and regions still suffer from the pandemic, which imposes a significant and continuing public health burden. Clinical and public health resources have been stretched, and the resulting shortage of medical resources has inevitably affected the outcomes of COVID-19 treatment. Moreover, the high prevalence of severe cases has caused great pressure on medical services due to the shortage of intensive care resources.

Early and accurate identification of severe COVID-19 and timely treatment is the key in reducing the fatality rate.<sup>[3,4]</sup> Many

potential factors are related to the prognosis of COVID-19, including age, biochemical parameters,<sup>[5-11]</sup> comorbidities,<sup>[12-14]</sup> and diet.<sup>[15]</sup> Artificial intelligence (AI) approaches have been deployed to build the prognostic models using laboratory and clinical parameters to assist in clinical decision-making.<sup>[16]</sup> For instance, Yan *et al.* used the XGBoost decision tree model to determine the blood laboratory test parameters to predict

**Address for correspondence:** Prof. Xue-Zhong Zhou, Institute of Medical Intelligence, School of Computer and Information Technology, Xizhimenwai Street, Haidian District, Beijing Jiaotong University, Beijing 100063, China. E-mail: xzzhou@bjtu.edu.cn

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

**For reprints contact:** WKHLRPMedknow\_reprints@wolterskluwer.com

© 2023 World Journal of Traditional Chinese Medicine | Published by Wolters Kluwer - Medknow

**Received:** 31-10-2021, **Accepted:** 11-12-2021, **Published:** 23-03-2023

**How to cite this article:** Chang K, Jia T, Zhou YN, Shu ZX, Liu JF, Sun J, *et al.* Validation and refinement of two interpretable models for coronavirus disease 2019 prognosis prediction. World J Tradit Chin Med 2023;9:191-200.

survival rates with 97% accuracy.<sup>[17]</sup> Using similar laboratory data, Wang *et al.* developed a multivariate logistic regression model for prognosis, with a high predictive ability of 93% accuracy.<sup>[18]</sup> Liu *et al.* investigated the prognostic effects of blood and biochemical laboratory test parameters on severe COVID-19 and its adverse clinical outcomes.<sup>[3]</sup> Shang *et al.* found that the neutrophil-to-lymphocyte ratio (NLR), C-reactive protein (CRP), and platelets could be used to effectively assess the severity of COVID-19, among which NLR was the best predictor of severe COVID-19.<sup>[19]</sup> In addition, He *et al.* investigated the role of tumor biomarkers for lung cancer as the predictive indicators for clinical outcomes in COVID-19 patients and found that the concentrations of carcinoembryonic antigen CYFRA21-1 and squamous cell carcinoma antigen could accurately predict the clinical outcomes.<sup>[20]</sup> Furthermore, Zhang *et al.* showed that male sex, comorbidity, lymphopenia, and elevated CRP levels were the independent risk factors for poor prognosis in COVID-19 patients, facilitating the early identification and stratification of high-risk COVID-19 patients.<sup>[21]</sup>

Machine learning often requires data preprocessing and cleaning to configure the data to fit within a specific model.<sup>[22]</sup> Despite the development of diagnostic and prognostic models for COVID-19, few have been externally validated, and their reported performances are often optimistic and incorporate high rates of bias, causing concerns about their real-world clinical use.<sup>[23]</sup> Developing a universal prediction model in the context of limited selected data lacks feasibility, particularly in applying it across countries or clinical settings for new diseases such as COVID-19. One effective method is to incorporate the human intelligence into a decision or prediction model, potentially making the prediction model generalizable for new cases when empirical knowledge is employed.

In this study, we validated two COVID-19 prognosis prediction models, a decision tree<sup>[17]</sup> and a logistic regression,<sup>[18]</sup> using the new, real-world clinical data of 944 confirmed COVID-19 patients derived from three hospitals in Wuhan, China. We evaluated their limitations in the clinical application as well as the factors affecting prediction errors.

We processed the first occurrence data (the first clinical test result after admission) and last occurrence data (the last clinical test result before discharge or in-hospital death) of patient cases and evaluated the influence of the two features on prognosis.

Based on our insights, we provided research ideas for the improvement and application of AI predictive models.

## MATERIALS AND METHODS

### Dataset and its preprocessing

Electronic medical records of 944 COVID-19 patients admitted to three hospitals in Wuhan, China between November 25, 2019, and March 18, 2020, were retrieved. According to the diagnosis and treatment protocol for novel coronavirus pneumonia (trial version 7),<sup>[24]</sup> patients were initially divided

into two groups on admission: severe and nonsevere. Of the cases, 865 recovered from COVID-19 within an average of 20.62 days and were discharged from the hospital, while the remaining 79 died within an average of 13.54 days [Table 1].

The raw data consisted of 138 variables or features including the blood test parameters and clinical information of the patients, with most patients having multiple blood samples collected during their hospital stay. Clinically, predicting patient prognosis as early as possible is vital in making appropriate clinical decisions in a timely manner, and using data collected at admission for treatment and effective management of resources in the hospital is particularly crucial. Across the dataset, some data were missing and 94 variables with >25% missing values were excluded. Missing values of the remaining 44 features were imputed using the predictive mean-matched average values. The data processing steps are summarized in Figure 1.

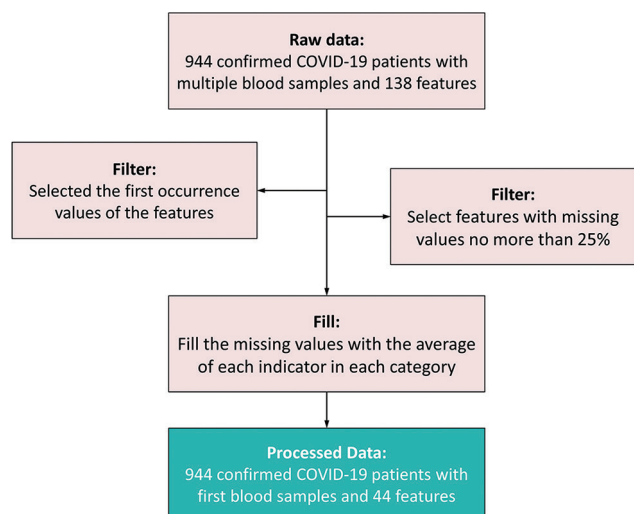
### Prediction models

We applied two representative approaches to COVID-19 prediction: (1) A decision tree and (2) logistic regression models. Decision trees are commonly used for the

**Table 1: The clinical characteristics of our data (944 cases)**

Characteristics	Overall
Age (years), mean (SD)	56.82 (15.6)
Gender, <i>n</i> (%)	
Male	489 (51.8)
Female	455 (48.2)
Severity, <i>n</i> (%)	
Severe group	188 (19.9)
Nonsevere group	756 (80.1)
Outcomes, <i>n</i> (%)	
Survival	865 (91.6)
Death	79 (8.4)
Hospital stay (days), mean (SD)	
All	20.03 (9.2)
Survival	20.62 (9.0)
Death	13.54 (9.3)
Number of people in each hospital, <i>n</i> (survival/death)	
Hospital 1	318 (265/53)
Hospital 2	377 (366/11)
Hospital 3	249 (234/15)
Comorbidities, <i>n</i> (%)	
Hypertension	267 (28.3)
Diabetes	133 (14.1)
Coronary heart disease	72 (7.6)
Stroke	43 (4.6)
Gastritis	34 (3.6)
Chronic bronchitis	33 (3.5)
Hyperlipidemia	33 (3.5)
Heart disease	25 (2.6)
Anemia	23 (2.4)
Arrhythmia	21 (2.2)

SD: Standard deviation



**Figure 1:** Summary of data processing: Step 1, raw data includes 944 confirmed COVID-19 patients with multiple blood samples and 138 features. Step 2, select the value of the first occurrence of each feature. Step 3, select features that were missing no more than 25%. Step 4, fill in the remaining missing data with the average of each indicator in different prognostic outcomes. Step 5, processed data includes 944 confirmed COVID-19 patients with first blood samples and 44 features. COVID-19: Coronavirus disease 2019

classification.<sup>[25]</sup> The most appreciated advantage of the decision-tree model is its interpretability, which offers a visible decision path for each new input case. The decision tree model is used to find an optimal feature and an optimal candidate value that divides the dataset into two sub-datasets according to the optimal candidate value. A decision tree was generated by repeating the above operations until the specified conditions were met. XGBoost is a high-performance machine learning algorithm for decision tree induction that has been widely used in the medical field as a recursive tree-based decision model.<sup>[13,17]</sup>

The logistic regression algorithm is a type of machine learning classification algorithm that is considered a generalized linear regression model for classification.<sup>[26]</sup> This algorithm has been used in many fields, such as economic forecasting and medical diagnosis. For independent variables  $X(x_1, x_2, \dots, x_i)$  and dependent binary variable  $Y(1 = \text{death}, 0 = \text{control})$ , logistic regression describes their relationship (expressed as the  $h$  function), which can be used to predict  $Y$ , and the thetas are estimated coefficients (log odds ratios) (formulas 1 and 2).

$$P(Y=1) = h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}} \quad (1)$$

$$\theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_i x_i, \quad (2)$$

## Evaluation

The performance of the prediction models was evaluated by assessing classification accuracy, precision, recall, F1 scores,

area under the receiver operating characteristic curve area under curve (AUC), and survival analysis. The formulae are as follows:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (3)$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (4)$$

$$F1_i = \frac{2 * \text{Precision}_i * \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \quad (5)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

The AUC value is a probability value indicating the reliability of the prediction model. When the value of AUC increases to 1, the prediction model has a more reliable prediction performance. Survival analysis was used to investigate the distribution of survival time under different factors. To further quantify the power of the prognosis models, we used the Kaplan-Meier (KM) method to estimate the survival probability from the observed survival time for each prognosis prediction model in our study.

## Experimental setting

### Evaluating the performance of applying existing models on our data

To evaluate the performance and identify the problems in using prediction models in the diagnosis and treatment of COVID-19, we first tested the performance of models by Yan *et al.* and Wang *et al.*, which used a decision tree and logistic regression model, respectively, on our three hospital datasets.<sup>[17,18]</sup> Undoubtedly, the use of the last appearance data had an improved prediction performance compared to the use of the first appearance data. However, regardless of the improved performance, the last appearance data, which were often shortly before discharge or death, had limited clinical use. The ability to accurately predict in-hospital mortality in patients at the time of admission could improve clinical and operational decision-making and outcomes.<sup>[27]</sup> Therefore, we chose the value of the first occurrence of each feature to test the performance of the Yan *et al.* and Wang *et al.* models.

### Evaluating the performance of the models retrained by our across-hospital data

To evaluate the performance of applying the machine learning algorithms to our new data, we randomly divided our across-hospital data ( $n = 944$ ) into training ( $n = 472$ ) and testing ( $n = 472$ ) sets. The training set (including 433 survival and 39 death cases) was used to retrain the decision tree and logistic regression model, and the test set was used to test these models.

### Evaluating the performance of the integrated model

We considered combining the multiple models and features to improve the usability of the model. The decision tree model



is widely used because of its good interpretability, which can easily obtain the characteristics of groups at high-risk of death. Therefore, we integrated the features in benchmark-dt on the basis of the new-dt model and slightly modified the threshold of the decision node to obtain a new model. We then tested our test set on the integrated model. The entire experimental setting is presented in Figure 2.

## RESULTS

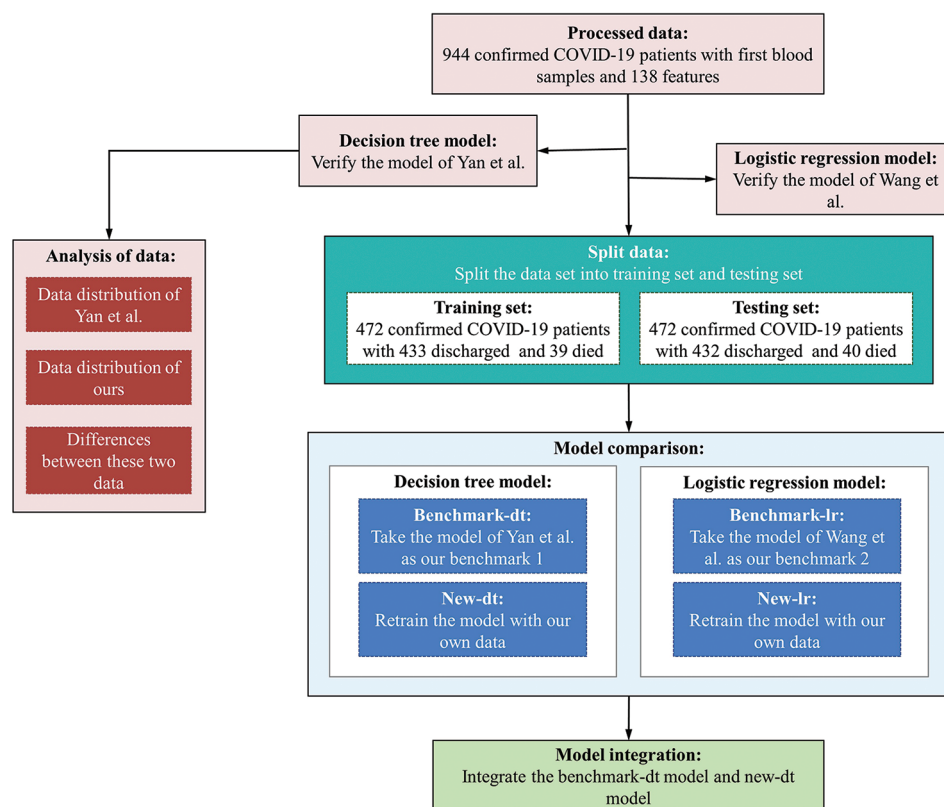
### Poor generalizability of the prediction model

The benchmark-dt model consisted of lactate dehydrogenase (LDH), lymphocytes (%), and high-sensitivity CRP (hs-CRP), while the benchmark-lr model contained neutrophils, lymphocytes, platelets and Interleukin-2 receptor (IL-2R). The new dt model had two features (LDH and Urea). The new model involved three features: (1) LDH, (2) lymphocytes (%), and (3) hs-CRP.

To further illustrate this, the actual performance of the benchmark-dt model regarding our data on the 944 cases is detailed in Table 2. An overall accuracy of 94.7% (precision of 99% for survival prediction and 62% for mortality prediction) was achieved using the last occurrence data. However, for the first occurrence data, an overall prediction accuracy of 83.6% was achieved (with 99% precision

for survival prediction and 32% precision for mortality prediction). The number of incorrectly predicted cases using the last occurrence data and the first occurrence data was 50 and 155, respectively [Table 3], in which the majority of mis-predicted patients (38/50) using the last occurrence data were also mis-predicted using the first occurrence data. Because the data were imbalanced, we chose the weighted average instead of the macro average to calculate the “Both” performance.

Although we obtained high performance (accuracy of 97%) in the original dataset, the benchmark-dt model achieved low performance in our across-hospital data (accuracy of 94.7% in the last occurrence records and 83.6% in the first occurrence records). Unsurprisingly, using the latest measurements may yield a higher accuracy in predicting an outcome. To further investigate the reason for the decrease in performance of the benchmark-dt model, we compared the difference in data distributions between the training data used in benchmark-dt and our study. We found that there were significant differences in LDH, lymphocytes, and hs-CRP levels ( $P < 0.0001$ ) that were employed in the decision tree model, which might explain the low performance of benchmark-dt and validate our new data. In addition, with similar validation steps, the benchmark-lr model (with 90.74% sensitivity and 94.44% specificity) achieved an overall prediction



**Figure 2:** The experimental setting. Step 1, 944 confirmed COVID-19 patients with blood samples (features first appeared). Steps 2 and 3, use our total data 944 to verify the existing two models (the decision tree and the logistic regression models). Step 4, analyze the data distribution of Yan *et al.* and ours, then compared the differences. Step 5, divide our processed data into training and test sets based on a ratio of 1:1. Step 6, use two models as our two benchmarks and retrain two new models with our training set. Step 7, integrate the two decision tree models. COVID-19: Coronavirus disease 2019

**Table 2: Performance of the benchmark-dt model on our new data (944 cases)**

Data record used	Outcome	Precision	Recall	F1 score	Support	Accuracy
Last occurrence data	Survival	0.99	0.95	0.97	865	0.947
	Death	0.62	0.92	0.74	79	
	Both	0.96	0.95	0.95	944	
First occurrence data	Survival	0.99	0.83	0.90	865	0.836
	Death	0.32	0.87	0.47	79	
	Both	0.93	0.83	0.86	944	

**Table 3: Comparison of the number of wrong predicted cases using the last occurrence data and the first occurrence data**

Data record used	Survival	Death	All
Last occurrence data	44	6	50
First occurrence data	145	10	155
Intersection	33	5	38

**Table 4: Performance of the benchmark-lr based on our first occurrence data**

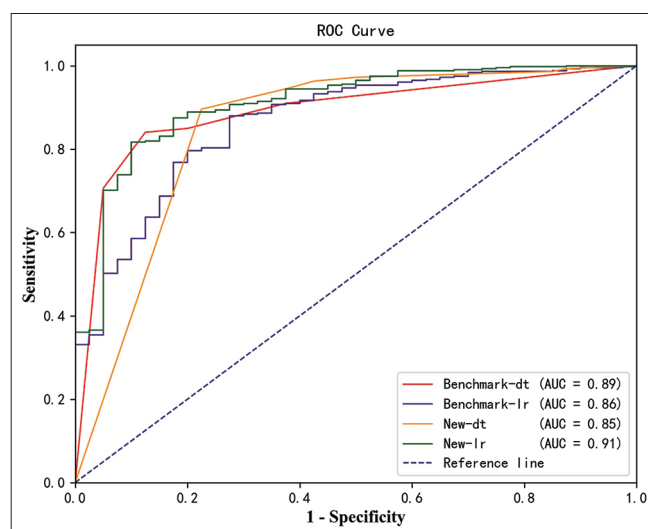
Outcome	Precision	Recall	F1 score	Support	Accuracy
Survival	0.97	0.90	0.93	865	0.875
Death	0.38	0.65	0.46	79	
Both	0.92	0.88	0.89	944	

accuracy of 87.5% (97% precision on survival and 38% on mortality) [Table 4] for the first occurrence data of the 944 cases.

The clinical utility of predictive models for decision-making may vary greatly as they are likely to perform differently across centers, settings, and time,<sup>[28]</sup> as well as requiring “external validation” on different datasets.<sup>[29]</sup> Our external validations of the two benchmark models showed a particularly low precision (<50%) for mortality predictions, which might partially be due to the high mortality ratio (46.4% in benchmark-dt and 50% in benchmark-lr) incorporated in the original datasets. This indicates that a derived dataset with a natural ratio of mortality, where the mortality ratio is comparable to the actual conditions of a specific disease, would be important for model construction and generalization.

### Improved performance with retraining on new data

To investigate the potential for performance improvement of AI models on new datasets, we retrained the two proposed models on our new data using first occurrence features with a 50:50 split of training (472 cases) and test (472 cases) sets [Table 5 and Figure 3]. Among the four models, the benchmark-dt model achieved the highest recall of 88% for mortality prediction, but the lowest precision of 34% for mortality prediction (472 cases), indicating that the benchmark-dt model was more likely to predict death. By retraining the benchmark-dt model on our training set, the prediction accuracy of the new-dt model increased from 84.3% to 93.2% in our test set. This result indicated that retraining a prediction model using new data can greatly improve the applicability of the model.<sup>[23]</sup>



**Figure 3:** ROC curves of different prediction models. Receiver operating characteristic analysis shows the performance of the four models (benchmark-dt, benchmark-lr, new-dt, and new-lr). benchmark-dt has the lowest AUC of 89% and new-lr has the highest AUC of 91%. ROC: Receiver operating characteristic, AUC: Area under curve

By retraining the benchmark-lr model on our training set, the precision of new-lr for mortality prediction increased from 38% to 74%. Additionally, the recall and F1-scores for survival prediction increased to 99% and 97%, respectively. The new-lr model achieved the highest performance among the four models, with 93.9% and 91% of accuracy and AUC, respectively.

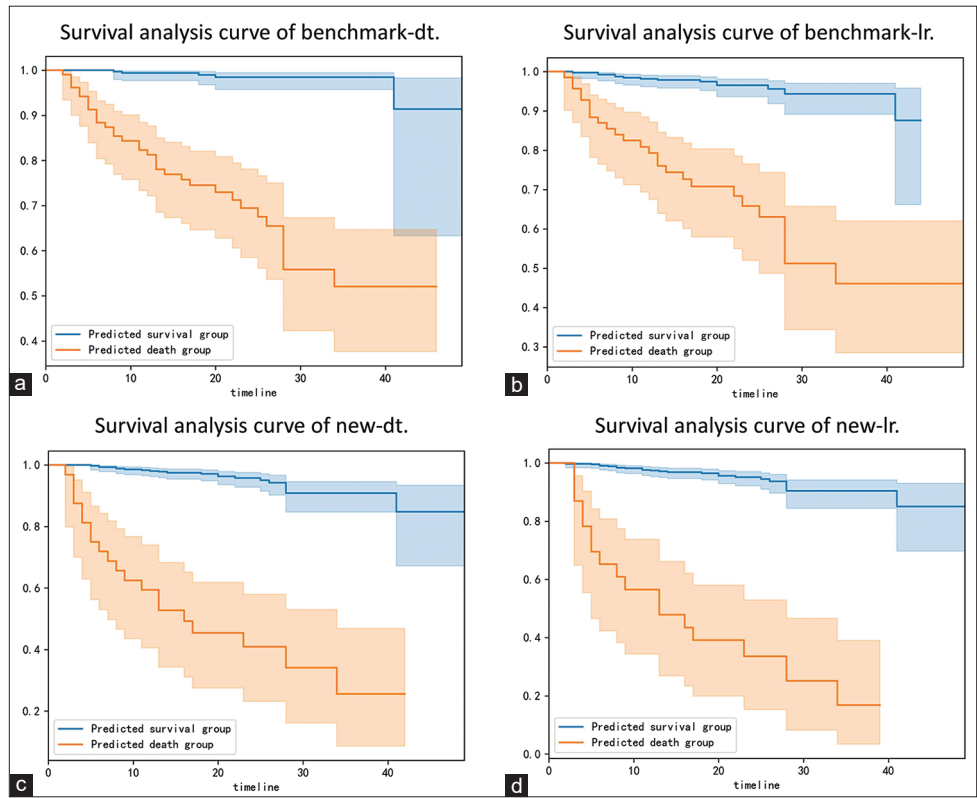
Survival analysis using the KM method indicated significantly consistent improvement from the benchmark model to the retrained models [ $P = 4.67\text{E-}34$  vs.  $8.91\text{E-}24$  for decision tree (DT) and  $2.55\text{E-}38$  vs.  $5.19\text{E-}20$  for logistic regression (LR), Figure 4]. In particular, the difference in the new-lr model was the most significant, with a log-rank test value of  $2.55\text{E-}38$ . A comparison of the four log-rank test results revealed that the retrained models, new-dt, and new-lr, had a greater difference between the predicted survival group and the predicted death group than the benchmark models, benchmark-dt, and benchmark-lr.

### Evolution of integrated models bring practical results

We integrated the lymphocyte (%) node of benchmark-dt into new-dt and incorporated more features, which may help discover more high-risk groups [Figure 5]. However, the first

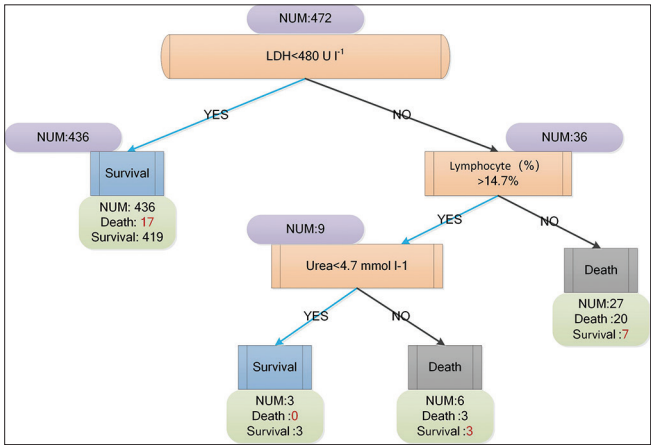
Table 5: The test performance of five coronavirus disease 2019 prognosis models on first occurrence data							
Model	Outcome	Precision	Recall	F1-score	Support	Accuracy	AUC
Benchmark-dt <sup>a</sup>	Survival	0.99	0.84	0.91	432	0.843	0.89
	Death	0.34	0.88	0.49	40		
	Both	0.93	0.84	0.87	472		
Benchmark-lr <sup>b</sup>	Survival	0.97	0.90	0.93	432	0.879	0.86
	Death	0.38	0.65	0.48	40		
	Both	0.92	0.88	0.89	472		
New-dt <sup>c</sup>	Survival	0.95	0.97	0.96	432	0.932	0.85
	Death	0.62	0.50	0.56	40		
	Both	0.92	0.93	0.93	472		
New-lr <sup>d</sup>	Survival	0.95	0.99	0.97	432	0.939	0.91
	Death	0.74	0.42	0.54	40		
	Both	0.93	0.94	0.93	472		
Integrated model	Survival	0.98	0.96	0.97	432	0.943	
	Death	0.70	0.58	0.63	40		
	Both	0.96	0.93	0.94	472		

<sup>a</sup>Decision tree model of Yan *et al.*, <sup>b</sup>Logistic regression model of Wang *et al.*, <sup>c</sup>Decision tree model of our study, <sup>d</sup>Logistic regression model of our study. AUC: Area under the curve



**Figure 4:** Survival analysis curves of four models. (a), Survival analysis curve of benchmark-dt (log-rank test of 8.91E-24). (b), Survival analysis curves of benchmark-lr (log-rank test of 5.19E-20). (c), Survival analysis curve of new-dt (log-rank test of 4.67E-34). (d), Survival analysis curves of new-lr (log-rank test of 2.55E-38). Each survival analysis chart consists of two survival curves according to the predicted outcomes, namely the predicted survival group and the predicted death group. The solid blue line is the predicted survival group, and the light blue band represents the 95% confidence interval. While the orange solid line is the predicted death group, and the light orange band represents the 95% confidence interval

type of high-risk population in the new-dt had the lowest number of deaths. When LDH was <481 U/L, only a small number of patients (19/442) died in the training set. When we made a more detailed division, the number of mis-predicted patients was reduced in the training set but increased in the test set, indicating that the stability of the branch was not high. To increase the stability of the model, we designated patients with LDH <481 U/L as those that survived. Although this may



**Figure 5:** Decision tree of model integration. The purple and green boxes represent the classification results of the model on our test set. NUM, the number of patients in a class; Death, the number of deaths in a class; Survival, the number of people discharged from hospital in a class

have reduced the predictive performance of the model for death patients, the accuracy and stability of the model for the entire dataset were enhanced.

In performing on our test data, the integrated model achieved the highest F1-scores for survival prediction and mortality prediction, at 97% and 63%, respectively [Table 5]. In addition, the prediction accuracy of our test set increased from the previous value of 93.2% to 94.3%. The number of mis-predicted patients in the integrated model and the previous four models on our training and test sets is presented in Table 6. The integrated model performed relatively well on both the training and test sets, with an optimal performance on the test set (a minimum number of mis-predicted patients of 27). Although new-dt had the best performance on the training set (minimum number of mis-predicted patients of 22), it was greatly reduced on the test set (32 mis-predicted patients), resulting in overfitting. Based on performances on the entire dataset, the integrated model demonstrated higher stability.

The integrated model incorporated LDH, urea, and lymphocyte (%) levels as the decision features. LDH levels reflect the extent of various pathophysiological processes,<sup>[30]</sup> and elevated LDH levels at admission are an independent risk factor in the severity and mortality of COVID-19. Therefore, LDH may assist in the early evaluation of COVID-19.<sup>[31]</sup> In addition, severe COVID-19 patients had higher levels of urea and urine protein, whereas uric acid levels were lower, reflecting poor kidney function.<sup>[32]</sup> Yan *et al.* also suggested that lymphocytes may serve as the potential therapeutic targets.<sup>[17]</sup> These studies indicate that LDH, urea, and lymphocyte (%) levels are the important risk factors in the severity and mortality of COVID-19.

### Dynamic evaluation of the retrained model's prediction bias

Notably, the number of mis-predicted patients with new-dt and new-lr had its own trend [Table 7]. The new-dt model predicted

Table 6: Number of misjudged patients of different models in the training set and test set		
Model	Training set	Test set
Benchmark-dt	81	74
Benchmark-lr	61	57
New-dt	22	32
New-lr	32	29
Integrated model	28	27

more survival patients as death patients, whereas the new-lr model tended to mis-predict death cases as survival cases better than the new-dt model. Thus, the new-dt model was more likely to predict death compared with the new-lr model.

The majority of cases misjudged by the new-dt model were in serious condition on admission, with poor physical conditions, abnormal blood sample indicators, and long hospital stays. However, with continuous treatment, the conditions of these patients significantly improved and they were ultimately discharged. We further evaluated improvements in the continued use of the prediction model and found that the prognoses of several patients were mis-predicted when admitted to the hospital [Figure 6], demonstrating the impact of treatment on the patient's final outcome as well as the possibility of using models dynamically to improve the therapeutic outcomes.

### DISCUSSION

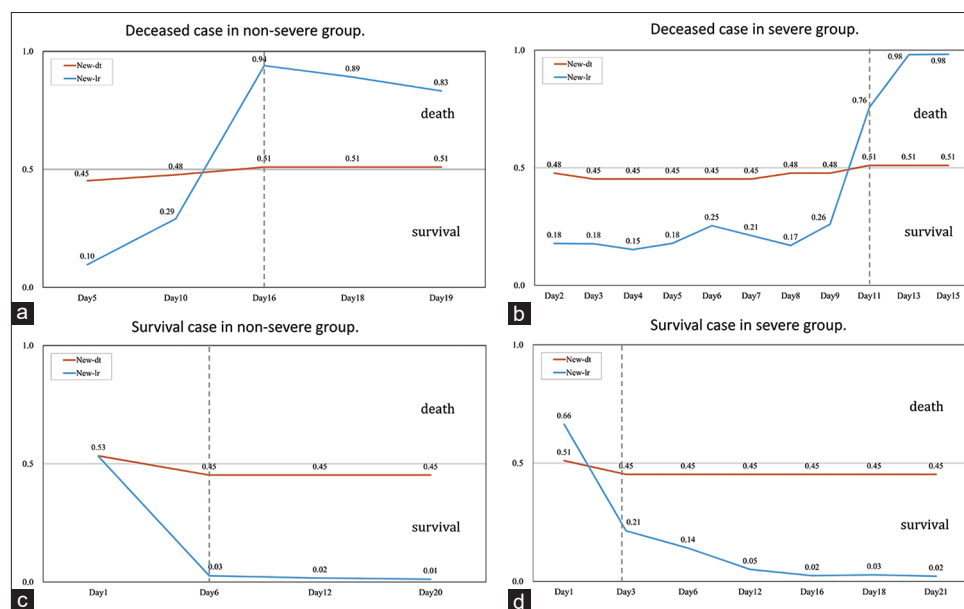
COVID-19 has put a massive strain on the clinical and public health resources. Critical decision-making and identifying patients at risk of developing severe disease are particularly important.<sup>[33]</sup> Clinical predictive models use patient data to determine the probability of a current disease condition and its outcomes,<sup>[23]</sup> and nontransferability of predictive models can cause significant concerns and misuse in clinical practice.<sup>[16]</sup> To ensure the clinical utility of prediction models, it is vital that they are rigorously developed, validated, and evaluated.<sup>[34,35]</sup>

For clinical prediction tasks, particularly for prognoses, an interpretable AI model may improve clinical adoptability and feasibility.<sup>[36]</sup> Here, we used two classical interpretable models (a decision tree and a logistic regression) for COVID-19 prognosis prediction. More than 50% of these models are logistic regression models, used to investigate concrete situations and potential solutions for model transfers to a new data setting. By utilizing a real-world, independent dataset and performing practical external validation experiments, our study highlighted and validated two well-recognized factors affecting the transferability of prediction models. This may, in turn, help improve their usability in clinical practice. First, limited and nonrepresentative homogenous benchmark data that incorporate patient subgroups with obvious clinical characteristics or bias is one of the main factors that make proposed prediction models unfeasible in real-world clinical settings. For instance, the significant differences of key laboratory parameters, such



**Table 7: The number of mis-predicted patients of new-dt and new-lr**

The number of mispredicted patients	Mild group			Severe group			All
	Survival	Death	All	Survival	Death	All	
new-dt	3	8	11	9	12	21	32
new-lr	2	9	11	4	14	18	29



**Figure 6:** The variability of prediction power using data from different time points. The x-axis represents the length of hospitalization in terms of days, and the y-axis represents the PMP. For both decision tree and logistic regression models, the case is predicted as death when the PMP is higher than 0.5, and otherwise predicted as survival. Results of four typical patients are presented. (a), deceased case in non-severe group. (b), deceased case in severe group. Both models produce wrong predictions for the time points of 5<sup>th</sup> and 10<sup>th</sup> days and later correct when using data from 16<sup>th</sup> day. (c), survival case in non-severe group. (d), survival case in severe group. Results indicate the importance and difference of using data from different time points during hospitalization. PMP: Predicted mortality possibility

as LDH ( $P = 3.49E-28$ ), lymphocytes ( $P = 4.32E-24$ ), and hs-CRP ( $P = 6.22E-19$ ), between the dataset of Yan *et al.* and our data confirmed this idea. In addition, this may explain why the model parameters (i.e., the threshold of partitioning in the decision node for the DT model) of Yan's prediction model could not be used directly for our data. Moreover, the high mortality rate in Yan *et al.*'s data indicated that their clinical data were mainly derived from severe COVID-19 cases, which does not represent the real situation of COVID-19 in China and worldwide. Second, prognosis models specifically require the capability of early predictions (e.g., 2 weeks earlier for risk prediction). The benchmark-dt model was mainly derived from the last occurrence data points, which were close in time to the final outcomes (i.e., deceased or discharged), and lacked usability in clinical practice as no therapies or treatments that affect outcomes could be utilized in as narrow of a window. By incorporating empirical knowledge on the laboratory parameters found in COVID-19 literature, we proposed a manually tuned new DT model for COVID-19 prognosis prediction with comparably high performance at 94.3% accuracy and robustness with both training and test sets, where the average length of hospital stay is 20.03 days.

This model also had early prediction capabilities, indicating the feasibility and importance of incorporating the latest evidence or background knowledge into an AI model to improve usability and performance. However, although laboratory parameters (e.g., LDH) have acceptable prediction capabilities for prognosis, as demonstrated through the investigation of mis-predicted cases, a multi-model fusion system that includes factors such as patient characteristics (e.g. age, gender, and comorbidities) and clinical manifestations is needed to improve COVID-19 predictions. Furthermore, differences in the data distributions across sites may lead to the need for training models for different patient groups (e.g., age-based or ethnicity-based).<sup>[37]</sup> However, as there may not be sufficient data for each patient group to train multiple models, a new transfer learning would be useful in continuously adjusting and updating models based on new data.<sup>[38]</sup>

This study had several limitations. First, only two representative models were used, and the integration of more effective models integration into a multi-model fusion system to support clinical decision making may be useful. A systematic framework for validating, comparing, improving, and updating prediction

models should be developed further rather than the creation of new models.<sup>[16]</sup>

## CONCLUSIONS

The use of AI predictive models to assist in diagnoses and treatment decision-making has been effective with COVID-19. However, an urgent need to validate prediction models that were trained using small data, to enable their applicability in clinical practice remains. Appropriate and effective use of these models may lead to preemptive treatment, prevention of severe outcomes, and improvements in prognoses.

In this study, we proposed new methods to improve the effectiveness of these models. Knowledge of the dataset characteristics used in the training of the model is vital, in order to understand the differences in the training data from that used in clinical decision-making. In addition, effective model fusion is beneficial in improving model performance. This study also provides research ideas regarding the application of AI predictive models to health management and early intervention, particularly for other chronic and emergent diseases in the future. There have been calls for sharing anonymized clinical data at the patient level,<sup>[16]</sup> which, along with the establishment of large datasets, would greatly benefit the continuous improvement of prediction models and lead to improvements in clinical treatments that save the lives.

## Financial support and sponsorship

This research was financially supported by the Natural Science Foundation of Beijing (No. M21012), National Natural Science Foundation of China (No. 82174533), and Key Technologies R and D Program of the China Academy of Chinese Medical Sciences (No. CI2021A00920).

## Conflicts of interest

There are no conflicts of interest.

## REFERENCES

- Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;395:497-506.
- The Lancet. Emerging understandings of 2019-nCoV. *Lancet* 2020;395:311.
- Liu X, Shi S, Xiao J, Wang H, Chen L, Li J, *et al.* Prediction of the severity of the coronavirus disease and its adverse clinical outcomes. *Jpn J Infect Dis* 2020;73:404-10.
- Peng Y, Xu B, Sun B, Han G, Zhou YH. Importance of timely management of patients in reducing fatality rate of coronavirus disease 2019. *J Infect Public Health* 2020;13:890-2.
- Lagadinou M, Salomou EE, Zareifopoulos N, Marangos M, Gogos C, Velissaris D. Prognosis of COVID-19: Changes in laboratory parameters. *Infez Med* 2020;28:89-95.
- Soraya GV, Ulhaq ZS. Crucial laboratory parameters in COVID-19 diagnosis and prognosis: An updated meta-analysis. *Med Clin (Barc)* 2020;155:143-51.
- Liu SP, Zhang Q, Wang W, Zhang M, Liu C, Xiao X, *et al.* Hyperglycemia is a strong predictor of poor prognosis in COVID-19. *Diabetes Res Clin Pract* 2020;167:108338.
- Sahu BR, Kampa RK, Padhi A, Panda AK. C-reactive protein: A promising biomarker for poor prognosis in COVID-19 infection. *Clin Chim Acta* 2020;509:91-4.
- Liao D, Zhou F, Luo L, Xu M, Wang H, Xia J, *et al.* Haematological characteristics and risk factors in the classification and prognosis evaluation of COVID-19: A retrospective cohort study. *Lancet Haematol* 2020;7:e671-8.
- Hu L, Chen S, Fu Y, Gao Z, Long H, Ren HW, *et al.* Risk factors associated with clinical outcomes in 323 coronavirus disease 2019 (COVID-19) hospitalized patients in Wuhan, China. *Clin Infect Dis* 2020;71:2089-98.
- Ayanian S, Reyes J, Lynn L, Teufel K. The association between biomarkers and clinical outcomes in novel coronavirus pneumonia in a US cohort. *Biomark Med* 2020;14:1091-7.
- Wu J, Song S, Cao HC, Li LJ. Liver diseases in COVID-19: Etiology, treatment and prognosis. *World J Gastroenterol* 2020;26:2286-93.
- Toraih EA, Elshazli RM, Hussein MH, Elgaml A, Amin M, El-Mowafy M, *et al.* Association of cardiac biomarkers and comorbidities with increased mortality, severity, and cardiac injury in COVID-19 patients: A meta-regression and decision tree analysis. *J Med Virol* 2020;92:2473-88.
- Guo W, Li M, Dong Y, Zhou H, Zhang Z, Tian C, *et al.* Diabetes is a risk factor for the progression and prognosis of COVID-19. *Diabetes Metab Res Rev* 2020;36:e3319.
- Jayawardena R, Misra A. Balanced diet is a major casualty in COVID-19. *Diabetes Metab Syndr* 2020;14:1085-6.
- Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* 2020;369:m1328.
- Yan L, Zhang H, Goncalves J, Xiao Y, Wang M, Guo Y, *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nat Mach Intell* 2020;2:283-8.
- Wang F, Hou H, Wang T, Luo Y, Tang G, Wu S, *et al.* Establishing a model for predicting the outcome of COVID-19 based on combination of laboratory tests. *Travel Med Infect Dis* 2020;36:101782.
- Shang W, Dong J, Ren Y, Tian M, Li W, Hu J, *et al.* The value of clinical parameters in predicting the severity of COVID-19. *J Med Virol* 2020;92:2188-92.
- He B, Zhong A, Wu Q, Liu X, Lin J, Chen C, *et al.* Tumor biomarkers predict clinical outcome of COVID-19 patients. *J Infect* 2020;81:452-82.
- Zhang J, Yu M, Tong S, Liu LY, Tang LV. Predictive factors for disease progression in hospitalized patients with coronavirus disease 2019 in Wuhan, China. *J Clin Virol* 2020;127:104392.
- L'Heureux A, Grolinger K, Elyamany HF, Capretz MA. Machine learning with big data: Challenges and approaches. *IEEE Access* 2017;5:7776-97.
- Sperrin M, Grant SW, Peek N. Prediction models for diagnosis and prognosis in Covid-19. *BMJ* 2020;369:m1464.
- Zhao JY, Yan JY, Qu JM. Interpretations of "diagnosis and treatment protocol for novel coronavirus pneumonia (Trial Version 7)". *Chin Med J (Engl)* 2020;133:1347-9.
- Kotsiantis SB. Decision trees: A recent overview. *Artif Intell Rev* 2013;39:261-83.
- de Menezes FS, Liska GR, Cirillo MA, Vivanco MJ. Data classification with binary response through the Boosting algorithm and logistic regression. *Expert Syst Appl* 2017;69:62-73.
- Brajer N, Cozzi B, Gao M, Nichols M, Revoir M, Balu S, *et al.* Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA Netw Open* 2020;3:e1920733.
- Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: How can we know it works? *J Am Med Inform Assoc* 2019;26:1651-4.
- van der Ploeg T, Nieboer D, Steyerberg EW. Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *J Clin Epidemiol* 2016;78:83-9.
- Wu MY, Yao L, Wang Y, Zhu XY, Wang XF, Tang PJ, *et al.* Clinical evaluation of potential usefulness of serum lactate dehydrogenase (LDH) in 2019 novel coronavirus (COVID-19) pneumonia. *Respir Res* 2020;21:171.
- Li C, Ye J, Chen Q, Hu W, Wang L, Fan Y, *et al.* Elevated Lactate Dehydrogenase (LDH) level as an independent risk factor for the severity and mortality of COVID-19. *Aging (Albany NY)* 2020;12:15670-81.

32. Gao M, Wang Q, Wei J, Zhu Z, Li H. Severe Coronavirus disease 2019 pneumonia patients showed signs of aggravated renal impairment. *J Clin Lab Anal* 2020;34:e23535.
33. Liang W, Yao J, Chen A, Lv Q, Zanin M, Liu J, *et al.* Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* 2020;11:3543.
34. Cowley LE, Farewell DM, Maguire S, Kemp AM. Methodological standards for the development and evaluation of clinical prediction rules: A review of the literature. *Diagn Progn Res* 2019;3:16.
35. Riley RD, Ensor J, Snell KI, Harrell FE Jr., Martin GP, Reitsma JB, *et al.* Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
36. Liao W, Zou B, Zhao R, Chen Y, He Z, Zhou M. Clinical interpretable deep learning model for glaucoma diagnosis. *IEEE J Biomed Health Inform* 2020;24:1405-12.
37. Wynants L, van Smeden M, McLernon DJ, Timmerman D, Steyerberg EW, Van Calster B, *et al.* Three myths about risk thresholds for prediction models. *BMC Med* 2019;17:192.
38. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: The Achilles heel of predictive analytics. *BMC Med* 2019;17:230.